Audiovisual Perception of Mandarin Lexical Tones

Rui Wang

A thesis submitted to the Faculty of the Graduate School of
Bournemouth University, in partial fulfilment of the
requirements for the degree of Doctor of Philosophy

October 2018

ABSTRACT

It has been widely acknowledged that visual information from a talker's face, mouth and lip movements plays an important role in speech perception of spoken languages. Visual information facilitates speech perception in audiovisual congruent condition and even alters speech perception in audiovisual incongruent condition. Audiovisual speech perception has been greatly researched in terms of consonants and vowels, and it has been thought that visual information from articulatory movements conveys phonetic information (e.g. place of articulation) that facilitates or changes speech perception. However, some research give rise to another type of visual information which conveys non-phonetic information (e.g. timing cue), affecting speech perception. The existence of these two types of visual information in audiovisual integration process suggests that there are two levels of audiovisual speech integration in different stages of processing. The studies in this dissertation focused on audiovisual perception of Mandarin lexical tones. The results of the experiments which employed behavioural and event-related potential measures provided evidence that visual information has an effect on auditory lexical tone perception. First, lexical tone perception benefits from adding visual information of corresponding articulatory movement. Second, the duration perception of lexical tones is changed by incongruent visual information. Moreover, the studies revealed that there are two types of visual information—timing (non-phonetic) cue and tone duration (phonetic/ tonetic) cue—involving in audiovisual integration process of Mandarin lexical tone. This finding further supports that audiovisual speech perception comprises non-phonetic and phonetic-specific levels of processing. Non-phonetic audiovisual integration could start in an early stage while phonetic-specific audiovisual integration could occur in a later stage of processing. Lexical tones have not been paid much attention in the research of audiovisual speech perception. The current studies fill the gap in the research of Mandarin lexical tone perception, and the findings from these experiments have important theoretical implications for audiovisual speech processing.

Copyright Statement

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my three supervisors: Dr. Biao Zeng, Dr. Xun He and Dr. Bernhard Angele.

First of all, I am deeply indebted to my main supervisor Dr. Biao Zeng for his continuous support, motivation, inspiring suggestions and giving me great freedom during my PhD studies, and also for offering me wonderful academic opportunities to enrich my research experience.

I would also like to sincerely thank Dr. Xun He for his professional guidance which helped me go through the difficult research and writing of this dissertation. Without his support, it would not be possible to conduct some of the studies in this dissertation.

I am deeply grateful to Dr. Bernhard Angele, especially for his encouragement, patience and insightful comments and suggestions which motivated me to elevate my research and writing.

I also thank Dr. Nan Jiang for allowing me to record his face and voice that I can use as important experiment materials in some of my studies. I would like to thank all the participants for their dedications.

My deep appreciation goes to my PhD colleagues Pree Thiengburanathum, Pengcheng Liu and Yan Wang. Thank you for being there for me along my PhD journey. I would not have been gone through the hard times without you.

And finally, my deepest gratitude goes to my parents. Thank you for your unconditional love and support. I would like to dedicate this dissertation to you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC: accuracy

ANOVA: analysis of variance

AO: auditory-only

AO’: difference between audiovisual and visual-only brain response

AV: audiovisual

CP: categorical perception

CV: consonant-vowel

CVV: consonant-vowel-vowel

DI: discrimination index

EEG: electroencephalography

EOG: electrooculograms

ERP: event-related potential

F0: fundamental frequency

ISI: inter-stimuli interval

ITI: inter-trial interval

MEG: magnetoencephalography

MMN: mismatch negativity

ms: millisecond

RMS: root mean square

ROI: region of interest

RT: reaction time

SNR: signal-noise-ratio

SPL: sound pressure level

STS: superior temporal sulcus

T: tone

TBU: tone-bearing-unit

VO: visual-only

VOT: voice onset time

Author’s Declaration

I hereby declare that the work presented in the thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

SUMMARY OF THE DISSERTATION

In face-to-face communication, speech perception involves audiovisual processing. Visual information from a talking face (movements of the mouth, lips, jaw, eyebrows and head) is beneficial to the listener. Speech perception can be significantly improved when adding visual speech along with corresponding auditory speech (audiovisual benefit effect). The perception of auditory speech can also be altered by visual speech information consequently generating an illusory percept (McGurk effect). Two types of visual information from visual speech movements are considered to be factors of the visual effect in audiovisual speech processing: phonetic visual information (lip-reading information, e.g. place of articulation) and non-phonetic visual information (timing information, e.g. visual information predicting the timing of the auditory signal). Audiovisual speech processing has been studied extensively in terms of consonants and vowels. However, research on audiovisual lexical tones is very limited in the literature, and the effect of visual information on auditory lexical tone processing is unclear.

The studies presented in this dissertation investigated Mandarin audiovisual lexical tone perception with the audiovisual benefit effect and the McGurk effect. Both behavioural and event-evoked potential (ERP) methods were employed.

The behavioural studies in Chapter 2 found that phonetic visual information (possibly tone duration) facilitates lexical tone perception. The studies examined Mandarin lexical tones with audiovisual, auditory-only and visual-only modalities in identification and discrimination tasks. To test the audiovisual benefit effect, audiovisual lexical tones were compared with auditory lexical tones in identification and discrimination tasks in clear and noise conditions. The results showed that lexical tone perception was better in the audiovisual modality than in the auditory-only modality in noise. The audiovisual benefit was consistently stronger in the contrast between the dipping tone (Tone 3) and the falling tone (Tone 4). It suggests that these two tones have more salient visual cues than other tones, and this visual cue could be tone duration. Additionally, the results for lexical tone identification and discrimination in the visual-only condition showed that lip-reading

performance was above chance level, suggesting the availability of phonetic or tonetic visual cues in Mandarin lexical tones.

The ERP studies in Chapter 3 measured the effect of visual lexical tones on auditory lexical tone processing in the early components N1 and P2. There is evidence that auditory N1/P2 is reduced in amplitude and shortened in latency with the influence (aid) of visual speech. The ERP results found that auditory N1 evoked by audiovisual lexical tones was reduced in amplitude and accelerated in latency compared to that evoked by auditory-only lexical tones. This result suggests that visual information affects the sensory processing of auditory tones at an early stage (about 188 ms) before phonetic (categorical) processing. The N1 reduction of lexical tones showed a clear right lateralisation, suggesting that the audiovisual interaction of lexical tones in N1 is non-phonetic processing.

In studies of the McGurk effect regarding Mandarin lexical tones in Chapter 4, Tone 3 and Tone 4 were chosen to combine audiovisual incongruent lexical tone stimuli (e.g. visual Tone 4 paired with auditory Tone 3) due to their salient tone duration feature. In the behavioural study in Section 4.1, the lexical tone identification rate was compared between congruent and incongruent audiovisual lexical tones, which only differed in visual input. The results showed that the response time for the incongruent lexical tone (auditory Tone 3 paired with visual Tone 4) was shorter than the congruent lexical tone (Tone 3). This result suggests that perception of auditory tone duration is modified by mismatched visual duration information. That is, under the influence of visual Tone 4, auditory Tone 3 is perceived to be shorter than the original duration. This indicates that the incongruent audiovisual lexical tone led to an illusory lexical tone different in duration. This visual modification could be processing on a non-phonetic level because the mismatched visual did not alter the identification (i.e. categorical tone perception).

The ERP study of the McGurk effect on lexical tones in Section 4.2 investigated the (mismatch negativity) MMN activation of incongruent audiovisual lexical tone processing. MMN is used as an index for pre-attentive detection of an infrequently appearing stimulus from a frequently presenting stimulus. Even without any physical change to the auditory inputs between the incongruent and congruent lexical tones, the incongruent audiovisual lexical tone deviance activated MMN in the frontal electrodes bi-laterally within 180–280 ms after auditory tone onset. This suggests that incongruent visual information modulates the perception of auditory tones in the pre-attentive stage of processing.

In summary, current studies have found behavioural and electrophysiological evidence that visual information influences (facilitates or alters) the perception of auditory lexical tones on both non-phonetic and phonetic levels of processing. The findings for the audiovisual processing of lexical tones support the hypothesis of multiple-stage audiovisual integration in speech.

Chapter 1 General Introduction

1.1 Research aims and significance

In everyday life, a substantial amount of event information is always around us. Amazingly, our brains can accurately and effortlessly perceive these multiple pieces of information as one united percept. This cannot be completed through merely processing every single event via a particular sensory perceiver one at a time. Instead, these events are processed in parallel and interact at various process levels. Speech perception is the most common event involving multisensory processing or, more commonly, audiovisual processing. In face-to-face communication, listeners always involuntarily watch interlocutors' faces in order to hear them better, especially in a noisy environment. This is because speech perception and the comprehension of auditory speech, along with facial gestures, are always more informative than that of auditory speech alone. A substantial amount of research on language processing is fundamentally auditory processing-based, ignoring the fact that speech perception in the natural environment is essentially an audiovisual process. Understanding the process of audiovisual integration/ interaction in speech enables us to have a better understanding of the mechanisms involved in speech perception.

In natural audiovisual speech events, visual speech information from a talking face (e.g. articulatory mouth/ lips movement, head and eyebrows movement and even glottal movement) plays an important part in interacting with auditory speech signals during speech perception. Two typical phenomena provide strong evidence that visual speech information influences auditory speech perception: the audiovisual benefit effect and the McGurk effect. The audiovisual benefit effect refers to the fact that speech intelligibility is better when presenting auditory speech along with visual speech information (i.e. audiovisual speech) compared to perception during auditory only speech (e.g. Sumby & Pollack, 1954). The McGurk effect refers to the fact that mismatched visual speech changes the perception of dubbed auditory syllables of certain incongruent audiovisual syllables.

McGurk and MacDonald (1976) demonstrated that when presenting auditory speech, /ba/ was paired with visual speech /ga/, and perceivers often heard a novel syllable /da/ or /tha/.

Audiovisual perception has been extensively studied in segmental speech (consonants and vowels), especially in consonants of English language, but audiovisual studies of Mandarin lexical tones are very few in the literature. Hence it is hard to know whether results based on audiovisual perception studies of segmental speech are also compatible with the perception of audiovisual lexical tones. Lexical tones are an essential part of tonal languages, which takes up 60–70 per cent of the of the world's languages, and Mandarin (Standard Chinese), as one of these tonal languages, is the most spoken language in the world (c. 850 m. speakers) (Yip, 2002). When generalising theoretical implications from studies of audiovisual speech perception, these are restricted if we only discuss the implications of various audiovisual speech studies based on segmental speech without involving audiovisual perception of lexical tones. Furthermore, from a practical perspective, to understand audiovisual perception of lexical tones is advantageous to second language learning students and hearing impairment patients. In recent years, the need to learn Mandarin Chinese has been steadily increasing (Department of Education, 2017). There is a consensus that, for those students whose native languages are not tonal, acquiring lexical tones in Mandarin is one of the biggest obstacles in their learning process (Kiriloff, 1969). For Mandarin speakers who have hearing difficulties, even with the help of cochlear implants, accurately perceiving auditory lexical tone information remains a problem (Huang et al., 2005). Learning and hearing difficulties might be relieved if Mandarin learners and speakers are able to exploit the visual cues of lexical tones, hence improving their perception and production in communication.

The main goal of this dissertation is to fill this gap in studies of audiovisual speech perception through investigating Mandarin audiovisual lexical tone perception via the audiovisual benefit effect and the McGurk effect. The audiovisual benefit effect is generally believed to be associated with the visibility and predictability of visual speech contributing to auditory speech perception, and the McGurk effect refers to how incongruent visual speech changes auditory speech perception. Certain speech sounds can be visually recognised based on mouth movements during production, such as consonants with different places of articulation and vowels with various degrees of lip roundedness or openness. For natural speech, lip movements articulating speech sounds precede the actual sounds produced. The speech information conveyed by lip movements predicts the phonetic information of upcoming speech sounds. For example, the place of articulation of a bilabial (the action of lips closing) predicts the consonants /b/, /p/ and /m/. Unlike consonants and vowels, the production of lexical tones is less visible and mouth movements are less predictive of upcoming auditory tones. This is because producing lexical tones relies on vocal fold vibration of the larynx, which is less detectable from mouth movements. The way of articulating lexical tones renders them auditorily dependent. However, this does not necessarily exclude the possibility that the perception of lexical tones can be influenced by the visual information of lexical tones. Congruent visual input from lexical tones can be useful to facilitate tone perception as visual consonants and vowels operate, and incongruent visual input of lexical tones could even alter auditory tone perception on a certain level like the McGurk effect.

More specifically, in this project, a series of studies were conducted to investigate the audiovisual benefit effect and the McGurk effect on Mandarin lexical tones with behavioural and electrophysiological measures. Through these studies, the questions of to what extent visual speech can influence auditory speech in audiovisual lexical tones, and whether the audiovisual integration of lexical tones is different from that of segmental speech, will be answered. Moreover, based on the findings from these studies, the possible theoretical implications for the mechanism of audiovisual speech integration will be discussed.

This introductory chapter starts with some background information about lexical tones in Mandarin. As audiovisual speech perception is closely related to acoustic perception in auditory speech and the corresponding visual correlates derived from articulatory movements, before going into more detail about audiovisual speech perception, the mechanism of lexical tones articulation, and the perception and acquisition of Mandarin lexical tones in the auditory domain, will be discussed. In terms of the literature on audiovisual speech perception, the chapter reviews relevant studies that focus on the audiovisual benefit effect and the McGurk effect in both segmental (consonants and vowels) and prosodic speech (intonation, stress), and then audiovisual studies on lexical tones. Finally, the chapter raises a series of research questions and propose hypotheses for audiovisual lexical tone perception in current studies, which will be answered during the experiments in the following chapters.

1.2 Background information on Mandarin lexical tones

A lexical tone is the fundamental frequency (F0) or pitch variation over a syllable that can distinguish the lexical or grammatical meaning of a word. F0 is an acoustic signal indicating the number of the pulses per second that the signal contains, and F0 is measured in Hertz. The F0 of a speech signal can be perceived as pitch (Yip, 2002). Variations in pitch can be perceived as different lexical tones or intonation. A lexical tone is different from intonation that uses pitch to convey pragmatic meaning in many languages. Intonation, for example in ‘yes’ with rising or falling pitch, can indicate different pragmatic meanings or attitudes in English (e.g. rising pitch “yes?” would be used to indicate asking for affirmation/ negation or replying to someone and expecting further conversation, while falling pitch “yes!” would be used to indicate assent or to confirm that someone has received an instruction). Lexical tones, on the other hand, change the core meaning of a word. For example, in Mandarin Chinese, the syllable /ma/, when produced with a high-level tone means *mother*, while /ma/ in a rising tone means *hemp*, /ma/ in a dipping tone means *horse* and /ma/ in a falling tone means *to scold* (see Table 1.1). Lexical tones exist in many East Asian languages, such as Mandarin Chinese, Thai, Vietnamese, etc. This dissertation focuses particularly on the audiovisual perception of lexical tones in Mandarin Chinese. In order to better understand the audiovisual aspect of Mandarin lexical tones, this section first introduces general background knowledge about Mandarin lexical tones in terms of acoustic characteristics and auditory perception.

1.2.1 Acoustic characteristics of Mandarin lexical tones

1.2.1.1 Categories of Mandarin lexical tones

There are four tone categories, based on different types of pitch variation: Tone 1 (T1), which is a high-level tone; Tone 2 (T2), which is a high-rising tone; Tone 3 (T3), which is a low-falling-rising tone; Tone 4 (T4), which is a high-falling tone. As in the example mentioned previously, the monosyllable /ma/ in four tones can convey at least four different meanings, and these pitch variations can be transcribed, as shown in Table 1.1, below. Chao (1930) proposed a 5-scale system to describe the pitch range of Mandarin lexical tones. In the ‘Pitch’ column, the numbers are representations of the pitch range in 1–5 digits, in which 1 refers to the lowest pitch and 5 to the highest one. Two different digits refer to the pitch at beginning and end of a syllable (for example, 35 — high rising tone / T2 — refers to the pitch at scale 3 at the beginning and the pitch at scale 5 at the end of the syllable (Reetz & Jongman, 2009).

Table 1.1 Sample monosyllable /ma/ with four lexical tones in Mandarin and three ways of transcribing them. The table is adapted from Reetz and Jongman (2009).

Tones	Description	Pitch	Tone marker	Tone letters	Gloss
Tone 1	high-level	55	/mā/	ma˥	'mother'
Tone 2	high-rising	35	/má/	ma˧˥	'hemp'
Tone 3	low-falling-rising	214	/mă/	ma˨˩˦	'horse'
Tone 4	high-falling	51	/mà/	ma˥˩	'scold'

1.2.1.2 Tone-Bearing Unit

A Mandarin lexical tone, like the lexical tones in other East-Asian tonal languages, is a separate unit from segmental and prosodic (intonation, stress) aspect of speech. It has to attach to a syllable or a mora to be realised and pronounced, and a syllable or a mora that carries a lexical tone is called a Tone-Bearing Unit (TBU). TBUs for Chinese language (including dialects) are syllabic-based, in that any syllable can carry lexical tones (Yip, 1995, 2002). An individual lexical tone is always realised through a monosyllable, which is also a morpheme or an independent word in Mandarin. Most experiments on lexical tone perception use syllable stimuli with different consonant and vowel combinations, and different lexical tones.

1.2.1.3 Pitch Contours

Mandarin lexical tones are known as contour-based tones with multiple contour tones in the tonal inventory (Zhang, 2014). Contour tones have rich pitch variation (rising or falling) over the course of a syllable (Duanmu, 2007; Yip, 2002; Zhang, 2002). Due to complex pitch movements, pitch height alone (high/low) is unable to fully defined pitch variation in Mandarin lexical tones. As can be seen from the plot of the F0 contours of the four tones in the monosyllable /ma/ in Figure 1.1 below, with the exception of T1 which is considered to be a level tone, T2, T3 and T4 have significant pitch movements or trajectories in F0 within the syllable. T1 starts at a higher pitch and remains at a similar height until the offset. However, T1 is not perfectly level as it "wobbles" slightly towards the end. T2 appears to be lower than T1 at the beginning. It drops a little early on and then gradually rises until the end. Similarly, T3 drops significantly in the middle of the trajectory and then rises until the syllable offset. T3 is generally lower and more complex than T2 over the course of its production. T4 is highest at the onset but falls quickly to the lowest pitch by the end.

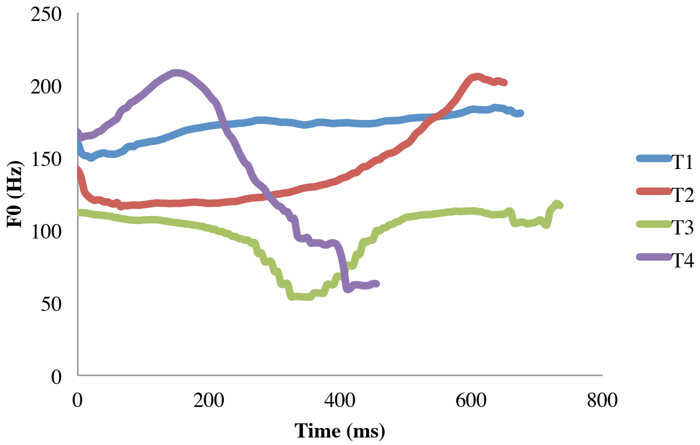


Figure 1.1 F0 contours of the four Mandarin lexical tones of monosyllable /ma/ in isolation.

The syllable /ma/ was selected from a corpus which was recorded by one of the speakers in the current project. The pitch contours of the four lexical tones were measured and plotted with Praat (Boersman & Weenink, 2013).

1.2.1.4 Duration and intensity

Mandarin lexical tones in isolation monosyllables are systematically different in duration. Among the four tones, T3 usually has the longest duration, T4 has the shortest duration, and the duration of T1 and T2 are often in the middle (Whalen & Xu, 1992; Xu, 1997). In terms of the intensity or loudness of the four tones in isolation, T4 has the highest average intensity while T3 has the lowest intensity, and T1 and T2 are at an intermediate level (Chang & Yao, 2007).

1.2.2 Production and perception of Mandarin lexical tones

1.2.2.1 Articulation of lexical tones

The articulation of lexical tones is fundamentally determined by the vocal fold vibration frequency in the larynx (Hayes, 2009; Yip, 2002). The laryngeal mechanism is composed of vocalic muscles and several sets of cartilage. The vocalic muscles (vocal folds) are two strips of tissue with an opening (called the glottis) in-between. The arytenoid cartilage that attaches to the rear of the vocal fold can control the distance between the vocal folds (see

Figure 1.2). If the vocal folds get closer together, air is allowed to pass through the narrowing glottis from the lungs into the mouth, and then vocal folds vibrate in order to generate sounds. The thyroid cartilage that attaches to the front of the vocal folds can tighten or loosen the vocal folds, hence changing the frequency of vocal fold vibration and, as a result, changing the level of the pitch. If the vocal folds are tensed, the pitch rises. If the vocal folds are relaxed, the pitch drops.

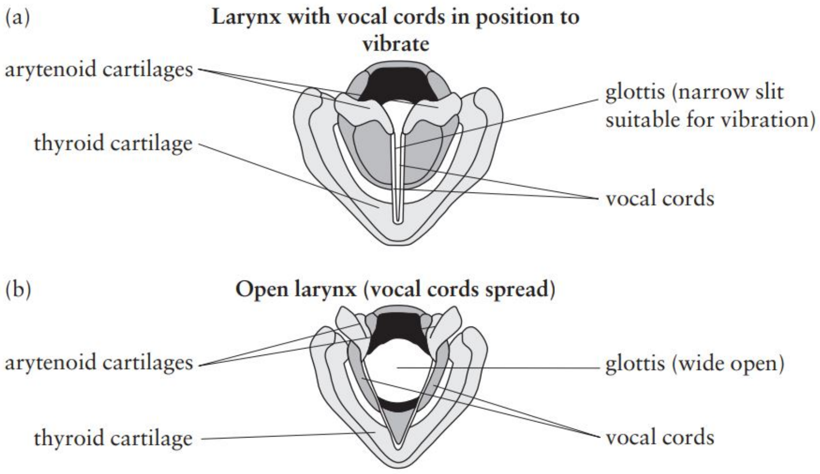

(a)
Larynx with vocal cords in position to vibrate
arytenoid cartilages
thyroid cartilage
glottis (narrow slit suitable for vibration)
vocal cords
(b)
Open larynx (vocal cords spread)
arytenoid cartilages
thyroid cartilage
glottis (wide open)
vocal cords

Figure 1.2 The larynx, adapted from Hayes (2009).

1.2.3 Perception of Mandarin lexical tones

Primary cue: fundamental frequency

The (auditory) perception of lexical tones is closely associated with how listeners catch variation in the F0 information in a syllable. Early studies have shown that F0 information (including F0 height and F0 contour) is the most crucial cue for lexical tone perception. Howie (1976) compared synthesised stimuli with pitch patterns to monotonous whisper-like stimuli, and the results showed that tone identification was significantly better for stimuli with a pitch pattern. Fok (1974) and Abramson (1979) also found that native speakers achieved high discrimination rates for lexical tones when F0 information remained intact and other information was removed. Gandour (1984) further broke down the F0 variation into two cues: F0 height and F0 contour. He found that native speakers of Mandarin preferred to put more weight on F0 contour than F0 height, while English speakers preferred to use F0 height as the main cue, rather than F0 contour. Among the four tones, the high rising tone (T2) and the low falling rising/dipping tone (T3) are known to be easily confused due to similar pitch contours. Even native speakers often misperceive them with one another. According to Moore & Jongman (1997), native speakers distinguish T2 and T3 by using the turning point and ΔF0 of the pitch. The turning point is the point when the pitch contour starts rising from falling, and ΔF0 refers to the pitch change from the onset to the turning point. The turning point for T2 is earlier than that of T3, and the ΔF0 of T2 is smaller than that of T3.

Secondary cues: amplitude and duration

In addition to F0 information as the major cue for lexical tones, other acoustic information sources, such as amplitude and duration, also contribute to lexical tone perception as secondary cues. Whalen and Xu (1992) reported that listeners achieved about 80% identification accuracy with the help of amplitude envelope information when F0 information was removed. They found that the amplitude envelope was highly correlated with the F0 contour for T3 and T4. Fu and Zeng (2000) also found that T3 and T4 recognition benefited more from temporal envelope information when no F0 signal was available.

Liu and Samuel (2004) investigated the secondary cues of Mandarin lexical tones by neutralising F0 information through a signal processing method and producing tones in a whisper. The results showed that tone recognition was above chance level for stimuli devoid of F0 information. Interestingly, the results showed that T3 appeared to have the highest recognition. Correlation of tone duration with tone recognition was found in the human whisper condition (no F0 presented), hence they proposed that tone duration could be a cue contributing to tone recognition when F0 is absent.

Categorical Perception

Speech sounds are perceived in a categorical manner, for example, when perceivers distinguish stop consonants in the /b/-/p/ continuum, they only hear either /b/ or /p/, but no intermediate sounds (Liberman et al., 1957). This phenomenon is known as categorical perception (CP). Identification studies on lexical tones show that perceiving lexical tones is also categorical. Wang (1976) found that Mandarin native speakers showed a clear tone category boundary between high-level tone (T1) and high rising tone (T2). Peng et al. (2010) also found a categorical boundary in their results for tone identification, but their results for tone discrimination showed a weaker pattern for CP. Francis et al. (2003) tested the CP of lexical tones in Cantonese, they reported a similar result, i.e. that tone identification had a shape tone category boundary but not in tone discrimination. Peng et al. (2010) also tested the CP of lexical tones with non-native speakers (German), who did not recognise lexical tones based on CP, but rather relied more on the detection of psychoacoustic differences between tone contrasts. The CP of Mandarin lexical tones seems to be affected by task paradigm, language experience or phonological knowledge.

Reaction time for lexical tone perception

In terms of processing time, there is some evidence showing that lexical tones are perceived later, compared to segmental speech. Cutler and Chen (1997) measured the discrimination accuracy and response time for consonants, vowels and lexical tones in Cantonese. The results revealed that lexical tone discrimination had the poorest accuracy and the longest response time. Yip (2002) explained that the slow reaction time might be due to perceivers being unable to perceive tone contour movement until they hear almost a whole syllable, therefore processing a lexical tone requires more time than segmental speech.

1.2.4 Lexical tone acquisition in infants

The studies of lexical tones perception discussed above are based on adult native speakers. Mandarin monolingual babies acquire lexical tones as early as 6 months of age. Tsao (2008) examined tonal discrimination by Taiwan-Mandarin-learning babies from 10–12 months old. He found that T1-T3 discrimination achieved the highest discrimination rate (73%), while T2-T3 (61%) and T2-T4 (58%) discrimination rates were comparatively lower. Shi et al. (2017) tested T2-T3 discrimination by 4–13-month-old Mandarin monolingual infants, and the result showed that T2-T3 was successfully categorised (also see Shi et al, 2010). Wong et al. (2005) examined the tone recognition of toddlers (3 years old) who were in the one- to multi-word stage of language acquisition, and they found that all four tones were accurately recognised (nearly 90%), except for T3 (69%). In their data, T3 was frequently misperceived as T2. Electrophysiological studies provide evidence that infants are able to detect pitch differences in early infancy. Cheng and Lee (2018) measured the MMN component of lexical tones, they found that a T1/T3 tone contrast elicited an adult-like MMN in 12-, 18- and 24-month-old toddlers in Taiwan. T2/T3 tone contrast failed to elicit adult-like MMN in 12- and 18-month-old infants. The results indicate that Mandarin-learning babies are able to detect lexical tone categories before they can produce in the later stage in the first two years of life. However, the data also suggest that T2 and T3 may be acquired later than T1 and T4. Based on these studies, infants' perception of Mandarin lexical tones starts before the verbal stage (6 months), and all four lexical tones can be perceived as different tone categories by about 2–3 years of age. The four lexical tones are perceived in a particular order: T1 and T4 are acquired before T2 and T3. T3 may be the most difficult tone to learn and produce, which could be due to the late acquisition of T3, a physical limitation in articulation.

According to the articulatory mechanism and acoustic contrastive features of Mandarin lexical tones introduced above, the visibility of lexical tones is clearly weaker than for consonants and vowels. Since lexical tone perception greatly depends on the pitch variation that is produced through vocal vibration, lexical tones are difficult to recognise from mouth/ lip movements (i.e. less lip-readable). Because of this, one might suspect that visual information can contribute to the perception of lexical tones. Yet, there is evidence supporting that visual articulatory movement of lexical tones can improve the perception of Mandarin auditory tones, suggesting that visual cues for Mandarin lexical tones are available (Mixdorff et al., 2005b; Smith & Burnham, 2012). Moreover, there is evidence showing that the visual speech influence on auditory speech perception is not only determined by distinctive phonetic-specific visual information but also by timing information from the mouth/ lip movements (e.g. Bernstein et al., 2004; Grant & Seitz, 2000; Schwartz et al., 2004). The next two sections will review the literature on audiovisual speech perception and discuss the availability of visual cues for lexical tone perception in terms of two types of visual speech information.

1.3 Audiovisual speech perception

In many speech perception studies, the auditory aspects of a particular language have generally been the default modality for investigating speech processing in the human mind. This is not surprising, given that the speech signal that an individual receives relies heavily on acoustic waveforms and they process speech information through their auditory system. However, there has been evidence for a long time indicating that listeners do not only use auditory information but also visual information whenever it is available (e.g. Sumby & Pollack, 1954). Visualising an interlocutor's face, particularly the mouth area, can have an effect on auditory speech perception. It improves the perception, intelligibility and comprehension of speech (the audiovisual benefit effect) (e.g. Arnold & Hill, 2001; Bernstein et al., 2004; Grant & Seitz, 2000; Jesse & Janse, 2012; MacLeod & Summerfield, 1987; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollack, 1954). Additionally, visualising incongruent visual articulation paired with auditory speech information also results in illusory speech perception in certain cases (the McGurk effect) (McGurk & MacDonald, 1976). These visual effects could be due to two types of visual information from visual speech: phonetic-specific visual information (speech articulatory movements) and non-phonetic visual information (timing information predicting upcoming auditory speech), which integrate with auditory speech signals at different levels of processing, possibly with different time courses (Eskelund et al., 2011; Kim & Davis, 2014; Klucharev et al., 2003; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004; Soto-Faraco & Alsius, 2009; Stekelenburg & Vroomen, 2007).

In the literature on audiovisual speech perception, most research focuses on the segmental aspect of speech (consonants and vowels), and some focuses on prosodic speech (stress and intonation); therefore, the theories and hypotheses regarding audiovisual speech integration processing that have been developed are also based on the findings of these studies (mainly on consonants). Therefore, it is essential to first review some important audiovisual studies on segmental and prosodic speech before starting to discuss the main theme of this dissertation: audiovisual lexical tone perception. This section will introduce the audiovisual benefit effect and the McGurk effect, and then review relevant audiovisual studies on segmental and prosodic speech (stress and intonation) and their theoretical implications for audiovisual speech perception associated with 'the two-visual-cue hypothesis'.

1.3.1 Two audiovisual effects

1.3.1.1 Audiovisual benefit effect in speech

The audiovisual benefit effect has been studied extensively in the past few decades. The audiovisual benefit effect is robust across various sizes of speech units, from syllables to words, sentences and passages. Compared to performance in the auditory-only condition, watching speakers' faces while listening to their utterances improves speech perception, word intelligibility, comprehension and the detection of sensitivity to auditory signals, especially in adverse audition conditions (e.g. Arnold & Hill, 2001; Bernstein et al., 2004; Grant & Seitz, 2000; Jesse & Janse, 2012; MacLeod & Summerfield, 1987; Ross et al., 2007; Schwartz et al., 2004; Sumby & Pollack, 1954).

First, audiovisual syllables are recognised more accurately and faster, particularly in noise. The better performance of audiovisual syllable recognition actually reflects the audiovisual benefit effect in consonants, in which visual articulatory movements (lips, teeth and tongue movements) provide critical consonantal features (e.g. place of articulation), which complements hearing, especially in noise (e.g. MacLeod & Summerfield, 1987; Summerfield, 1987; Summerfield & McGrath, 1984). Even when mouth/ lip articulatory movement does not provide any speech information (e.g. identifying two consonants having the same place of articulation), an audiovisual benefit effect can still be found (Schwartz et al., 2004). The audiovisual benefit effect is not only demonstrated in syllable recognition, but it has also been shown in a detection task, where presenting audiovisual signals increased the detectability of auditory speech in noise (Bernstein et al., 2004; Grant & Seitz, 2000; Kim & Davis, 2004; Lalonde & Holt, 2016).

Additionally, word identification is also more accurate in the audiovisual modality than in the auditory-only modality in a noisy environment (e.g. de la Vaux & Massaro, 2004; Ross et al., 2007; Sumby & Pollack, 1954). Word recognition processing can involve perceiving a physical (acoustic or visual) signal and interacting with phonological-lexical representation in long-term memory (Rönnberg et al., 2008). In auditory speech perception processing, a set of phonological neighbours of a target word can be activated, which then compete with each other in a selection process (Luce & Pisoni, 1998). In audiovisual word recognition, candidates that are activated by auditory speech and those that are activated by visual speech overlap, so there are fewer phonological-lexical competitors to select from. That is, the provision of visual speech input limits the number of phonological candidates, hence more effectively and accurately identifying target words (Peelle & Sommers, 2015; Tye-Murray et al., 2007).

The audiovisual benefit is also found in sentence and passage comprehension. Grant & Seitz (2000) tested visual enhancement in spoken sentences in noise. They found that adding visual input decreased the detection threshold of the auditory signal (and increased auditory signal sensitivity). They proposed that the visual enhancement of sensitivity could be due to the temporal co-occurrence of lip opening and dynamic auditory envelope change. That is, visual input (lips opening) functions as an indicator of the time of the amplitude envelope of auditory signals. Even without using auditory masking (e.g. noise), the audiovisual benefit effect can still be found in longer speech materials. Reisberg et al. (1987) and Arnold and Hill (2001) reported that the comprehension of auditory passages was significantly improved when presenting speakers' faces when the auditory signal was complete.

1.3.1.2 McGurk effect

Adding visual speech information not only enhances speech perception and intelligibility, it also alters auditory perception when visual articulation is mismatched with auditory speech. Compelling evidence for this is the McGurk effect, in which mismatched visual articulation biases the perception of auditory syllables, which results in hearing illusory novel syllables (McGurk & MacDonald, 1976). There are two types of response in the McGurk effect. The fusion response of the McGurk effect, where a visual palatal consonant (e.g. /ga/) is paired with of an auditory bilabial (e.g. /ba/), leads to a fusion consonant /da/ (McGurk and MacDonald, 1976). Another type is a combination response of the McGurk effect. For example, when a visual bilabial (e.g. /ba/) pairs with an auditory palatal (e.g. /ga/), perceivers frequently hear a consonantal cluster (e.g. /bga/) (MacDonald & McGurk, 1978). The response to the McGurk effect is an involuntary process that does not require extra attention because it is unaffected even though perceivers are informed that the audiovisual input is incongruent. In the fusion type of the McGurk effect, the discrepancy between auditory and visual speech inputs is barely noticeable (Summerfield & McGrath, 1984). However, in the combination type of the McGurk effect, inconsistency in audiovisual input can be clearly detected, yet the illusory consonant cluster is still perceived (Soto-Faraco & Alsius, 2007).

The McGurk effect is also found in incongruent audiovisual vowels. First, incongruent audiovisual vowels can have audiovisual fusion analogous to the illusory percept in consonants. Traunmüller and Öhrström (2007)) reported that the Swedish auditory syllable /geg/ (open) paired with visual /gyg/ (rounded) was perceived as /gøg/ (open and rounded). Valkenier et al. (2012) also found incongruent audiovisual vowels in Dutch, leading to illusory vowels. For example, when auditory /Y/ (rounded mid-high) was dubbed into visual /e/ (unrounded mid-high), the vowel was perceived as /I/ (unrounded high). Yet, at the same time, the incongruence/ incompatibility between auditory and visual vowels can be clearly detected by perceivers (Summerfield & McGrath, 1984), and it greatly disrupts vowel identification in either modality (Valkenier et al., 2012).

In addition to mismatched visual speech information for consonants and vowels, there are other non-speech-specific factors provided by visual speech input that can change speech perception. Green and Miller (1985) reported that the visual speaking rate altered the perception of the voice onset time (VOT) of voiced and voiceless consonants. Furthermore, presenting the gender of speakers' faces affected the perception of the categorical boundary between vowels /ʊ/ and /ʌ/ continuum (as in *hood-hud*) (Johnson et al., 1999)

1.3.2 Visual speech information

The audiovisual benefit effect and the McGurk effect provide strong evidence that visual speech information interacts with auditory speech processing, no matter whether the auditory signal is impaired or intact. Accounts of the visual effect on auditory speech perception are related to the two types of information provided by visual speech: phonetic visual information (e.g. visemic information) and non-phonetic visual information (e.g. timing information) (Eskelund et al., 2011; Kim & Davis, 2014; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004; Soto-Faraco & Alsius, 2009). Articulatory movements of the mouth, lips and even teeth can provide phonetic-specific information that complements the speech cues that are vulnerable in auditory signals (e.g. place of articulation for consonants, roundedness/ openness for vowel identity, etc.) (Campbell, 2008; Massaro & Jesse, 2007; Summerfield, 1987) and reduces the numbers of lexical neighbourhoods, hence constraining lexical completion (Peelle & Sommers, 2015; Tye-Murray et al., 2007). Additionally, visual speech input provides non-phonetic visual information (e.g. temporal or spatial information) that increases the sensitivity for detecting auditory signals (Bernstein et al., 2004; Grant & Seitz, 2000; Kim & Davis, 2004; Lalonde & Holt, 2016; Schwartz et al., 2004), because it decreases the uncertainty in auditory signal detection and relieves cognitive demands (Moradi et al., 2013; Moradi et al., 2017; Rönnberg et al., 2013).

1.3.2.1 Phonetic information provided by visual speech

Visual cues of segmental speech

The phonetic information provided by visual speech (articulatory movements) is relatively easy to observe in the production of segmental speech (i.e. consonants and vowels). With the production of a speech sound, the movements of lips, teeth, tongue and jaw convey some of the features for auditory speech. For instance, the consonantal place of articulation (e.g. bilabial) can be easily visualised. In noise, the acoustic place of articulation is easily impaired. The visual place of articulation can supplement the corresponding deficient acoustic signal (Summerfield, 1992). For instance, /m/ and /n/ are confusable in the acoustic signal, but they can be easily distinguished when adding articulatory movements. In addition to the complementarity of visual cues, redundant information from visual inputs also benefits auditory speech perception (Campbell, 2008; Massaro & Jesse, 2007).

Vowels, as consonants, are perceived better when articulatory movement is present (Moradi et al., 2017; Robert-Ribes et al., 1998). The critical cues for acoustic vowel identification are vowel height, the backness of the tongue position in the mouth and the roundedness of the lips (Grant & Walden, 1996; Kent, 1997). The visual articulation of vowels provides information about mouth roundness that is associated with those acoustic cues (Traunmüller & Öhrström, 2007), therefore enhancing vowel perception. However, the recognition of vowels benefits much less from visual input compared to the degree of the audiovisual benefit in consonants (Kim et al., 2009). Moradi et al. (2017) reported a similar result, i.e. that the audiovisual benefit effect was stronger for consonants than for vowels. This might be due to the visual features of vowel being less salient relative to consonantal visual features (Kim et al., 2009; Moradi et al., 2017).

Visual cues of prosodic speech

In addition to the visual effect on segmental speech, there are other studies focusing on the visual effect of suprasegmental speech (intonation, stress, lexical tone). Suprasegmental information conveys meanings, attitudes and emotions by changing pitch, intensity, duration and vowel space on the lexical or phrasal levels. Intonation variation can define a sentence as a statement or a question. For example, in English, statements have stable falling F0, while (yes/no) questions usually have rising F0 on the final word (Cvejic et al., 2010; Scarborough et al., 2009). Lexical stress and phrase stress/ prosodic focus emphasise novel or important information in a word/ phrase in a sentence (Cvejic et al., 2010; Scarborough et al., 2009). Acoustically, the word/phrase that is emphasised within a sentence usually has a higher pitch, stronger intensity and longer duration (Krahmer & Swerts, 2001). As mentioned in Section 1.2, lexical tones distinguish the lexical or grammatical meaning of a word by changing F0, but they only function as phonemic information and often come with minimal pairs or other sets (e.g. minimal quadruplet tones in Mandarin) (Hayes, 2009). Although prosodic speech (stress and intonation) and lexical tones are functionally different, given that they have the same acoustic perceptual cues, visual cues of prosodic speech are more comparable with the visual cues of lexical tones.

The key acoustic features for perceiving prosodic difference are F0, duration, intensity and vowel quality (Ladd, 1996). As described previously, F0 variation is associated with the movement between the laryngeal muscles and the thyroid cartilage, which is difficult to see from the mouth region (lips, teeth and tongue). However, some studies have found that movements beyond the mouth region, such as eyebrow movement (Cavé et al., 1996) and rigid head movement (Burnham et al., 2006; Yehia et al., 2002), correlate with F0 variation to some extent. Swerts and Krahmer (2008) found that prosodic prominence in Dutch was more easily perceived when adding the upper part of the facial expression than the bottom part. Lansing and McConkie (1999) found that intonation judgement greatly decreased without presenting the upper face. Cvejic et al. (2010) reported that prosodic focus discrimination remained high accuracy in both video-only and audiovisual conditions in which only the top half of the speaker's head and face (in either full texture or outline) was presented. Their results support head movement as a visual cue that contributes to prosodic perception.

In contrast with eyebrow and head movements, some studies have reported that prosodic visual cues are located in the mouth region (including lips, jaw and chin), which might be visually correlates to acoustic intensity or duration, since stressed syllables/ words with greater intensity have larger, longer and faster lip and jaw movements (Krahmer & Swerts, 2001; Scarborough et al., 2009). Dohen and Lœvenbruck (2005, 2006) recorded and analysed facial data from the production of contrastive focuses in French, and the authors reported that lip protrusion was the visual movement correlating most with contrastive focus. Dohen and Lœvenbruck (2009) investigated detection scores and reaction times of audiovisual prosodic focus in whisper, and their results showed a clear improvement when adding speakers' lower face articulation. Scarborough et al. (2009) tested the perception of English lexical and phrasal stress location (i.e. which word was stressed/ emphasised) in a video-only condition without presenting auditory signals, and the results showed visual perception was better than a chance level. They analysed visual correlates from production stimuli and found that Chin Opening Displacement (chin downward gesture) was the most important cue for visual stress perception.

The visual information from visual segmental and visual prosodic speech described above largely focuses on the various kinds of phonetic information that contribute to audiovisual speech perception. Visual cues, such as the place of articulation of consonants, the lip-roundedness/ openness of vowels and the F0 of prosody, complement acoustic cues during perception. However, other than phonetic visual information, increasing numbers of studies show that there is another type of visual information that is not phonetic-specific in audiovisual speech perception. Non-phonetic visual information could be related to the temporal aspect of visual speech, relative to auditory speech.

1.3.2.2 Non-phonetic information provided by visual speech

In addition to phonetic information, visual speech can serve as a temporal indicator for the onset and offset time and the dynamic change of the amplitude envelope of the concurrent acoustic signal. Especially in noisy environments where the auditory signal is compromised, available visual speech can predict when coming speech sounds start and end, so that it directs perceivers' attention to auditory speech and enhances perceivers' sensitivity to detecting auditory signals (Bernstein et al., 2004; Eskelund et al., 2011; Grant & Seitz, 2000; Kim & Davis, 2004; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004). This type of audiovisual benefit effect is independent of whether visual speech can provide content for phonetic identity (i.e. non-speech-specific), it may therefore reflect more general perceptual processing.

Schwartz et al. (2004) demonstrated a case where the audiovisual benefit effect did not rely on lip-read cues from visual speech. In their first experiment, audiovisual syllables in identification were similar in articulatory gestures (e.g. voiced and unvoiced plosives: /ty/ vs. /dy/). The results showed that identification was significantly better in the audiovisual modality compared to the auditory-only modality. In a second experiment, in order to remove residual lip-read information, visual input was replaced by fixed-lip gestures about 100–240 ms ahead of auditory signals, so that all syllables had the same visual speech. Again, the results showed a significant audiovisual benefit effect. Because visual speech was the same across syllables, it was impossible for lip-read information from visual speech to contribute an audiovisual benefit. The authors proposed that the effect was possibly due to an increase in the sensitivity to acoustic signals contributed by the temporal cue of the onset time of articulatory movement.

A few studies have directly investigated the audiovisual benefit in detection tasks by testing whether auditory detection can be improved (i.e. whether auditory sensitivity can be increased) by adding visual speech along with auditory speech. Grant and Seitz (2000) employed a two-interval two-alternative force-choice task to examine the detection threshold of spoken sentences in noise. They found that the detection threshold was lower when adding matched visual speech. They also found a correlation between visual articulation movement (changes in the area of lip opening) and the speech amplitude envelope, and when the correlation was higher, the speech detection threshold was lower. The authors explained that presenting visual speech movement provides information about when and where to expect signal energy, hence preventing the target speech from interference by noise (reducing uncertainty) (also see Grant, 2001; Kim & Davis, 2004).

In addition to sentence detection, the audiovisual benefit effect has also been found in syllable detection (Bernstein et al., 2004; Tye-Murray et al., 2011). Furthermore, these studies show that the audiovisual benefit effect to the detection threshold does not require phonetic information provided by visual speech either. For example, Bernstein et al. (2004)

compared the auditory detection threshold in a natural visual articulation condition and a synthesised visual stimuli condition (static and dynamic rectangle bar, Lissajous figure) (also see Tye-Murray et al., 2011). Synthesized visual stimuli did not contain any speech-specific cues (e.g. place of articulation), but they temporally aligned with auditory signals. The results showed that natural audiovisual stimuli had the lowest detection threshold and the other synthesised audiovisual stimuli also led to a lower detection threshold relative to the auditory-only condition. The authors explained that the improved detection threshold was associated with the temporal relationship between preliminary visual speech gestures and the auditory signal at a lower level of processing. Their findings further suggest that the audiovisual benefit effect does not necessarily require higher-level processing of visual speech features.

Kim and Davis (2014) compared the audiovisual benefit effects of two types of visual information: phonetic information (presenting full-face gestures) and timing information (presenting face gestures with the mouth area covered). For the timing condition, although the mouth area was covered, timing information could still be seen from face movements around the mouth and jaw. The results showed an audiovisual benefit effect in both conditions; specifically, timing information improved the response time of auditory detection. The researchers suggested that visual speech temporally prepares the upcoming auditory signal, therefore increasing the sensitivity of the auditory cortex.

In addition to evidence for the audiovisual benefit effect, studies of the McGurk effect have also proved the existence of non-phonetic visual information in audiovisual speech integration. Soto-Faraco and Alsius (2007) employed a temporal order task in which participants needed to judge the syllables they heard and the order of auditory and visual inputs. They found that participants could still perceive /bda/ even when auditory /da/ was correctly detected as the preceding signal against visual /ba/ in the audio lead condition. Soto-Faraco and Alsius (2009) then confirmed that perceivers were able to detect audiovisual asynchrony yet also experienced the McGurk illusion (also see Munhall et al., 1996). The authors proposed that detecting the temporal order of auditory and visual signals and audiovisual phonetic integration may be two different processes. In other words, audiovisual integration might comprise an audiovisual temporal process and an audiovisual phonetic process.

As mentioned earlier, there is evidence showing that the visual speaking rate/ visual duration can trigger the McGurk effect without providing actual phonetic information (Green & Miller, 1985). The visual rate is defined by the duration of visual speech. Visual speech longer in duration is perceived as a slower visual rate, and vice versa. Green and Miller (1985) employed fast and slow visual articulations (/bi/ or /pi/) dubbed into an auditory /bi-pi/ (voiced and voiceless) continuum spoken at intermediate speed. The results showed a shift in the categorical voicing boundary along the /bi/-/pi/ continuum due to the change in the visual rate.

As can be seen in the above literature, visual speech can influence auditory speech through different visual cues (phonetic or temporal) in the audiovisual perception of consonants, vowels and prosodic speech. Especially considering the evidence for prosodic speech, movements of eyebrows, head and mouth seem to provide useful cues related to the acoustic features of F0, duration and intensity, that are also important for lexical tones. Moreover, the temporal aspect of visual cue indicates that visual influence in speech does not necessarily require the speech to be salient. This suggests that lexical tones, as less visualised speech units, are also likely to be affected by visual speech in terms of perception.

1.4 Audiovisual perception of lexical tones

1.4.1 Visual information facilitates auditory tone perception

Compared to audiovisual perception research on the segmental and prosodic aspect of speech, there is only a limited number of studies about the audiovisual perception of lexical tones in the literature. This may be because the articulation of lexical tones largely depends on laryngeal movement, which is less visible to perceivers, therefore the perception of audiovisual lexical tones is considered to involve auditory dominant processing (Liu et al., 2014; Sekiyama, 1997). However, visual features being less visible does not necessarily suggest that visual cues for lexical tones do not exist. Several studies have provided evidence that visual speech information can improve auditory lexical tone perception (Burnham et al., 2014; Burnham et al., 2001; Burnham et al., 2006; Burnham et al., 2011; Chen & Massaro, 2008; Mixdorff et al., 2005a; Mixdorff et al., 2005b; Mixdorff et al., 2006; Smith & Burnham, 2012).

The first study of audiovisual lexical tones was conducted by Burnham et al. (2000). They tested the identification rate of Cantonese lexical tones in three modalities: auditory-only, visual-only and audiovisual modalities, and they found that the identification in the video-alone condition was much worse than in the auditory-only and audiovisual conditions, but it was slightly yet significantly higher than a chance level. Their findings suggest that visual cues for lexical tones (at least in Cantonese) do exist, although the effect is rather weak. However, they failed to find an audiovisual benefit effect, i.e. the identification of audiovisual lexical tones was no better than that of auditory lexical tones. Burnham et al. (2001) later tested Cantonese lexical tone discrimination with non-native perceivers (Thai speakers and English speakers). They found that audiovisual discrimination was better than auditory-only discrimination when audio-masking (multi-talker babble noise) was added. Their findings indicate that visual speech information of lexical tones can have an influence on auditory tone perception (at least for tone-experienced speakers), and also suggest that the noise condition is important to induce an observable audiovisual benefit effect in behavioural studies. When the auditory signal is impaired by noise, the perceivers are forced to give more weighting to the visual speech signal; and furthermore, the ceiling effect (the performance in both audiovisual and auditory-alone conditions is nearly 100%), which hides potential differences across conditions when the auditory signal is intact, can be avoided.

In terms of audiovisual perception of Mandarin lexical tones, Mixdorff et al. (2005b) compared audiovisual and auditory Mandarin lexical tones in an identification task with native speakers. They set up multiple auditory degradation conditions [e.g. amplitude modification and babble noise at signal-noise-ratio (SNR) levels of -3 dB, -6 dB, -9 dB, -12 dB]. An audiovisual benefit effect was found in the amplitude modified condition and SNR -9 dB and -12 dB conditions, and there was a tendency for the audiovisual benefit effect to be stronger with decreasing SNR. Similar results were found for Thai (Mixdorff et al., 2005a) and Vietnamese lexical tones (Mixdorff et al., 2006), as well. Smith and Burnham (2012) investigated cross-language (Mandarin and English) audiovisual perception of Mandarin lexical tones with a discrimination task in a condition where the auditory signal had cochlea implant simulation. An audiovisual benefit effect was found in both language groups in the cochlear implant stimulation condition, and the lexical tone discrimination in visual-only (lip-reading) was above a chance level in both groups. These studies support visual cues being available for Mandarin lexical tones and that they can be captured to improve auditory tone perception.

1.4.2 Audiovisual benefit effect for individual lexical tones

Furthermore, the degree of audiovisual benefit effect is not equal across individual tones in identification or discrimination. Smith and Burnham (2012) found that tone contrasts with a dipping tone (T3) tended to have a stronger effect. The authors explained that acoustic T3 in Mandarin had a more complex and dramatic F0 movement that could easily be visualised as a distinctive visual cue. Interestingly, in the report of Mixdorff et al. (2005b), T3 and T4 had better identification rates in a devoice condition (no F0 information), which also indicated T3 and T4 could be more visually distinctive compared to other tones. The various degree of the audiovisual benefit effect for individual tones does not only appear in Mandarin lexical tone perception, it is also found in Thai tone discrimination. Burnham et al. (2011) found that both Mandarin and Thai speakers showed that they experienced a greater audiovisual benefit effect when discriminating contour-contour tone contrast rather than level-level or level-contour tone contrast in Thai lexical tones. There are only two contour tones in the Thai tone inventory:[1] falling tone (Tone 231) and rising tone (Tone 315) (the other three are level tones). These two tones have opposite F0 movements, which could have more readily visual correlates.

The unequal magnitude of the audiovisual benefit across individual Mandarin tones indicates that some tones, such as T3, could be more visually salient than others. The

[1] The five Thai lexical tones comprise: 1. Mid-level tone T33; 2. Low level tone T11; 3. Falling tone T231; 4. High level tone T55 and 5. Rising tone T315. T231 and T315 are contour tones, while the others are level tones.

distinctive acoustic features of these ‘more visible’ tones may be transferred as potential visual cues. Returning to the study of Smith and Burnham (2012), they ran a subsequent regression analysis between the acoustic features of the stimuli and discrimination performance, and the results showed that the strongest correlation was with duration and intensity. The authors suggest that lexical tones that are acoustically distinctive in duration and intensity could be more visually salient. As described in Section 1.2, for Mandarin lexical tones in isolation, the longest tone in duration is the dipping tone (T3) due to its complex F0 movement, while the shortest tone is the falling tone (T4). T3 has rather weak intensity while T4 usually has the strongest intensity. T3-T4 tone contrast is the most readily distinguishable pair due to their contrastive features. However, in the visual-only results in Smith and Burnham (2012), T3-T4 contrast was the worst one among other contrasts. This result seems to contradict the conclusion that visual cue of lexical tones is duration-based. Therefore, more experiments are required to address this issue further.

Although Smith and Burnham (2012) support visual speech information facilitating auditory tone perception when adding cochlear implant stimulation, Liu et al. (2014) reported that they failed to find an audiovisual benefit effect in real cochlear implant Mandarin-speaking patients. They investigated the recognition of Mandarin phonemes and lexical tones in audiovisual and auditory modalities by cochlear implant patients and normal hearing native speaker adults. No audiovisual benefit effect was found for lexical tones in either group, though it was found in phonemes in both groups. They concluded that lexical tone perception was largely auditory-signal-dependent.

1.4.3 Possible visual cues for lexical tones

Does the negative result for an audiovisual benefit effect for lexical tones suggest the visual cues for lexical tones are very limited? Chen and Massaro (2008) suggested that this is due to Mandarin speakers underusing visual cues during perception. In their study, Mandarin speakers were trained to pay attention to possible visual cues from the movements in lexical tone production, such as head/chin movement, neck muscle bulge and durational differences in tones reflected in mouth movement. After training, their visual recognition of Mandarin lexical tones significantly improved. Smith and Burnham (2012) pointed out that Mandarin speakers underused the available visual cues for lexical tones compared to English speakers, because they found that tone inexperienced (English) speakers outperformed tone native (Mandarin) speakers in visual-only tone discrimination.

Some of the studies above proposed that possible visual cues for Mandarin lexical tones are associated with head, chin or mouth movements (e.g. Chen & Massaro, 2008; Smith & Burnham, 2012), but no perceptual tasks have directly tested these possible visual cues for lexical tones; as of now, what exact visual cues participate in audiovisual lexical tone perception is unclear. In parallel with perception studies, the analysis of lexical tone production gives a hint as to which motions have stronger lexical tone visual correlates. In Burnham et al. (2006), a motion capture technique was used to track kinetic facial motion data for Cantonese lexical tone production. They found that the correlation between visual movement and acoustic F0 involved rigid head movements (global head movements), suggesting rigid head movements provide important visual cues for lexical tones in Cantonese. This result is consistent with the findings for visual prosodic speech reported by Cvejic et al. (2010), where top of the head movement provided enough visual cues for accurate prosodic judgement. However, in a series of audiovisual lexical tone studies (Mixdorff et al., 2005a; Mixdorff et al., 2005b; Mixdorff et al., 2006), video clips of lexical tone production only presented speakers' lower faces (from below the eyes to the upper neck) to perceivers, and all the studies (on Mandarin, Thai and Vietnamese lexical tones) have found an audiovisual benefit effect in noise. Based on their results, crucial visual cues may also largely come from mouth, chin or jaw movements, rather than head movement. Acoustically, in addition to F0 movement, Mandarin lexical tones are systematically contrastive in duration and amplitude, which are secondary but crucial acoustic cues especially when F0 information is not available (Liu & Samuel, 2004). When articulating lexical tones, the mouth opening time from onset to offset is consistent with the durational difference across individual lexical tones as well, therefore visual lexical tones are able to provide duration cues in the process. The degree of mouth opening and closing (or jaw movement) also informs the intensity/ loudness of speech sound (Grant & Seitz, 2000; Kim & Davis, 2001; Kim & Davis, 2004), which could be a useful cue to the intensity of lexical tones.

These potential visual cues for lexical tones function as phonetic visual cues, like the visual place of articulation for consonants or lip roundedness/ openness for vowels. Visual cues complement auditory speech for phonetic/ tonetic information, such as F0, duration and amplitude, hence contributing to audiovisual speech perception. However, evidence reveals that audiovisual perception in speech does not completely depend on phonetic information from visual speech (see Section 1.3.2.2). Visual speech can provide non-phonetic information (timing cue) for a temporal relationship in audiovisual speech dynamics, which could be more important for speech that is less visually salient, such as lexical tones.

1.4.4 Non-phonetic visual information for lexical tones

The studies reviewed previously show that audiovisual integration processing does not necessarily involve the phonetic decoding of visual speech. The temporal aspect of visual speech input can influence auditory speech perception. Despite the process of temporal information audiovisual interaction not being fully understood, this processing is believed to be non-phonetic or pre-phonetic processing for audiovisual speech processing, which is very likely to take place in the early stages of stimulus processing (Kim & Davis, 2014; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Soto-Faraco & Alsius, 2009).

In terms of lexical tones, the proposed visual cues (e.g. head movement) mentioned have been largely associated with the acoustic characteristics of lexical tones, such as F0, amplitude and duration. There are few audiovisual studies concerning the role that the temporal aspect of visual cues plays in audiovisual lexical tone processing; therefore, the question of whether visual timing cues also affect the audiovisual integrating process of lexical tones remains unclear. However, according to the properties of timing cues, the temporal effect does not involve phonetic-decoding processing and it may be a more generic processing that reflects the temporal relationship between visual and auditory signals in natural audiovisual events. As mentioned previously in Section 1.2.1.2, a lexical tone has to be realised through a syllable which usually contains at least one vowel in Mandarin, such as /ǎ/ or /bǎ/ (syllable /a/ or /ba/ in T3). Even though actual visual tone articulatory movement does not precede the auditory tone signal per se, the visual vowel/

initial consonant movement of the syllable occurs ahead of the auditory vowel/ consonant and tone signals; thus, mouth/ lip movements are able to predict the ensuing auditory signals, including auditory lexical tones. That is, auditory tone detection can still be improved by preceding mouth/ lip movement, even if the movement is not tone-specific. Therefore, it is reasonable to assume that the timing of visual cues is very likely to play an important role in the early stages of audiovisual lexical tone processing as well.

1.4.5 The extent of using visual speech information by Mandarin speakers

In addition to the availability of visual speech information, another important factor in audiovisual speech perception is whether perceivers have the ability to use visual speech information in audiovisual speech. Even though it has been argued that audiovisual speech processing appears to follow a universal principle across languages (Chen & Massaro, 2004; Massaro, 1998), there are a certain number of studies suggesting that the extent of using visual speech information or audiovisual integration processing varies due to various factors, such as language background, culture (Sekiyama, 1997; Sekiyama & Tohkura, 1993), age (Chen & Hazan, 2009; Sekiyama & Burnham, 2008), speakers and even individual differences (Chen & Hazan, 2009; Hazan et al., 2010). For Mandarin speakers, whether they have the same extent of using visual speech information as speakers of other languages (e.g. English speakers) is still being debated. Most studies have employed the McGurk paradigm to examine the weight of visual processing in audiovisual speech perception across language groups. Some studies have reported that Mandarin speakers had a lower McGurk effect, therefore making less use of visual speech information (Sekiyama, 1997), whereas other studies have reported Mandarin speakers having the same degree of the McGurk effect as their English counterparts (Chen & Hazan, 2009; Magnotti et al., 2015). The main reasons to account for the lower use of visual information lie in lexical characteristics (reliance on auditory), cultural conventions (face-avoidance) and possibly exposure time to foreign languages.

Sekiyama (1997) compared Mandarin speakers with American English speakers in terms of the strength of the McGurk effect, and the results revealed that Mandarin speakers demonstrated a much weaker McGurk effect, suggesting they tended to use less visual information. The author also pointed out that Mandarin speakers relied heavily on auditory information as they had a lower weighting of visual information, even under the influence of the 'non-native speaker effect'. The 'non-native speaker effect' refers to the phenomenon that perceivers use more visual information when the speakers who deliver speech are non-native (Sekiyama et al., 2003; Sekiyama & Tohkura, 1993). That is, on seeing Japanese and English speakers' facial gestures in a task, Mandarin speakers are supposed to give a higher weighting to visual information. However, the results did not show an increase in the use of visual information.

This lower visual weighting is also thought to be associated with the characteristics of lexical tones in Mandarin. Sekiyama (1997) explained that lexical tones are more informative in auditory speech than in visual speech, consequently this fosters Mandarin speakers to predominantly use auditory information in speech perception. This account basically suggests that tonal speakers are likely to give lower weighting to visual information in audiovisual speech perception. Burnham and Lau (1998) also reported that Cantonese had a strong reliance on auditory information when perceiving McGurk syllables compared to an English group. Although there is no direct evidence for the 'tonal speakers hypothesis' as proposed by Sekiyama (1997), their findings also seem to support the notion that tonal speakers tend to give a lower weighting to visual information. Another account for the weak McGurk effect among Mandarin speakers may be that Asian culture discourages eyes gazing at speakers during conversations (Sekiyama, 1997). Mandarin speakers may have a cultural convention of face avoidance, whereby people avoid directly watching speakers' faces, especially when they speak to someone who has higher social status (Sekiyama, 1997). Additionally, in Sekiyama (1997), the data also showed that the degree of using visual information correlated with the time that Mandarin speakers were exposed to Japan. Speakers who stayed for a short time tended to have a weaker McGurk effect than those who stayed longer. Therefore, the author suggested that monolingual Mandarin speakers may increase their use of visual information when they are exposed to a foreign language environment.

However, some of the interpretations above for the weaker McGurk effect among Mandarin speakers have their limitations. First, the 'tonal hypothesis' account for the weaker visual effect is not compatible with empirical findings, because a number of studies have provided evidence that lexical tones can be visually distinguished by native speakers to some extent (Burnham et al., 2000; Burnham et al., 2001; Burnham et al., 2011; Mixdorff et al., 2005a; Mixdorff et al., 2005b; Mixdorff et al., 2006; Smith & Burnham, 2012). Furthermore, Mandarin speakers do not rely only on lexical tones but also on segments. The phonetic inventory in Mandarin is as rich as in English; thus, visual information can be largely used for distinguishing consonants and vowels in Mandarin. As for the cultural convention of face avoidance, it may be true for the older generation, but this may have changed among the young due to the heavy influence of popular culture (i.e. mainly Western culture) in recent decades. Additionally, the study of Sekiyama (1997) only investigated a small number of participants (14 Mandarin speakers). Given the great variation in individual difference in the McGurk effect, their speculations need to be further verified by testing more participants.

Some studies have re-analysed the McGurk effect in Mandarin speakers, and the results were different from what Sekiyama (1997) reported. Chen and Hazan (2009) compared the degree of the visual effect in Mandarin speakers and English speakers, and no significant difference was found between the two groups. Later, more studies reported that Mandarin speakers had the same degree of visual effect across conditions (Hazan et al., 2010; Magnotti et al., 2015). Magnotti et al. (2015) tested a large number of Mandarin speakers (*N* = 162) and English speakers (*N* = 145). Their results revealed the both groups had a similar pattern in McGurk syllable identification, further suggesting Mandarin speakers have the same extent of using visual information in audiovisual speech perception as do their English counterparts. It is possible that increasing second language (English) exposure enhances Mandarin speakers' ability to use visual cues. In Chan & Hazan (2009), for example, although the Mandarin participants had never lived in a foreign country for more than six months, English as a second language had been taught since primary school. The familiarity with English language might be related to the increase in weighting of visual information for Mandarin speakers. Generally, these findings strongly support Mandarin speakers being able to use visual information in audiovisual speech perception as much as English speakers.

1.4.6 Behavioural experiments in audiovisual lexical tone perception

Up to this point, what is known from the relevant literature is that, although lexical tones are difficult to lip-read, presenting visual lexical tones with auditory tone signals facilitates lexical tone perception in noise, suggesting that visual cues are in fact available in Mandarin lexical tones. There is no strong evidence for what specific visual cues contribute to perception. The potential visual cues that are proposed could be related to head movements, neck muscle tension (F0 visual correlates) and mouth/ lip movement (duration/ intensity visual correlates). Besides that, there might exist non-phonetic visual cues, such as timing cues that, preceding visual movements, predict the timing of the auditory signal. Additionally, Mandarin native speakers appear to be capable of making use of available visual cues in either audiovisual lexical tone perception or the McGurk effect for consonants.

This dissertation aims to resolve some questions about audiovisual lexical tone perception going beyond existing findings. For example, whether the audiovisual benefit effect of lexical tones is due to distinctive visual speech-specific information or visual non-speech information has not been explored; whether the time course of audiovisual lexical tone integration processing is the same as segmental speech; or whether incongruent audiovisual lexical tones can induce the McGurk effect as segmental speech. These questions will be answered through the experiments conducted in Chapters 2–4, in which both behavioural and electrophysiological methods are employed. The following section will discuss behavioural experiments investigating the audiovisual benefit effect and the McGurk effect of lexical tones, and electrophysiological experiments including relevant studies will be covered in a separate section (Section 1.5).

1.4.6.1 Audiovisual benefit effect in discrimination and identification

First, in Experiments 1–4 in Chapter 2, two behavioural paradigms: the identification and discrimination of Mandarin lexical tone perception, are employed to test the audiovisual benefit effect in Mandarin lexical tones. In previous audiovisual lexical tone research, these two paradigms have been applied in several studies (Burnham et al., 2001; Burnham et al., 2011; Mixdorff et al., 2005b; Smith & Burnham, 2012). They are widely used for probing speech perceptual performance in either unimodal or bimodal stimuli. In current experiments, the performance of audiovisual lexical tones and that of auditory lexical tones are compared in identification and discrimination, and the lip-reading performance of lexical tones in a visual-only condition is also measured. The audiovisual benefit effect of lexical tones can be verified when audiovisual lexical tones are better identified or discriminated than auditory lexical tones. Tone lip-reading performance directly links to whether any tone-specific visual features can be extracted for tone perception. Furthermore, if the distinctiveness of the visual cue varies across four lexical tones, the strength of the audiovisual benefit effect of lexical tones may differ across individual tones as well. Lexical tones that are visually distinctive could benefit more from visual speech input; therefore, these tones would achieve a stronger audiovisual benefit effect, and they should also be easier to lip-read than other tones in visual-only condition. The variation in the audiovisual benefit effect across different tones could reflect what potential phonetic/ tonetic-specific visual cues facilitate lexical tone perception by looking into those tones' acoustic properties. For example, in Mixdorff et al. (2005b), T3 was highlighted among other tones for its better audiovisual benefit effect and higher lip-reading rate. Hence, the distinctive duration feature of T3 was inferred to be a potentially crucial visual cue.

In terms of experimental paradigms, the task per se can be one of factors of audiovisual effect. The degree of the audiovisual benefit effect of lexical tones may be different in two tasks because identification and discrimination tasks engage different cognitive processes during speech perception. In behavioural audiovisual speech studies, these two paradigms are frequently used to probe the extent to which visual speech facilitates auditory perception. They are believed to differ in the levels of cognitive processing (Aslin & Smith, 1988; Lalonde & Holt, 2015; Lalonde & Holt, 2016). In a discrimination task, perceivers distinguish given alternatives based on their salient differences, therefore the task often requires a representation of the physical properties of the stimuli but does not necessarily require a phonetic or lexical representation of speech input (Aslin & Smith, 1988). In the discrimination of audiovisual stimuli, task responses may largely reflect low level processing (acoustic or visual physical features). Yet, the duration of inter-stimuli intervals

(ISI) in a discrimination task can affect access to the processing level. van Hessen and Schouten (1992) reported that discrimination performance based on phoneme labelling increased when the ISI was longer. It also been proposed that a discrimination task can involve a higher level processing if the ISI lasts for a longer time (e.g. 1,500 ms) (Werker & Tees, 1984). On the other hand, the identification paradigm relates to the ability to label speech received, and it often requires perceivers to map the target speech to phonetic or lexical representation in long-term memory (Aslin & Smith, 1988). Compared to discrimination tasks, the responses driven by identification tasks reflect a late stage of processing, therefore identification tasks are more sensitive to phonetic visual information (Kim & Davis, 2014).

Previous studies on Mandarin audiovisual lexical tones have shown that an audiovisual benefit effect was found in both task paradigms (Mixdorff et al., 2005b; Smith & Burnham, 2012). However, these two paradigms differ in the degree to which perceivers get access to visual phonetic or lexical representation, so they could be useful tools to explore whether the audiovisual lexical tone benefit effect results from low level processing of physical features or from higher level processing of phonological representation, or both. The degree of audiovisual lexical tone benefit might be larger in an identification task than in a discrimination task if visual lexical tones encourage additional process (e.g. top-down processing) at a higher level. Because lexical tones are used to distinguish meaning by changing pitch level/ movement, syllables with different lexical tones often vary in meaning. Therefore, it is probable that lexical tone perception has a top-down influence. Conversely, if visual lexical tones only provide physical feature information (e.g. visual salient features), the audiovisual benefit might be similar in the two tasks.

1.4.6.2 McGurk effect for lexical tones

From the literature, we know that congruent visual information facilitates lexical tone perception, but what about the effect of incongruent visual information on lexical tone perception? Does incongruent visual information bias auditory tone perception as in the traditional McGurk effect? To my knowledge, the audiovisual perception of incongruent lexical tones or the McGurk effect of lexical tones has not been studied. The existing research on the McGurk effect in tonal speakers (especially Mandarin speakers) (Chen & Hazan, 2009; Magnotti et al., 2015; Sekiyama, 1997) only focused on addressing the issue of whether language background/ experience can influence the magnitude of the McGurk effect, instead of whether mismatched visual lexical tones can bias auditory tone perception as in the classic McGurk illusion found for consonants. In those studies, the incongruent audiovisual speech stimuli used were nonsense consonant-vowel (CV) syllables (e.g. auditory /ba/ pairing with visual /ga/), just like the stimuli in the majority of McGurk studies. That is, those studies of the McGurk effect in Mandarin speakers actually investigated the McGurk effect in audiovisual consonants but not the McGurk effect in audiovisual lexical tones. Although those studies have provided strong evidence that Mandarin speakers have a similar McGurk effect (in consonants) to the one among English speakers (Chen & Hazan, 2009; Magnotti et al., 2015), whether Mandarin speakers can also experience the McGurk effect through incongruent audiovisual lexical tones remains unexplored.

As stated previously, the articulation of lexical tones is determined by vocal fold vibration in the larynx, which suggests that lexical tones are difficult to lip-read. Even though visual-only lexical tones can be perceived, the discriminated or identification rate is still very low (slightly above chance level) (see Chen & Massaro, 2008; Mixdorff et al., 2005b; Smith & Burnham, 2012). In an incongruent audiovisual lexical tone syllable, the mismatch between visual and auditory tone can be even subtler. If generating the McGurk effect requires a greater discrepancy between auditory and visual speech in mouth/lip movement (e.g. place of articulation or roundedness/ openness of the mouth), then incongruent audiovisual tone is very likely to be perceived the same as congruent tone (i.e. incongruent visual information does not change the auditory tone).

However, as mentioned earlier in Section 1.3.2.2, auditory speech perception is not only biased by mismatched mouth/ lip articulatory movement, it can also be changed by the speaking rate of visual speech. Green and Miller (1985) reported that when the fast or slow visual syllable /bi/ or /pi/ paired with the auditory continuum /bi/-/pi/ at a moderate rate, the visual rate (speaking rate) changed the VOT perception of auditory speech, thus shifting the categorical boundary between /bi/ and /pi/. In order to make sure the dubbed visual stimuli did not contain phonetic information, only speed information, the authors also pretested the discrimination between /bi/ and /pi/ in the visual alone condition and the result was negative. In the auditory perception of stop consonants, VOT is a critical acoustic cue for distinguishing voiced and voiceless features (Reetz & Jongman, 2009), and changing the auditory duration can influence VOT perception. By adding articulatory movement at a fast or slow rate, the original auditory duration or speaking rate can be changed accordingly, hence influencing auditory speech phonetically. That is, the visual rate effect is actually the result of audiovisual integration, specifically in terms of speed. Incongruent audiovisual speech at a speaking rate in their studies primarily suggests that this type of audiovisual integration process is non-phonetic because it reflects that the mismatched visual rate modifies the perception of duration of auditory speech.

An interesting question that should be raised is whether this visual rate can affect audiovisual lexical tones in Mandarin. Regarding the speaking rate of lexical tones, durations are systematically different across the four tones. As mentioned in Section 1.2.1.4, the longest tone is the dipping tone (T3), and the shortest is the falling tone (T4). According to Green and Miller (1985), duration is perceived as the speaking rate of a syllable. A longer syllable is perceived as a slower one depending on the speaking rate, and vice versa. Then, in lexical tones, a longer tone can be perceived as a slow tone, and a shorter tone as a fast tone. In incongruent audiovisual lexical tones, when a long tone is paired with a short tone, audiovisual integration can lead to illusions in tone duration. For instance, when a long visual tone (T3) pairs with a short auditory tone (T4), one might end up hearing a longer T4 compared to the original auditory tone. In auditory perception, tone duration is an important cue to identify lexical tones in addition to F0 cues. If the perception of tone duration can be changed by a mismatch in visual information, then tone identification might consequently be influenced by mismatched visual information. Therefore, through testing the perception of incongruent lexical tones, a possible McGurk effect caused by mismatched tone duration can be probed.

In the behavioural experiments mentioned above, the audiovisual benefit effect and the McGurk effect of lexical tones can be observed through perceivers' physical responses to tone tasks across different conditions. However, the behavioural response has its own limitations. For example, the cognitive process is completed when perceivers make responses; consequently, the online processing of audiovisual integration is impossible to observe. To understand more about audiovisual integration in real time processing, electrophysiological experiments are also employed to investigate the neural mechanisms of the audiovisual benefit effect and the McGurk effect of lexical tones.

1.5 Electrophysiological evidence of audiovisual speech perception

In addition to behavioural studies with discrimination and identification paradigms, an electroencephalography (EEG) technique is adopted to investigate the audiovisual benefit effect in Chapter 3 and the McGurk effect of incongruent lexical tones in Section 4.2. This section will explain the rationale for using an EEG technique and review relevant literature on EEG or event-related potential (ERP) studies in audiovisual speech processing.

1.5.1 Advantages of event-related potential research

In recent years, EEG/ERP has been applied in studies of audiovisual speech perception, via which audiovisual speech processing unfolding over time can be more directly observed and measured. EEG refers to variations in the electrical activity recorded from the human brain. EEG signals contain neural responses that are associated with sensory, cognitive and motor events. By averaging time-locked EEG responses according to specific experimental conditions, responses can be extracted from EEG signals (Luck, 2005). The biggest advantage of ERP responses is their excellent temporal resolution. ERP can capture real-time brain activities in milliseconds over temporal stages without the requirement for a manual response. For research on audiovisual speech integration, the time stage of audiovisual speech signal integration has always been a debatable issue, both empirically and theoretically. ERP responses can signify early stages where auditory and visual sensory signals integrate before they are analysed as phonetic information, and later stags where auditory and visual information integration occurs after they are processed as phonetic information (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007).

Additionally, although behavioural discrimination and identification tasks can engage different levels of stimulus processing, it is difficult to completely avoid involving multiple effects (e.g. decision-making) during perception; as a result, it is difficult to determine specific effects from behavioural responses. It is comparatively easier to investigate levels of processes separately according to the components evoked in certain time windows. For example, the auditory N1-P2 complex is often considered to be an obligatory exogenous response at an early stage (at about 100–200 ms) that represents the physiological detection of auditory stimuli (Alain & Tremblay, 2007). Furthermore, ERP responses are independent of overt responses from participants. In behavioural studies (e.g. audiovisual speech identification in noise), participants' performance across audiovisual and auditory conditions is very likely to approach a ceiling in a quiet environment; consequently, accuracy comparisons are unlikely to be significantly different due to the ceiling effect. ERP can measure the effect in quiet conditions regardless of the ceiling effect. In a noise condition, participants' physical responses may be biased by their conservative judgement in a same-different discrimination task. For example, perceivers might tend to respond to trials as 'different-type' only if they are very confident about their judgement (Gerrits & Schouten, 2004). An ERP response can be directly evoked without relying on physical responses. In many cases, ERP can measure the activity of target stimuli that do not need to physically react. ERP components in early latency (before about 200 ms) after stimulus onset are highly sensitive to changes in the physical properties of stimuli, such as the frequency, duration and intensity of auditory stimuli (Näätänen & Winkler, 1999) and the contrast, spatial frequency and luminance of visual stimuli (Luck, 2005). For lexical tones, visual features are less distinctive compared to the visual features of segmental speech; consequently, perceivers might have difficulty in detecting subtle visual differences as a useful cue during perception. ERP may be particularly useful for lexical tones because it is more sensitive to small variations in visual features.

1.5.2 ERP evidence for the audiovisual benefit effect in speech

Reduction in N1 and P2 components

Several ERP studies on audiovisual speech processing in early auditory components N1 and P2 evoked by audiovisual speech stimuli provide strong evidence that visual speech information facilitates auditory speech processing at an early stage of the process (Besle et al., 2004; Klucharev et al., 2003; Knowland et al., 2014; Pilling, 2009; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). The auditory components N1 and P2 are considered to be early brain activities that reflect the physical properties of an auditory stimulus before categorisation of the stimulus (Näätänen & Winkler, 1999). The N1 response usually maximises at fronto-central sites of the scalp at about 100 ms after stimulus onset. Following N1, the P2 component peaks at about 175 ms and is distributed in the central area of the scalp (Näätänen & Picton, 1986). Compared to N1 and P2 evoked by auditory-only stimuli, N1 and P2 were found to be weaker and earlier when visual speech articulatory gestures were available. This N1/P2 reduction in ERP activity was interpreted as audiovisual facilitating speech perception at an early stage of the process before phonetic information starts being decoded, whereupon the processing of visual speech information eases off and the processing of auditory signals accelerates (Besle et al., 2004; Klucharev et al., 2003; Knowland et al., 2014; Pilling, 2009; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005).

Specifically, Besle et al. (2004) first reported that ERPs evoked by audiovisual speech syllables (e.g. /pa/) decreased within the auditory time range compared to the ERP summation of auditory and visual (A+V) responses. According to the law of the superposition of electric fields, if two current sources are independent of each other, then the current produced by the two sources is the sum of the currents produced by the individual sources; therefore, it is additive. Following this logic, if the bimodal potential is not the linear sum of the individual unimodalities, then the sources interact in the brain. In other words, if the audiovisual response is not equivalent to the sum of the response of the auditory and visual modalities (AV ≠ A+V), audiovisual interaction takes place. The result of the study in Besle et al. (2004) revealed that the difference in activity of AV and A+V, which represents audiovisual integration activity, maximised bilaterally at the fronto-central electrodes between 120 ms and 190 ms after auditory stimuli. They proposed that early audiovisual integration was speech-specific and associated with phonetic pre-activation from preceding lip movement in the auditory cortex through the poly-modal area superior temporal sulcus (STS).

van Wassenhove et al. (2005) also found an N1/P2 reduction effect in audiovisual speech processing. More importantly, their results confirmed that the shortening latency was related to the phonetic feature salience (place of articulation) of visual speech. They discovered that the most visually salient /p/ (bilabial) had the greatest N1/P2 latency facilitation, while the least visually salient /k/ (velar) had the smallest latency facilitation, but N1/P2 amplitude did not have the speech-specific effect. Their findings agree with the predictability of the phonetic visual speech feature leading to an early reduction effect in latency.

Some studies have claimed that the processing of N1/P2 reduction involves non-speech-specific audiovisual processing. Knowland et al. (2014) replicated the reduction effect in ERP responses by using monosyllabic words as stimuli in both congruent and incongruent (e.g. the auditory word *lake* paired with the visual word *rose*) conditions. They also found the reduction in amplitude was congruent-independent but the reduction in latency was sensitive to the congruency of audiovisual syllables. They claimed that the reduction in N1/P2 amplitude and latency reflects two separate processes. The N1/P2 effect of amplitude could represent competition between auditory and visual speech, so as to evaluate the consistency of cross-modality inputs. In contrast, the effect of latency could indicate that visual speech predicts the phonetic identity and timing of ensuing auditory signals.

Klucharev et al. (2003) reported a reduction effect in vowels. They found the reduction effect of amplitude maximising as early as 85 ms after auditory onset, regardless of the congruency of the stimuli; however, the difference response between congruent and incongruent stimuli started later at 155 ms. The authors proposed that the earlier processing was non-phonetic and reflected a temporal-spatial property of audiovisual integration as it is congruence-independent, while the later processing was phonetic-related audiovisual integration.

Setekelenburg and Vroomen (2007) also pointed out that the reduction effect in an early time course (auditory N1) was possibly non-phonetic. They investigated that early reduction effect with speech (e.g. /fu/), including congruent and incongruent speech, and non-speech stimuli (e.g. sawing wood, tearing paper). They found the N1/P2 reduction effect in both types of audiovisual events, but it was stronger in non-speech stimuli. However, when the preceding visual movement was removed in non-speech stimuli, the reduction effect disappeared. In the speech condition, the N1 reduction in both amplitude and latency was not influenced by the congruency of audiovisual speech, but the P2 reduction in amplitude was larger in incongruent than congruent audiovisual stimuli. The authors proposed that the N1 reduction effect was not due to visual signals predicting the content of the auditory signal but due to visual signals temporally preceding the auditory signal. In other words, the early audiovisual integration process reflects the temporal relation between auditory and visual inputs. The later reduction effect in P2 could suggest audiovisual integration on a phonetic or semantic level of processing.

Some other EPR studies have tested the reduction effect by controlling the degree of audiovisual integration and their findings confirmed that the N1/P2 reduction effect was caused by the audiovisual integration process, not because of other cognitive effects (e.g. attention shifting to the visual modality). For example, Miki et al. (2004) conducted a magnetoencephalographic (MEG) study and reported a negative result for the reduction in M100 (its equivalent in ERP is N1) when a still image of vowel articulation was presented simultaneously with auditory vowels. In Huhn et al.'s (2009) and Pilling's (2009) studies, they manipulated the synchrony of auditory and visual speech (e.g. auditory speech was 200 ms ahead of visual speech) in order to prevent favourable audiovisual integration, and they found the N1 amplitude reduction effect in the audiovisual condition was absent or weakened.

It is well known that speech processing, such as the processing of phonetic sounds, words and sentences, is asymmetrically dominant in the left hemisphere (Broca, 1861; Wernicke, 1874). In terms of the hemispheric lateralisation of audiovisual integration (AV vs A + V) in speech, different results have been reported, and there seem to be no consistent findings on the hemispheric lateralisation of audiovisual speech integration in the ERP studies above. Klucharev et al. (2003) reported right hemisphere dominance for audiovisual vowel interaction (A + V − AV) at 125 ms. In addition, Davis et al. (2008) discovered that the magnitude of M100 reduction (A – AV) was lateralised more to the right hemisphere. Furthermore, Besle et al. (2004) reported that the activity of [AV – (A + V)] was stronger at C3, while Setekelenburg and Vroomen (2007) did not observe the dominance of any hemisphere (central maximised). Huhn et al. (2009) found that a high-pass filter for ERPs affected the scalp distribution of different waves (AV – A – V). The activity of integration in N1 showed strong right hemisphere dominance with a 1–30 Hz band-pass filter.

According to the studies discussed above, the reduction effect in auditory components N1 and P2 for amplitude and latency should be considered as an indicator of the audiovisual integration process, in which visual speech facilitates the early auditory speech process. Most studies agree that the reduction is related to the degree of predictability of visual speech information, but whether the visual speech predicts the content (phonetic identity) or timing of the upcoming auditory speech signal remains in dispute. One possible interpretation of the reduction in amplitude and latency is that preceding lip movement (articulation) provides phonetic information and thus alleviates the processing load of auditory modality, thereby reducing the auditory amplitude. This is true for consonants and vowels whose lip movements start a few hundred milliseconds ahead of the following auditory signal; therefore, visual speech might prime auditory speech processing (Besle et al., 2004; van Wassenhove et al., 2005). Another assumption is that the early reduction effect may not be related to the phonetic decoding processing duration of audiovisual integration; instead, it may largely reflect two different processes: non-phonetic (temporal or spatial) processing in earlier integration and phonetic in later integration (Klucharev et al., 2003; Stekelenburg & Vroomen, 2007).

N1/P2 reduction effect in lexical tones

The findings for the early reduction effect discussed above were mainly based on consonants and vowels. In terms of ERP studies of audiovisual lexical tone processing, these are very limited in the literature. The process specifying how auditory and visual information integrates is practically unknown. It would be interesting to discover whether a reduction effect can be found in lexical tones as well. Moreover, lexical tones are ideal speech material to test whether early audiovisual integration involves phonetic or non-phonetic processing. In an audiovisual lexical tone syllable, the preceding mouth movement before the auditory signal starts is unlikely to be recognised as tonal. The production of Mandarin lexical tones is realised through syllables which contain at least a vowel, such as the consonant-vowel (CV) syllable /mǎ/ (/ma/ in T3), where audiovisual lexical tone syllables actually contain preceding articulatory movement and auditory sound. However, with visual articulation it is difficult to predict tone information of the upcoming auditory sound, since pitch production relies on the vocal fold rather than lip movement. Even though the visual tone duration of the time from mouth opening to mouth closure might be helpful for lip-reading lexical tones, a useful time point for cueing critical information about lexical tones occurs later in the processing (e.g. in the middle or towards the end of a lexical tone). That is, visual speech is unlikely to predict upcoming auditory tone information, but it can predict the timing of audiovisual asynchrony onset. If the early audiovisual integration process relies on the phonetic pre-activation of visual speech, it will be less likely to find a reduction effect in lexical tones or the reduction effect would much weaker compared to segmental speech. If early audiovisual integration only reflects the timing of visual information process, then a reduction effect would appear in lexical tones and might be similar to that in consonants or vowels. Therefore, testing the reduction effect of lexical tones can help to better understand the role that visual information plays in lexical tone perception; moreover, it can help to answer the question of whether the reduction effect is specific to phonetic visual information in the early stages of integration or not.

1.5.3 Mismatch Negativity of the McGurk effect

In additional to the reduction effect of lexical tones, the ERP method is also applied to investigate incongruent audiovisual lexical tone perception, specifically, whether incongruent visual information can modulate the perception of auditory lexical tones. The component used to probe incongruent audiovisual lexical tones in ERP experiments is mismatch negativity (MMN).

MMN is a negative ongoing component observed at about 160–220 ms in the early processing stage. MMN is considered to indicate the result of automatic discrimination between the sensory representation of infrequent stimuli (deviant) and the memory representation of standard stimuli (standard) in a sequence (Näätänen, 1990; Näätänen & Alho, 1997; Näätänen et al., 2007; Näätänen & Picton, 1986). Traditionally, MMN was used to study auditory event processing. It can reflect the behavioural discrimination threshold and is often used for testing categorical perception in speech (Näätänen et al., 2007). MMN is also used to explore audiovisual integration processing with incongruent audiovisual speech stimuli (particularly McGurk stimuli). A number of studies have reported that McGurk illusion syllables evoke MMN even if auditory speech input remains the same across conditions (Colin et al., 2004; Colin et al., 2002b; Möttönen et al., 2002; Saint-Amour et al., 2007; Sams et al., 1991).

Compared to the reduction effect of N1/P2 reviewed previously, an MMN approach could investigate audiovisual integration processing on a perceptual level in later time windows. The N1/P2 reduction is the result of a comparison between audiovisual and auditory responses, and it is considered to index that visual speech facilitates a corresponding auditory speech process in the early stages of processing. Although incongruent audiovisual speech (mostly fusion McGurk syllables) is also used in this paradigm, the N1/P2 reduction effect (in terms of amplitude) seems to be independent of the congruency of audiovisual speech in some reports (e.g. van Wassenhove et al., 2005). On the other hand, MMN is an effective tool to probe the extent to which incongruent visual speech modulates auditory speech perception by comparing it with the processing of congruent syllables. For incongruent audiovisual lexical tones, the McGurk effect can be difficult to observe in behavioural paradigms due to the less salient visual speech information of lexical tones. MMN could be sensitive enough to detect the weaker visual effect of lexical tone in the responses of the brain.

Since the activity of MMN in audiovisual speech integration is associated with the MMN of speech in unimodalities, before looking into more detail at MMN studies of audiovisual speech processing, the following sections will start with auditory MMN and visual MMN.

1.5.3.1 Auditory MMN

Traditional MMN is attained via an oddball paradigm, in which a ‘standard’ sound is presented frequently while a deviant sound is presented infrequently in a random order. The response of the deviant stimulus evoked a larger negative deflection compared to the response of the standard stimulus between 160–220 ms (N2 time window) in the frontal scalp area. However, if perceivers attend to the deviant stimulus, it can evoke N2b or P3 component which often overlaps with the MMN response, hence contaminating the MMN observation. To avoid the attention effect, a passive oddball paradigm is commonly used in MMN experiments, where perceivers are required to perform a task that is irrelevant to detecting the deviant stimulus (e.g. watching a silent movie) (Näätänen et al., 2007). The MMN response has been employed to probe the discrimination of acoustic sounds (e.g. in frequency, duration, intensity etc.) and the contrast of phonetic categories in a larger number of speech perception studies. The MMN response reflects the property of stimuli, for example, speech-related MMN tends to be larger in the left hemisphere of the brain (Näätänen et al., 2007).

Auditory MMN of lexical tones

In auditory perception studies of Mandarin lexical tones, the MMN approach is also used to probe the levels of lexical tone processing and the time course of these levels of processing. Lexical tones are an interesting speech material. They can be perceived as acoustic signals (pitch), especially by non-tonal speakers, and can also be perceived as speech signals phonetically or phonologically by native speakers. How lexical tones are perceived acoustically or phonologically affects MMN activation in lateralisation.

It is well known that language-related processing in the brain is more left-lateralised (Broca, 1861; Wernicke, 1874). However, some studies have demonstrated that lexical tone processing tends to be right-lateralised. Lou et al. (2006) found that, for Mandarin speakers, the MMN evoked by Mandarin lexical tones was larger in the right hemisphere, which was the opposite pattern to consonant-evoked MMN. The authors proposed a two-stage processing for lexical tones, containing low-level acoustic processing in the right hemisphere in an early time window (about 200 ms) and high-level acoustic processing with tone categorical information mapped into a semantic representation in the left hemisphere in a late time window (300–500 ms). Xi et al. (2010) compared the MMNs evoked by the contrast of between-category lexical tones (e.g. T1 vs T2) and by the contrast of within-category lexical tones (pitch variation within a lexical tone category) in Mandarin. They found that the MMN of between-category tones was generally larger relative to the response of within-category tones, and it was mainly in the left hemisphere. The time courses of the two MMNs were not different. Their results suggest that the acoustic and phonological lexical tone information is processed in different hemispheres, but in parallel, rather than across two stages.

The lateralisation of lexical tone MMN is sensitive to acoustic and phonological processing. It could also be useful to explore at what level auditory lexical tone processing is modulated by visual information. In an audiovisual lexical tone syllable, if incongruent visual information alters auditory tone perception, the brain response of the auditory tone process could be different from the response of auditory tone processes with congruent visual input, therefore activating MMN. The hemispheric dominance pattern can indicate whether the perception of auditory lexical tones is modulated on the acoustic level or on the phonetic/ phonological level.

1.5.3.2 Visual MMN

It is widely acknowledged that MMN elicitation is generally from auditory stimuli. However, analogous MMN is not impervious to visual stimuli. There is a growing number of studies claiming that visual MMN in analogous to the auditory MMN that is found by using either non-speech (e.g. colours) or speech stimuli (written words, articulatory gestures) with a passive oddball paradigm (for a review, see Czigler, 2014). In contrast to auditory MMN, visual MMN tends to be the largest in the N2 time range with a modality-specific distribution, and visual MMN requires a larger deviant distance threshold between infrequent and frequent stimuli (for a review, see Pazo-Alvarez et al., 2003). In visual speech (lip-reading), MMN is not easily evoked. In several studies, MMN driven by visual speech was not found (see Colin et al., 2004; Colin et al., 2002b; Saint-Amour et al., 2007). However, Files et al. (2013) reported that the contrast of articulation movements can elicit MMN on the phonological level. The authors found that MMN was evoked by visual speech /zha/ vs. /ta/ ('near' feature condition[2]) and /zha/ vs. /fa/ ('far' feature condition) in the posterior temporal area. MMN in the 'far' condition was larger in the left posterior temporal area, and the latency lasted longer (e.g. 200–500 ms after stimulus onset). The authors proposed that the left posterior temporal was sensitive to speech-specific features, while the right posterior temporal may be sensitive to the contrast of face gestures (non-speech feature).

1.5.3.3 MMN of audiovisual speech integration

MMN evoked by McGurk fusion

In MMN studies of audiovisual speech processing, McGurk stimuli are frequently used to examine the visual modality effect on auditory representation. Some studies have found an auditory-like MMN evoked by McGurk stimuli (Colin et al., 2004; Colin et al., 2002b; Möttönen et al., 2002; Saint-Amour et al., 2007; Sams et al., 1991) and by incongruent stimuli leading to the ventriloquist illusion (Colin et al., 2002a). These McGurk-MMN

[2] In Files et al. (2013), the 'near' feature condition refers to the perceptual distance in visual features between the consonant /zh/ vs /t/ being closer. In contrast, the 'far' feature condition indicates that the perception distance in visual features between the consonants /zh/ vs /f/ is farther.

studies provide evidence that visual speech modulates auditory speech perception at an early processing time before MMN is elicited.

Colin et al. (2002b) used the syllables /bi/ and /gi/ in three modalities (auditory-only, visual-only and audiovisual) with a passive oddball paradigm. In the audiovisual condition, A/bi/V/gi/ and A/gi/V/bi/ were used as deviant stimuli in separate sequences with congruent syllables as standard stimuli. The auditory speech input was identical across deviance and standards. The results showed that a clear MMN response was evoked at Fz in the auditory-only and audiovisual modalities, and the time window covered the early components P1, N1 and P2. Similar results were also found in Colin et al. (2004). Saint-Amour et al. (2007) tested MMN in audiovisual and visual-only modalities. They found a large MMN in the AV – V condition (visual response subtracted from audiovisual response). This MMN was evoked by the A/ba/V/va/ incongruent syllable (illusorily perceived as /va/) as a deviant stimulus compared to the /ba/ congruent syllable as a standard stimulus. MMN included three phases: 175–225 ms with a left hemisphere distribution; activity maximisation at 290 ms in the bilateral fronto-central area; 350–400 ms in the left hemisphere.

The mismatch field (MMF), which is analogous to MMN in ERP studies, is also reported in audiovisual MEG studies. Similar to MMN driven by the McGurk stimulus, Sams et al. (1991) found MMF in the left auditory cortex evoked by the McGurk syllable A/pa/V/ka/ (which is perceived as /ta/ or /ka/) presenting with the congruent syllable /pa/. Möttönen et al. (2002) also reported that MMF was evoked by both audiovisual and visual alone speech. They used both incongruent audiovisual syllable A/ipi/V/iti/ which is perceived as /iti/ and congruent syllable /iti/ as deviant stimuli presented among congruent syllable /ipi/ as standard stimuli in audiovisual condition, and they used the same stimuli (deviant /iti/ and standard /ipi/) in visual alone condition. In the audiovisual condition, both incongruent and congruent syllables elicited MMFs. In the visual alone condition, the visual speech deviant also evoked MMF. However, audiovisual MMF was earlier than visual MMF in latency, suggesting that audiovisual speech processing was faster than visual speech processing in discrimination.

Negative result for McGurk-related MMN

Kislyuk et al. (2008) showed that MMN could not be evoked when presenting incongruent syllable (A/ba/V/va/) as a deviant stimulus along with congruent syllable /va/ as a standard stimulus. With the incongruent syllable, under visual speech influence, the auditory syllable /ba/ was perceived as /va/, which is perceptually the same as the auditory incongruent syllable. When the perceptual difference in the auditory syllable between the two conditions was eliminated, the brain failed to discriminate the two, even though they were physically different. Their findings can be seen as further evidence for the McGurk MMN, indicating the visual modulation of auditory speech in the early processing stage.

Hessler et al. (2013) used an active oddball paradigm, where participants responded directly to deviant stimuli. The study compared MMN, N2b and P3 in pure tone, auditory-only speech, visual-only speech and audiovisual speech (congruent and McGurk) conditions. Possibly because of the elicitation of N2b and P3 by the attention effect caused by the task, MMN was not found in any speech conditions. Specifically, in the McGurk condition, the response of the deviant stimulus (incongruent A/pa/V/ka/) was more negative to the response of the standard stimulus (congruent /pa/) in all time windows (120–160 ms, 200–240 ms, 360–400 ms), but the negativity was in the occipital electrodes instead of the frontal electrodes where traditional auditory MMN is distributed. Despite MMN not being found, their results still support visual speech influencing auditory speech processing, at least in early time windows in the analysis.

Apart from the task paradigm, it appears that McGurk MMN can be influenced by the strength of the McGurk effect. Eskelund et al. (2015) manipulated the face of visual speech input in McGurk syllables and they found that McGurk MMN was only evoked by the incongruent audiovisual syllable (A/ba/V/va/) with a normal talking face, and no MMN was found in the same syllable with the face configuration manipulated, including an inverse normal face, an upright Thatcherised face (a face with eyes and mouth inverted) and an inverse Thatcherised face.

MMN evoked by incongruent audiovisual combinations

McGurk fusion is a special case of audiovisual integration where visual speech alters auditory speech perception without participants being aware. The perception of auditory speech perception is modified before MMN is evoked, therefore McGurk fusion MMN might merely reflect auditory processing rather than ongoing audiovisual integration processing (Besle et al., 2009). As mentioned earlier, in Section 1.3.1.2, another type of McGurk effect is an audiovisual combination where the discrepancy between auditory and visual speech can be clearly detected (e.g. /bd/); and yet, auditory and visual information still integrate. The process of audiovisual integration might be different in these two types of the McGurk effect. If McGurk MMN reflects audiovisual integration processing, then MMN evoked by two different types of McGurk stimuli could be different as well.

There is evidence that the brain activity in the combination type of McGurk effect is different from McGurk MMN. Colin et al. (2002b) tested MMN in both fusion (A/bi/V/gi/ which is perceived as /di/) and combination (A/gi/V/bi/ which is perceived as /bgi/) types of McGurk stimuli. MMN was early in fusion McGurk and lasted longer in latency. Colin et al. (2004) reported a different result in a follow-up MMN experiment, in which a changed methodology was employed to reduce the activity due to the physical difference between deviant and standard. Specifically, MMN was yielded by subtraction between the response of the deviant presenting among the standards and the response of the deviant presenting alone. The results showed that the MMN was only found in the McGurk fusion condition.

Kushnerenko et al. (2008) conducted an infant ERP study of the McGurk effect, which found evidence that fusion and combination McGurk syllables were processed differently in 5-month-old infants' brains. An incongruent syllable (A/ga/V/ba/) (combination) induced a response deviating from the response evoked by congruent (/ba/ and /ga/) and McGurk illusion A/ba/V/ga/ (fusion) syllables. Negative activity was found over the temporal area, and positive activity was also found over the frontal area, maximising at 360 ms after auditory signal onset. The authors postulated that the deflection evoked by the combination McGurk stimulus was likely to reflect activity for detecting audiovisual discrepancies rather than audiovisual integration.

In a more extreme case of audiovisual stimulus (e.g. a non-speech artificial audiovisual event), the MMN evoked by an audiovisual deviant is modality-specific. Belse et al. (2005) used a circle (visual) and tone (auditory) combination to examine the MMN evoked by non-illusory audiovisual integration. They tested four deviant conditions: audiovisual deviants differed from audiovisual standards in both auditory and visual inputs (audiovisual deviant) (standard: A1V1; deviant A2V2), in auditory input (auditory deviant) (standard: A1V1; deviant: A2V1), in visual input (visual deviant) (standard: A1V1; deviant: A1V2) and visual-only stimulus. The result showed that, in the visual deviant condition, the MMN was different from the traditional auditory MMN or McGurk-MMN reported in previous studies. Instead, it was similar to the MMN observed in the visual-only condition, which was largest in the occipital area rather than in the frontal area. The audiovisual deviant MMN was different from the sum of the MMNs evoked by the auditory deviant and the visual deviant. These findings suggest that the degree of audiovisual integration of the deviant stimulus might be a crucial factor in deciding MMN elicitation. Even though their results showed that audiovisual integration occurs, this integration could be much weaker than natural audiovisual events, such as audiovisual speech. The auditory and visual inputs could be processed separately, because the auditory and visual inputs of an artificial non-speech audiovisual event are not naturally related.

MMN evoked by incongruent audiovisual lexical tones

In MMN studies of audiovisual speech integration, most research has focused on the McGurk consonant effect. Research on other segmental speech (e.g. vowels) and supra-segmental speech (e.g. lexical tones) is almost non-existent. Whether McGurk MMN can be evoked by incongruent audiovisual lexical tones and whether the McGurk MMN of lexical tones is similar to that reported in the literature is unclear. By examining the McGurk MMN of incongruent lexical tones, whether mismatched visual information can modify auditory tone processing in the pre-attentive stage of processing can be investigated.

The McGurk effect in incongruent lexical tones could be different in processing from the McGurk effect found in consonants. Due to the articulation of tones being difficult to lip-read, mismatched visual information is less likely to influence auditory tones through visual phonetic features (e.g. place of articulation). The McGurk effect in lexical tones (if it is available) might rely on another type of visual information, such as duration. If visual information can modulate auditory tone perception, then McGurk MMN could be found in incongruent lexical tones. Additionally, if visual modulation is on the phonological level (changing lexical tone perception categorically), then MMN could show a left-lateralised pattern. However, phonetic visual cues are more important than non-phonetic visual cues, due to their stronger visual effect on speech perception (Kim & Davis, 2014); therefore, the duration of visual cues in lexical tones could be weaker than place visual cues in consonants in terms of the strength of the visual effect on auditory speech perception. Consequently, the McGurk effect for lexical tones could be much weaker than that for consonants. Since the magnitude of the McGurk effect can affect the elicitation of McGurk MMN (Eskelund et al., 2015), the MMN evoked by incongruent audiovisual lexical tones could be rather weak as well.

1.6 Research questions and hypotheses

Audiovisual integration in speech is not solely determined by the processing that the articulatory configuration is decoded into phonetic information, hence visual phonetic information can map onto a phonological representation in long-term memory. Some other non-speech-specific information is extracted from visual speech input that plays a role in integration processing along with auditory signals. The existence of two types of visual information in audiovisual speech integration is more compatible with the theory of multiple-stage audiovisual speech processing (Eskelund et al., 2011; Kim & Davis, 2014; Klucharev et al., 2003; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004; Soto-Faraco & Alsius, 2009; Stekelenburg & Vroomen, 2007). The non-speech aspect of visual information is accessed in early integration, which predicts the timing of the auditory signal to increase detection sensitivity, so that auditory sounds can be better perceived. Phonetic-specific visual information plays the role of information that is complementary to the auditory signal in the later stage of integration. When it comes to lexical tones, based on the postulation of multiple-stage audiovisual integration in speech, visual lexical tones can affect auditory tone perception, regardless of the saliency of visual features. That is, lexical tones could have both an audiovisual benefit effect and a McGurk effect. To verify this hypothesis, a series of experiments were performed to test these two audiovisual effects in Mandarin lexical tones.

1.6.1 Audiovisual benefit effect of lexical tones

1) Does visual information facilitate auditory tone perception?

The first aim of the current study is to confirm that Mandarin lexical tones can have an audiovisual benefit effect, even though they are difficult to lip-read, with identification and discrimination tasks. In Experiments 1–3, identification and discrimination tasks are employed to examine the audiovisual benefit effect. In each task, Mandarin lexical tone syllables are tested in audiovisual, auditory-only and visual-only modalities in noise conditions. The audiovisual benefit effect can be observed by comparing lexical tone identification or discrimination performance in the audiovisual and auditory-only modalities. The most relevant studies have reported that adding visual information improves lexical tones at certain levels in auditory adverse conditions for Mandarin (Burnham et al., 2001; Burnham et al., 2011; Mixdorff et al., 2005a; Mixdorff et al., 2005b; Mixdorff et al., 2006; Smith & Burnham, 2012), but the study of Liu et al. (2014) found a negative result for the audiovisual benefit effect of lexical tone recognition in real cochlear implant patients. In the current experiments, the visual information facilitation of auditory tone perception is expected to be found in noise conditions. In terms of lip-readability, the perceptual performance of Mandarin lexical tones in the visual-only condition is only slightly higher than a chance level (Chen & Massaro, 2008; Smith & Burnham, 2012), which suggests that the visual cues in lexical tones are not salient enough for good tone perception, but it also suggests that visual cues for lexical tones are still available. In the current experiments, lip-reading performance in tasks is also expected to be better than a chance level. The hypothesis for the first research question can be split into two parts:

Hypothesis 1a: If visual lexical tones facilitate auditory lexical tone perception in Mandarin, the identification and discrimination performance of Mandarin lexical tones will be better in the audiovisual modality than in the auditory-only modality in noise condition.

Hypothesis 1b: If there are phonetic-/ tonetic-specific visual cues in Mandarin lexical tones, lip-reading performance in the visual-only condition will be higher than a chance level.

These hypotheses will be discussed in Experiments 1–3 in Chapter 2.

2) Is the audiovisual benefit effect stronger in Tone 3 and Tone 4?

Second, Experiments 1–4 in Chapter 2 aim to test whether the audiovisual benefit effect is different in individual lexical tones, and which tones or tone contrasts have a stronger audiovisual benefit effect. Based on previous findings, among the four tones, the identification of T3 (dipping tone) or the discrimination related to T3 (e.g. T2-T3 contrast) seems to benefit more when adding visual information (Mixdorff et al., 2005b; Smith & Burnham, 2012). Two studies have proposed that potential visual cues for Mandarin lexical tones could be related to tone duration. Smith and Burnham (2012) reported that acoustic duration correlated with discrimination in Mandarin lexical tones, which further suggests tone articulatory movement has visual correlates with auditory tone duration features. This is reasonable to explain why the audiovisual benefit effect of T3 stands out among other tones due to T3 having the longest duration. As mentioned earlier regarding auditory lexical tone perception, duration is a crucial cue when the primary cue F0 is less reliable, especially for a longer tone (T3). However, if T3 is distinctive because of its salient duration feature, T4 should also be visually distinctive because it is the shortest tone. Mixdorff et al. (2005b) showed evidence that T3 and T4 identification benefits more from visual information, and these two tones had the highest identification in a de-voiced (no F0) condition. However, in Smith and Burnham (2012), the discrimination of tone contrasts containing T4 had a worse audiovisual benefit effect in tone discrimination, especially for T3-T4 contrast, and it was the worst in a visual alone condition among other tone contrasts.

It is difficult to explain the inconsistent findings of these two studies with the postulation of a duration-based visual cue, so this requires further investigation.

The current experiments postulate that tone duration could be a potential visual cue to the audiovisual benefit effect of lexical tones, based on which variations in the audiovisual benefit effect in individual lexical tones or tone contrasts are retested. In an identification task, T3 and T4 are expected to have a stronger the audiovisual benefit effect; and in a discrimination task, T3-T4 contrast is expected to have a larger audiovisual benefit than the other tone-contrasts. This audiovisual benefit pattern of T3 and T4 should also be consistent with the pattern in a visual alone condition. For other tones or tone contrasts, it is difficult to make clear predictions since the duration of contrast is not as distinctive as T3 and T4.

Note that potential visual cues (mouth opening duration, head movement and neck muscle tension) are proposed based on the assumption that they may convey tone-specific information and thus enhance audiovisual tone perception. In the current studies, possible head movement of speakers is deliberately controlled during recording, and neck muscle tension is not emphasised. Given that the visual cue of neck muscle tension can be noticed only when one is trained to do so (Chen & Massaro, 2008), visual cues beyond the mouth/lip region, such as head, eyebrows and neck muscle movement, are not discussed in the experiments. The hypotheses for the second research question are:

Hypothesis 2a: If there is distinctive visual information that is specific to tone features in Mandarin lexical tones, then the strength of the audiovisual benefit effect will not be equal across the four individual lexical tones.

Hypothesis 2b: If visual tone duration is distinctive visual information in an identification task (Experiments 1 and 4 in Chapter 2), then T3 and T4 will have the strongest audiovisual benefit effect compared to other lexical tones; likewise, in a discrimination task (Experiments 2–3 in Chapter 2), T3-T4 tone contrast will have the strongest audiovisual benefit effect compared to other tone contrasts.

Hypothesis 2c: If visual information that contributes to the audiovisual benefit effect is phonetic-/ tonetic-specific, then lexical tones that have a stronger audiovisual benefit effect will also have higher lip-reading performance in the visual-only condition.

These will be investigated in Experiments 1–3 in Chapter 2.

3) Does visual information facilitate early auditory tone processing?

The aim of the ERP experiments in Chapter 3 is to further explore the early stage of processing where visual speech interacts with the perception of auditory speech in Mandarin lexical tones. In Experiment 5, ERP responses (N1 and P2) are compared in audiovisual and auditory-only lexical tones. In Experiment 6, N1 and P2 are compared in audiovisual and auditory-only consonants. Evidence suggests that the auditory early components N1 and P2 are reduced in amplitude and shortened in latency due to the facilitation of visual speech. This could be due to a priming effect (Besle et al., 2004) or the predictiveness (van Wassenhove et al., 2005) of phonetic visual features for upcoming auditory speech content. Alternatively, this could also be because visual speech onset predicts the timing of upcoming auditory signal (Stekelenburg & Vroomen, 2007). For lexical tones, the visual feature is less salient than that in consonants, based on which the reduction effect in N1 and P2 should be smaller in lexical tone response compared to that in consonant response. If the reduction effect exclusively reflects phonetic visual information processing (e.g. place of articulation), it may not even be found in lexical tones, because the preceding visual articulatory movement is less able to predict the following auditory tone signal. If early audiovisual integration processing reflects non-phonetic visual information interaction, then lexical tones and consonants should have a similar reduction effect. That is, any difference in the reduction effect in N1/P2 between lexical tones and consonant responses should be due to the phonetic-specific processing of visual input in audiovisual integration. The hypotheses for Experiments 5–6 are:

Hypothesis 3a: If visual information facilitates auditory tone perception in the early stage of processing, the amplitude of N1 and P2 will be smaller and their latency will be earlier *in the audiovisual modality compared to that in the auditory modality (reduction effect in N1 and P2).*

Hypothesis 3b: If the reduction effect in N1 and P2 reflects the audiovisual integration of non-phonetic visual information (timing information), the reduction of lexical tones and consonants will be the same.

Hypothesis 3c: If the reduction effect in N1 and P2 reflects the audiovisual integration of phonetic-specific visual information (e.g. place of articulation), the reduction of lexical tones will be weaker than that of consonants.

1.6.2 McGurk effect of lexical tones

4) Does incongruent visual information modify auditory tone perception?

If congruent visual information improves auditory tone perception, can incongruent visual information change auditory tone perception? In an incongruent audiovisual tone syllable, if mismatched visual information interacts with auditory tone perception and leads to an illusory tone different from the original auditory tone, like the McGurk effect, it indicates that incongruent audiovisual lexical tone perception involves the audiovisual integration of phonetic information. However, if mismatched visual information only modifies auditory tone perception, which results in the same tone but with a different duration, it suggests that incongruent audiovisual lexical tone perception involves the audiovisual integration of non-phonetic information.

In the 'traditional' McGurk effect, mismatched visual speech biases auditory speech perception in terms of the place of articulation. For Mandarin lexical tones, the visual place of articulation is the same in the four tones, but the visual durations are different. Based on the finding of Green and Miller (1985) that a mismatched visual/ speaking rate can change the auditory perception between voiced and voiceless consonants, a mismatched visual lexical tone should consequently be able to change the perception of an auditory lexical tone in duration. To test this McGurk effect on lexical tones driven by the visual duration effect, in Experiment 7, T3 and T4 were selected to combine and form an incongruent lexical tone due to their distinctive duration contrast. Lexical tone identification was compared between incongruent lexical tone syllables and congruent lexical tone syllables, where the auditory inputs remained the same, but the visual input was different across conditions. For example, a syllable with auditory T3 was paired with visual T4 ($A_{T3}V_{T4}$) and compared with the congruent T3 syllable, so any difference between $A_{T3}V_{T4}$ and T3 would be due to the effect of the mismatched visual lexical tone (T4). If the visual duration cue is crucial in facilitating auditory lexical tone perception when auditory and visual lexical tones are congruent, this could interact with the perception of auditory tones' duration in the incongruent condition. For example, $A_{T3}V_{T4}$ would be perceived as a shorter T3 than the original one. If tone duration can affect tone category identification when the auditory cue (F0) is not reliable in noise, then mismatched visual information might even alter auditory tone categorisation. For example, $A_{T3}V_{T4}$ would be perceived as T4. The former visual influence could reflect audiovisual integration on the non-phonetic (duration feature) level of processing, while the later visual influence could reflect audiovisual integration on the phonetic (categorical) level of processing. However, considering that the visual effect of duration is weaker than the visual effect of the place of articulation, the McGurk effect of lexical tones could be rather small compared to segmental speech. The hypotheses for the research question in Experiment 7 in Chapter 4 are as follows:

Hypothesis 4a: For the incongruent audiovisual lexical tone syllable $A_{T4}V_{T3}$, if the incongruent visual tone T3 modifies (lengthens) the perception of the auditory tone T4 in duration, then lexical tone identification of the incongruent syllable $A_{T4}V_{T3}$ will be worse than congruent T4 stimuli in noise condition, and the reaction time of $A_{T4}V_{T3}$ will be slower than that of congruent T4.

Hypothesis 4b: For the incongruent audiovisual lexical tone syllable $A_{T3}V_{T4}$, if the incongruent visual tone T4 modifies (shortens) the perception of the auditory tone T3 in duration, then lexical tone identification of the incongruent syllable $A_{T3}V_{T4}$ will be worse than congruent T3 stimuli in noise condition, and the reaction time of $A_{T3}V_{T4}$ will be faster than that of congruent T3.

5) Does incongruent visual information modify auditory tone processing in the auditory cortex?

Following the last question, if an incongruent visual lexical tone modifies auditory tone perception leading to an illusory lexical tone in behavioural responses, this visual effect could show in ERP responses. Experiment 8 in Section 4.2 focused on the MMN component to investigate the process of the McGurk effect of incongruent audiovisual lexical tones in the brain. From the evidence of MMN studies in audiovisual speech perception, auditory MMN can be evoked by an infrequent McGurk syllable (deviant) among frequent congruent syllables (standard), even without there being an actual difference between the auditory speech inputs in the two conditions (Colin et al., 2004; Colin et al., 2002b; Klucharev et al., 2003; Knowland et al., 2014; Möttönen et al., 2002; Saint-Amour et al., 2007; Sams et al., 1991). This suggests that visual speech interacts with the representation of auditory speech in sensory memory in the auditory cortex. If an incongruent visual lexical tone modifies auditory tone perception in a way that is analogous to the McGurk effect in consonants, it could evoke MMN that is similar to McGurk MMN in consonants.

In Experiment 8, incongruent audiovisual lexical tone syllables $A_{T3}V_{T4}$ were infrequently presented among frequent congruent lexical tone syllables T3, where only the visual inputs differed across conditions. If incongruent lexical tone syllables activate MMN within the N2 time window at the frontal electrodes, it indicates that the brain detects the difference in auditory tones between incongruent and congruent conditions. Because auditory tones in incongruent and congruent syllables are physically the same, any difference response between them should be due to the processing of the auditory lexical tone being modulated by visual information.

Hypothesis 5a: During the processing of incongruent lexical tone syllables $A_{T3}V_{T4}$, if an incongruent visual tone (T4) modifies the perception of an auditory tone (T3) in the auditory cortex, then the McGurk MMN evoked by the modified auditory T3 will be observed over the frontal electrodes of the scalp after auditory signal onset.

Hypothesis 5b: If incongruent visual information does not modify the perception of the auditory tone (i.e. the processing of incongruent and congruent lexical tones is the same), then there will be no ERP difference between the two conditions.

1.6.3 Structure of this dissertation

This dissertation consists of eight experiments, which are described in detail in the chapters that follow and answer the research questions stated above. Generally, the experiments seek to test the audiovisual perception of Mandarin lexical tones in the audiovisual benefit effect and the McGurk effect.

First, Chapter 2 includes four behavioural experiments (Experiments 1–4), in which two paradigms – the identification and discrimination of Mandarin lexical tones – are employed to compare the performance of lexical tones in three modalities: audiovisual, auditory-only and visual-only in clear and noise conditions. The aim of these experiments is to investigate whether an audiovisual benefit effect on lexical tone exists and whether the audiovisual benefit effect is stronger for specific individual tones (e.g. Tone 3 and Tone 4).

In Chapter 3 (Experiments 5–6), ERP method is used to investigate the visual facilitation of auditory lexical tone perception in early processing. Experiment 5 compares the early auditory components (N1 and P2) between the response of audiovisual and auditory-only Mandarin lexical tones. To compare the audiovisual processing of lexical tones with speech, which is more visually distinctive, the ERP response to consonantal stimuli in audiovisual and auditory is tested in Experiment 6.

Chapter 4 (Experiments 7–8) investigates whether the effect of incongruent visual information on the perception of auditory lexical tones generates the McGurk effect. Experiment 7 focuses on a comparison of identification performance between incongruent and congruent audiovisual Mandarin lexical tones that differ only in the visual dimension. Experiment 8 measures the auditory MMN component evoked by incongruent audiovisual lexical tones. MMN driven by the McGurk effect of lexical tones serves as an index for whether mismatched visual information interacts with auditory tones, thus modulating the auditory tone process in the auditory cortex.

Lastly, Chapter 5 summarises the main findings of the studies, based on which the theoretical implications of audiovisual speech integration and the practical implications for Mandarin lexical tone learning are discussed. At the end of the chapter, it will point out the studies' limitations and suggest directions for future research.

Chapter 2 Audiovisual benefit effect of lexical tones in identification and discrimination tasks

2.1 Introduction

This chapter investigates whether visual lexical tone information facilitates the perception of lexical tones in behavioural responses with identification and discrimination tone tasks. The performance of Mandarin lexical tone perception was tested in audiovisual (AV), auditory-only (AO) and visual-only (VO) modalities in different noise conditions. First, lexical tone perceptual performance in AO and AV was compared to examine the audiovisual benefit effect. Basically, if lexical tone identification and discrimination performance is better in the AV modality compared to perception in the AO modality, it indicates that adding visual input along with auditory tones enhances lexical tone perception. To encourage participants to pay more attention to visual speech and avoid a ceiling effect, noises with different levels of SNR were added in all modalities. In the VO modality, audiovisual lexical tones were presented by removing the auditory speech signal. VO performance reflects the degree of lip-readability of lexical tones and the availability of visual cues specific to tone features.

As mentioned in Section 1.6.1, the experiments in this chapter first aim to confirm that the audiovisual benefit effect exists in Mandarin lexical tones in noise conditions. That is, lexical tone perception in the AV modality is expected to be better than that in the AO condition. More importantly, the experiments aim to test which individual lexical tones have a relative stronger audiovisual benefit effect. According to the hypothesis of the visual duration cue proposed in previous research (Mixdorff et al., 2005b; Smith & Burnham, 2012), tones with more distinctive durational features should have a greater audiovisual benefit effect. Therefore, dipping tone (T3) and falling tone (T4) in identification, or T3-T4 contrast in discrimination, are expected to benefit more from presenting a visual input due to their distinctive visual duration feature. In addition to the audiovisual benefit effect, the experiments aim to test whether lexical tones can be lip-read and whether the lip-read pattern is consistent with the audiovisual benefit pattern for individual tones. If the tones or tone contrasts that have higher lip-read performance also have a stronger audiovisual benefit effect, this would support it being the phonetic visual cue that improves lexical tone perception.

Four behavioural experiments in this chapter will answer the questions above. In Experiment 1, the identification of Mandarin syllables with four lexical tones is tested in three modalities (AO, AV and VO) in both quiet and noise conditions. Experiments 2 and 3 employ a discrimination task (same-different judgement) where six possible lexical tone contrasts are presented. Again, Mandarin syllables in four tones were presented in AO, AV and VO modalities in both quiet and noise conditions. Experiment 4 tests two lexical tone contrasts: T2 vs T3 and T3 vs T4, with a two-alternative-choice identification task, in which stimuli are presented in AO and AV, in clear and noise conditions, respectively. Generally, in these experiments, the presence of the audiovisual benefit effect (AV > AO) of lexical tone perception is expected in the noisy condition. Variation in the degree of the audiovisual benefit effect across different individual lexical tone types/ contrasts is also anticipated. T3 and T4 or T3-T4 contrast is expected to stand out from other tones or tone contrasts.

2.2 Identification of audiovisual lexical tone perception (Experiment 1)

In Experiment 1, the identification accuracy rate of Mandarin syllables with four lexical tones is tested in AO, AV and VO modalities, in clear and noise conditions. In noise conditions there are two types of noise— pink noise and babble noise— at three levels of SNR: -6 dB, -9 dB and -12 dB. The purpose of using different types of noise at various SNR levels is to test whether the degree of audiovisual benefit is related to the availability of an auditory signal and to determine the most favourable noise type or noise level for the audiovisual benefit effect in lexical tones. Based on Mixdorff et al. (2005b), the audiovisual benefit effect of lexical tones increases with decreasing SNR levels, which suggests that the audiovisual benefit effect is larger when the auditory signal is weaker. Between the two types of noise, babble noise that contains multi-talker speech in Mandarin is more destructive to the auditory tone signal compared to pink noise (Dees et al., 2007). Therefore, the audiovisual benefit effect of lexical tones is expected to be larger in noise that has a stronger masking effect, that is, in the condition of babble noise with SNR -12 dB. In terms of individual tones, T3 and T4 are predicted to have a stronger audiovisual benefit effect than other tones due to their distinctive duration feature. T3 and T4 are also expected to have higher lip-read performance in the VO modality. Overall tone identification in the VO modality is predicted to be low, but it could be slightly higher than a chance level (i.e. 25% accuracy).

2.2.1 Method

2.2.1.1 Participants

Twenty-eight native speakers of Mandarin (aged 25.5 ± 4.1 years; 16 females) from Bournemouth University were recruited for this study. All participants reported normal or correct to normal visual acuity and had no previous hearing impairments. All participants were right-handed. The participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board.

2.2.1.2 Materials

Two sets of consonant-vowel-vowel (CVV) monosyllables /bai/ with four lexical tones (/bāi/, /bái/, /bǎi/, /bài/) and /dai/ with four lexical tones (/dāi/, /dái/, /dǎi/, /dài/) comprised the experimental materials (see Table 2.1). Three modalities of stimulus (AV, AO and VO) were presented in three blocks, respectively. The AV stimuli were presented as video clips of articulating faces. AO stimuli were present in auditory speech alone with a fixation cross at the centre of the screen. VO stimuli were derived from AV stimuli with the auditory tracks removed. A male native Mandarin speaker recorded the video materials with a Sony HDR-SR12E camera in a soundproof booth. The recordings were then digitised and edited in Adobe Premiere Pro CC (Adobe Systems, California) set to a resolution of 1280 × 720 pixels and at a frame rate of 29.97 frames per second. The auditory tracks were edited in Audacity (Crook, 2012), sampling at 48 kHz and 32 bits, and they were root mean square

(RMS) normalised to an amplitude of -12 dB. The speaker in the video clips only showed his head and the upper part of his neck (see Figure 2.1).

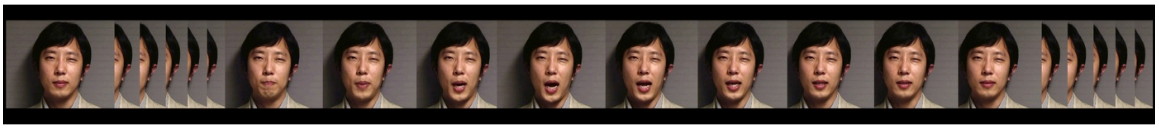


Figure 2.1 Articulation of the syllable /bāi/

In terms of the acoustic features of the lexical tones in the experiment, duration, F0, intensity and F0 contour were measured and plotted using Praat (Boersman & Weenink, 2013), as presented in Figure 2.2 and Table 2.1. Among the four lexical tones, the tone duration of T3 is the longest one, and T4 is the shortest one. In terms of F0, T3 is the lowest pitch; T1 and T4 are relatively higher. The intensity levels of the tones are not distinctively different. For the F0 movement of the four tones shown in Figure 2.2, T1 is a high-level pitch without much variation. T2 starts from a lower pitch and gradually rises towards the end. T3 falls at the beginning until the turning point (i.e. the point at which the pitch contour turns from falling to rising) and then gradually rises in the remaining part of the movement. T4 falls rapidly from a high pitch. The video duration of lexical tones was measured in picture frames from mouth opening to mouth closure and converted to milliseconds.

Table 2.1 Corpus of Mandarin syllables in Experiment 1. Syllables /bai/ and /dai/ with four tones with meanings, acoustic feature measurement: F0, intensity and duration, and visual mouth movement duration from opening to closure.

Syllable	Tone	Gloss	F0 (Hz)	Intensity (dB)	Audio duration (ms)	Visual duration (ms)
/bāi/	T1	to break	124	66	864	1285
/bái/	T2	white	116	69	790	1134
/bǎi/	T3	to display	95	62	957	1301
/bài/	T4	to defeat/failed	118	65	438	1084
/dāi/	T1	to stay/dull	128	68	689	951
/dái/	T2	NA	118	68	755	1034
/dǎi/	T3	to seize/bad	96	65	860	1084
/dài/	T4	to bring/belt	125	68	432	901

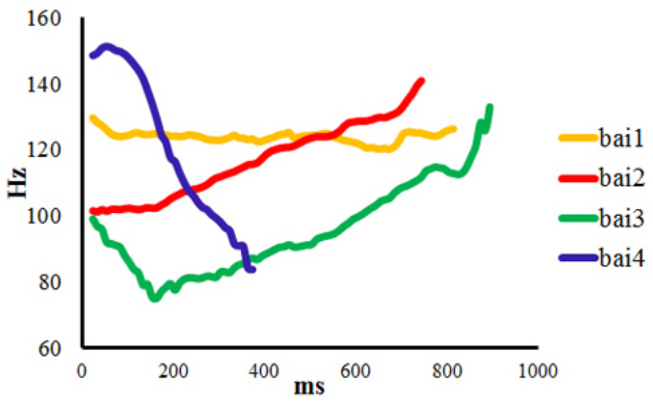


Figure 2.2 F0 movement of the syllable /bai/ in four tones

All stimuli were presented in clear and noise conditions, respectively. The noise condition comprised two types of noise: babble and pink noise. The babble noise consisted of the mixed voices of six Mandarin native speakers reading different sentences, and the pink noise was generated via Audacity (Crook, 2012). To make sure that the noise masked all the auditory signals within a syllable, the auditory tracks of the noise were consistently longer than the syllable duration. Additionally, to prevent participants predicting the onset time of AO stimuli in noise, the noise masking and the onset of the stimulus varied in each trial. Two types of noise were presented at three SNR levels: -6 dB, -9 dB and -12 dB, respectively. Therefore, the trials included three presentation modalities (AV, AO, VO) × 8 syllables (2 syllables × 4 tones) × 7 listening conditions (clear, 2 noise types × 3 SNR), which totalled 168 trials in the experiment.

2.2.1.3 Procedure

The participants were tested individually, and the presentation of stimuli was controlled by the psychology software E-Prime 2.0 (Psychology Software Tools, Sharpsburg) on a PC in a soundproof lab. Auditory sound was played using Sennheiser HD 280 (Sennheiser electronic GmbH & Co. KG, Wedemark) noise-cancellation headphones, and the output volume remained at about sound pressure level (SPL) 65 dB. Participants were instructed to identify the lexical tones of the presented syllables by pressing corresponding keys (one of four alternatives) on the keyboard. They had also been informed that the syllables were presented in AV, AO and VO in three separate blocks. Considering that VO stimuli were derived from AV stimuli, participants might develop a strategy for memorising linguistically irrelevant visual features (e.g. speaker's idiosyncrasies) from the AV syllables to identify VO syllables if the AV stimuli were consistently presented before VO stimuli. To avoid the bias of an order effect, the block sequence was counterbalanced for each participant. Prior to the real experiment, a practice session was given, and the stimuli in this session were not used in the current experiment. All the participants were told to watch the speaker's face in the video clips throughout the experiment, but they were not informed of which specific part of the face they should watch. The whole process lasted for about 20 minutes, and two breaks in-between the blocks were given.

2.2.2 Results

The statistical analysis consisted of three parts: 1) AO vs AV under a clear condition; 2) AO vs AV under noise conditions; 3) VO analysis. The dependent variable of the analysis was the lexical tone identification rate. If the identification rate of AV was higher than that of AO, then visual lexical tones improve lexical tone perception; that is, an audiovisual benefit effect is found. The modality comparison between AO and AV under noisy condition was the main interest, since the audiovisual benefit effect was expected to appear under noisy condition rather than clear condition, where ceiling effect would very likely to undermine observation of the difference between conditions. The analysis of VO identification aimed to test whether the lexical tones presented could be visually identified (lip-readability) compared to a chance level (25%). The calculation of significant effect was put into the Greenhouse Geisser correction (Jennings & Wood, 1976) wherever applicable, and the post hoc effect was adjusted with the Bonfferoni correction. In the analysis, the dependent variable was accuracy rate. Reaction time was not calculated in the data, because within each syllable the auditory signals started at random time points after the noise started playing.

Descriptively, identification accuracy in the AO, AV and VO conditions can be seen in Table 2.2 and Figure 2.3. Identification was the highest in the clear condition, and it greatly decreased with SNR levels in noise conditions. AV accuracy seems to be higher than AO accuracy, especially in noise conditions. VO performance was the worst among the three modalities in all conditions, yet it appears to be slightly better than chance level.

Table 2.2 Identification accuracy (%) (*N* = 28) of lexical tones in AO, AV, VO modalities under clear and different noisy conditions (babble noise at SNR-6 dB, -9 dB, -12 dB and pink noise at SNR -6 dB, -9 dB, -12 dB).

Mode	**Clear**		**Noise**												
			Babble						**Pink**					
			-6		**-9**		**-12**		**-6**		**-9**		**-12**	
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
AO	94.6	1.6	71.0	3.6	46.0	3.8	32.1	2.6	89.3	2.2	75.0	3.2	39.7	3.9
AV	96.4	1.3	76.3	3.5	71.9	2.9	50.0	4.0	90.6	2.6	82.6	3.6	53.1	3.4
VO	30.8	2.8	32.6	3.1	27.7	2.7	28.1	3.6	25.4	2.5	29.9	2.4	33.6	3.5

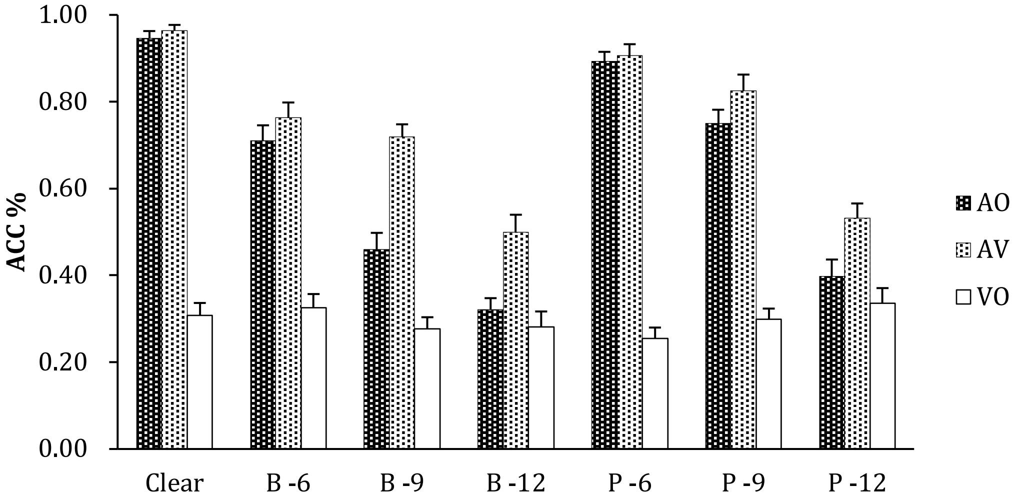


Figure 2.3 Identification accuracy (*N* = 28) of lexical tones in AO, AV, VO modalities in clear and different noisy conditions (babble noise at SNR-6 dB, -9 dB, -12 dB and pink noise at SNR -6 dB, -9 dB, -12 dB).

Clear condition

In the comparison between AO and AV in the clear condition, a two-way repeated measures analysis of variance (ANOVA) with the factors of Modality (AO, AV) and Tone type (T1, T2, T3, T4) was performed. The results showed that the main effect of Modality was not significant, *F*(1, 27) = 1.15, *p* = .29. The main effect of Tone type was significant, *F*(1.97, 53.28) = 4.5, *p* = .016, but Tone type had no interaction with Modality, *F*(1.59, 43.01) = 0.155, *p* = .93. AV was not significantly better than AO here, which means the audiovisual benefit effect was not significantly present in general lexical tone identification or in any individual tone identification. This is clearly due to ceiling effect that limited the observation of modality difference.

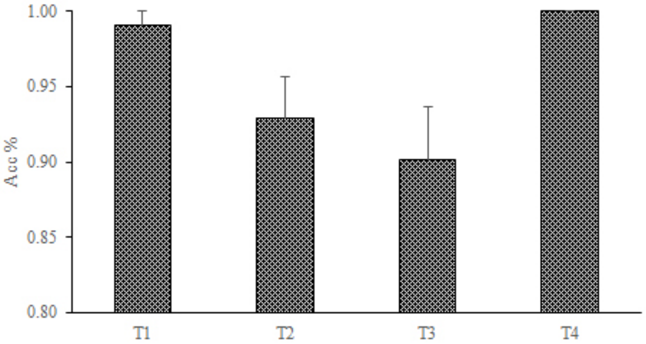


Figure 2.4 Identification accuracy of individual lexical tones in the clear condition, regardless of modality.

Post hoc comparisons of Tone type regardless of Modality showed that the most significant effect was found in the T3 and T4 comparison, which showed that T4 accuracy was significantly higher than T3 (*p* = .009). Additionally, T1 was significantly better than T3 (*p* = .022), and T4 was better than T2 (*p* = .018). With regard to the T2 vs T3 comparison (*p* = .523) and the T1 vs T4 comparison (*p* = 0.33), no significant effect was found. The T1 vs T2 comparison also failed to reach significance (p = .005). As shown in Figure 2.4, the accuracy was not equally good for each tone. Obviously, T1 and T4 were identified as significantly better than T2 and T3 in both AO and AV. The poorer T2 and T3 can be explained by their confusability in terms of acoustic features. In AO, lexical tone perception relies heavily on F0, including F0 high and F0 contour, and T2 and T3 have a similar F0 contour, which easily causes perceptual confusion between them (Jongman et al., 2006; Xu, 1997). The audiovisual results also showed that lower identification of T2 and T3 might indicate that the auditory pitch contour was utilised as a main cue even when visual information was provided in the clear condition.

Noise conditions

To analyse the audiovisual benefit effect in the noise condition, a separate four-way ANOVA with the factors Modality (AO, AV), Noise (pink, babble), SNR (-6 dB, -9 dB, -12 dB) and Tone type (T1, T2, T3, T4) was conducted. The results showed that the main effect of Modality was highly significant: $F(1, 27) = 32.16, p < .001$. The main effect of Noise [$F(1, 27) = 129.09, p < .001$], SNR [$F(2, 54) = 238.08, p < .001$] and Tone type [$F(3, 81) = 9.35, p < .001$] were also significant. In terms of two-way interaction, a significant effect was found in the interaction between Modality and Noise: $F(1, 27) = 10.54, p = .003$. The interaction between Modality and SNR was also significant: $F(2, 54) = 13.18, p < .001$. Three-way interaction of Modality × Noise × Tone type only reached a marginally significant level: $F(3, 81) = 2. 52, p = .064$. The other three-way interaction of Modality × SNR × Tone type had a small yet significant effect: $F(6, 162) = 2.50$, $p = .024$. The interaction effect of Modality × Noise × SNR was marginally significant: $F(2, 54) = 2.44$, $p = .097$. Similarly, the four-way interaction effect was only marginally significant: $F(4.44, 119.95) = 2.15, p = .072$.

First, the audiovisual benefit effect on general lexical tone perception (regardless of tone type) was found in both the babble and pink noise conditions. A pairwise comparison of Modality based on Noise showed that AV accuracy (66.1 ± 2.8%) was higher than AO accuracy (49.7 ± 2.8%) in babble noise ($p < .001$) and the same pattern was found in pink noise [AV: 75.4 ± 2.8% > AO: 68.0 ± 2.1% ($p = .003$)]. In the comparison of Modality based on SNR, the results revealed an audiovisual benefit effect in the SNR -9 dB condition ($p < .001$), where AV (77.2 ± 2.8%) was better than AO (60.5 ± 2.9%); and in SNR -12 dB ($p < .001$), where AV (51.6 ± 3.2%) was higher than AO (35.9 ± 2.8%) (see Figure 2.5).

To further compare the audiovisual benefit size (AV – AO) across conditions, a separate t-test comparison between pink and babble noise showed that the audiovisual benefit effect was significantly stronger in babble noise (16.4 ± 2.7%) than in pink noise (7.4 ± 2.3%)

[*t*(27) = 3.24, *p* = .003]. A one-way ANOVA of the degree of the audiovisual benefit at different SNR levels revealed that the effect was significantly different across SNR levels: $F(2, 54) = 13.18, p < .001$. The benefit effect was stronger at SNR -9 dB (16.7 ± 2.8%) and SNR-12 dB (15.6 ± 2.6%) noise levels, and there was no significant difference between them. The lowest benefit effect was at SNR -6 dB (3.3 ± 2.7%) (see Figure 2.5).

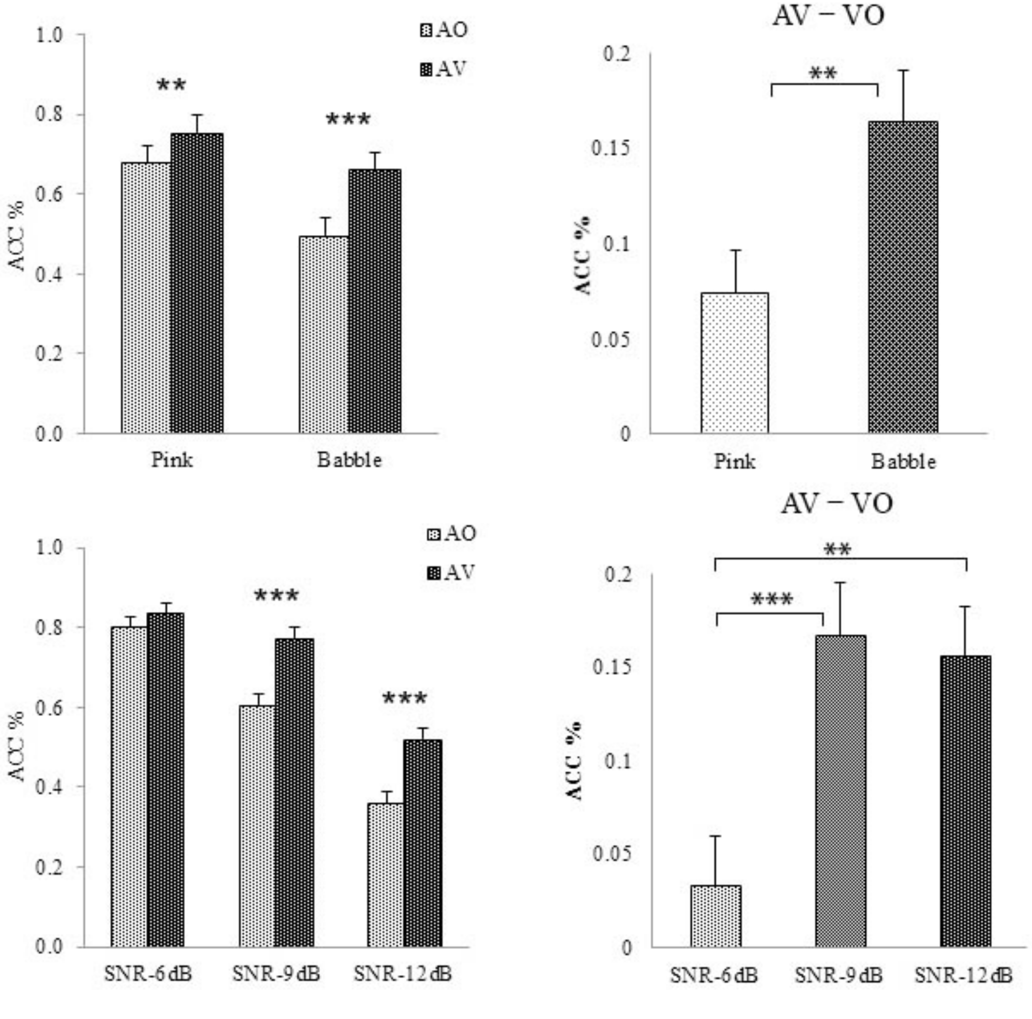


Figure 2.5 The identification accuracy of general lexical tones in AO and AV under noise conditions. The top left figure is AO vs AV for lexical tone accuracy in pink and babble noise conditions. The top right figure is the audiovisual benefit size (AV – AO) in pink and babble noise conditions. The bottom left is the accuracy of AO vs AV for lexical tones in SNR-6 dB, SNR-9 dB and SNR -12 dB. The bottom right represents the audiovisual benefit size for SNR levels of -6 dB, -9 dB and -12 dB conditions. **: *p* < .01; ***: *p* < .001.

Individual lexical tones

In terms of the audiovisual benefit effect for individual lexical tones, a pairwise comparison of Modality based on SNR and Tone type demonstrated that a significant audiovisual benefit effect for tone types was not found at SNR -6 dB, but it was found in the SNR -9 dB and SNR -12 dB conditions. In the SNR -9 dB condition, all tone types reflected a significant audiovisual benefit effect except for T3 (T1: $p < .001$; T2: $p = .012$; T3: $p = .41$; T4: $p < .001$). In the SNR -12 dB condition, only T1 ($p = .02$) and T2 ($p = .001$) had a significant benefit effect. Because there was no consistent pattern for the audiovisual benefit for individual tones across the three SNR levels, it is difficult to answer the question regarding which tone type benefits the most from visual information. In the comparison of Modality based on Noise type and Tone type, the audiovisual benefit effect pattern of the individual tones differed significantly in the two noise types. In the babble noise condition, all individual tones had a significant audiovisual benefit effect except for T3 (T1: $p < .001$; T2: $p = .005$; T3: $p = .264$; T4: $p < .001$). In the pink noise condition, a significant audiovisual benefit effect was only found in T2 ($p = .030$) (see Table 2.3 and Figure 2.6).

Table 2.3 Identification accuracy (%) of four lexical tones in noise conditions. Upper table: the identification accuracy of four lexical tone types of AO and AV for three SNRs. Bottom table: the accuracy of four lexical tones of AO and AV in babble and pink noise.

SNR	Tone	AO		AV	
		Mean	SE	Mean	SE
SNR-6	T1	71.4	5.3	75.9	4.5
	T2	83.0	4.1	79.5	5.0
	T3	83.0	4.3	89.3	3.7
	T4	83.0	2.9	89.3	3.5
SNR-9	T1	47.3	5.0	70.5	5.2
	T2	55.4	4.9	70.5	5.2
	T3	79.5	3.7	83.9	4.1
	T4	59.8	3.9	83.9	3.9
SNR-12	T1	32.1	5.4	44.6	5.9
	T2	31.3	4.0	58.0	5.9
	T3	52.7	4.1	64.3	5.2
	T4	27.7	4.9	39.3	5.1

Noise	Tone	AO		AV	
		Mean	SE	Mean	SE
Babble	T1	29.2	4.5	49.4	6.0
	T2	47.0	4.0	63.1	4.6
	T3	84.5	4.5	89.9	2.8
	T4	38.1	4.4	61.9	3.6
Pink	T1	71.4	5.1	78.0	3.9
	T2	66.1	3.8	75.6	5.0
	T3	58.9	4.1	68.5	4.8
	T4	75.6	3.0	79.8	4.2

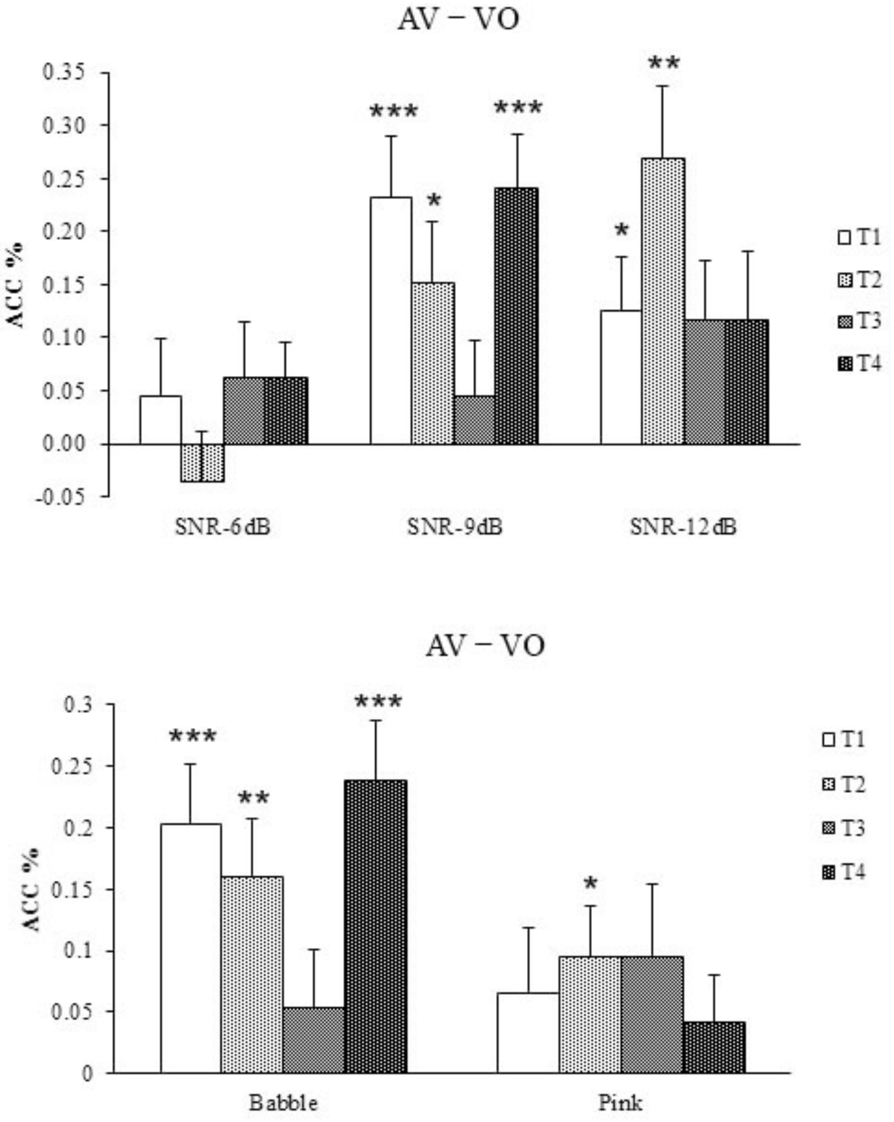


Figure 2.6 Audiovisual benefit for individual tones. Upper figure: the size of the audiovisual benefit effect (AV – AO) for four lexical tones in SNR-6 dB, SNR-9 dB and SNR-12 dB conditions. Bottom figure: the size of the audiovisual benefit effect for four lexical tones in babble and pink noise conditions. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Visual-only condition

In the analysis of identification in the VO condition, a two-way ANOVA with Listening condition (clear, 2 noise types × 3 SNR) and Tone type (T1, T2, T3, T4) showed no significant effect for any of the factors. It suggests that VO identification of lexical tones was not affected by noise and each tone had a similar accuracy. A t-test of a comparison of VO accuracy and a chance level (25%) showed that VO accuracy (29.7 ± 6.2%) was significantly higher than chance level: $t(27) = 3.18$, $p = .004$. As shown in Figure 2.7, the identification of all lexical tones in the VO modality was slightly higher than chance level. Despite the fact that the lip-read performance of VO lexical tones was the lowest compared to identification in the other conditions, the result indicated that tone-specific visual cues were available.

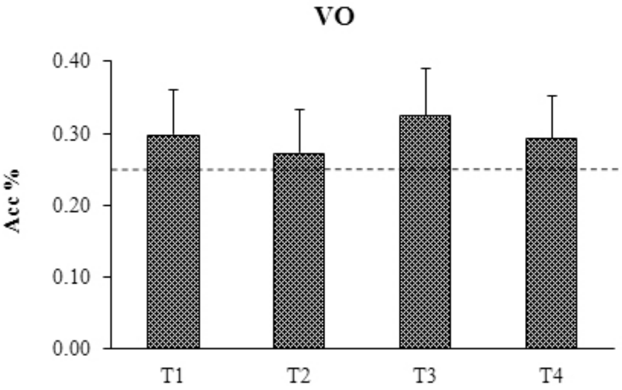


Figure 2.7 Average accuracy of lexical tones in the VO condition, regardless of clear or noise conditions. The dashed line refers to chance level.

2.2.3 Summary of the results

First, an audiovisual benefit effect of lexical tones was found in the noise condition rather than in the clear condition, as predicted, and the effect responded to both noise types (babble and pink) and all SNR levels (see Figure 2.5). The strength of the audiovisual benefit effect tended to be stronger in the condition where AO accuracy was lower. For example, in the results for noise type and SNR level, the audiovisual benefit effect was stronger in babble noise than in pink noise, and it was also stronger in lower SNR noise (-9 dB, -12 dB) than in SNR-6 dB noise. However, this does not necessarily indicate that the audiovisual benefit effect increases along with a decreasing auditory signal. In the three levels of SNR noise, the audiovisual benefit was indeed larger in lower SNR rather than higher SNR, but it was not statistically different in SNR-6 dB and in SNR-9 dB. In fact, the benefit tended to be slightly higher in SNR -9 dB than in SNR-12 dB.

In terms of individual lexical tones, the audiovisual benefit effect was not identical for each tone, but there was no particular lexical tone that was consistently better than others (see Figure 2.6). The audiovisual benefit for T1 and T2 was found in both SNR -9 dB and SNR -12 dB. The audiovisual benefit for T4 was only found in the SNR -9 dB condition. T3 benefited the least from visual information in any of the SNR conditions. The lower audiovisual benefit for T3 could be due to the auditory cues (e.g. F0) of T3 that remained available in the noise conditions, as T3 in AO was most accurately identified among the other tones in most of the noise conditions (see Table 2.3). Therefore, visual information might be relied on less during audiovisual perception of T3.

The VO results showed that lip-reading lexical tones was not easy but possible, because VO tone identification was slightly better than chance level. The identification in VO was not significantly different for individual lexical tones; therefore, a specific tone that is more visually distinctive than others could not be determined.

The findings in Experiment 1 confirmed that presenting visual information along with auditory tones improved lexical tone perception in noise. However, there were no clear results that indicate which tone has a stronger audiovisual benefit effect, and the VO results did not find that any tones having better lip-reading either. Hence, the current results cannot infer that which potential tone-specific visual cue was utilised during audiovisual lexical perception. The lack of a clear pattern of the audiovisual benefit for individual tones might be due to the difficulty of the task. In a four-alternative-choice identification task, it could be particularly difficult to make a judgement by ruling out a wider range of candidates in heavy noise conditions and the VO condition. Additionally, it might be due to an insufficient number of trials for the analysis of each tone type; hence, a stable effect in individual tones cannot be attained.

In the next two experiments, a same-different discrimination paradigm was employed to further test the questions left unaddressed from Experiment 1. In contrast to an identification task, perceivers only need to compare two lexical tones in a discrimination task, which is a more direct method to probe the visual feature contrast of given tones.

2.3 Cross-modality discrimination of audiovisual lexical tone perception (Experiment 2)

In Experiment 2, lexical tone discrimination was tested with a same-different judgement task in AO, AV and VO modalities. From the results of Experiment 1, the audiovisual benefit effect is confirmed, but T3 and T4 did not have a larger audiovisual benefit effect, as predicted. In the discrimination task, perceivers can focus more on a direct comparison between two tones, which might be boost the discrimination between tones that are contrastive and diminish the discrimination of tones that are similar in visual features. Thus, it might help to find variation in the audiovisual benefit effect in individual tone contrasts. In this experiment, an audiovisual benefit effect of general tone discrimination is expected, and T3-T4 is still predicted to have greater facilitation in AV modality and higher discrimination in VO modality.

2.3.1 Method

2.3.1.1 Participants

Twenty native speakers of Mandarin (aged 27.6 ± 5.1 years; 11 females) from Bournemouth University were recruited for this study. Most of the participants did not take part in the previous experiment. All participants reported normal or correct to normal visual acuity and had no previous hearing impairments. One of the participants was left-handed. The participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board.

2.3.1.2 Materials

The experiment employed the monosyllable /bai/ with four lexical tones (/bāi/, /bái/, /băi/, /bài/) from Experiment 1 as stimuli presented in three modalities: AO, AV and VO. The noise for the degrading auditory signal was babble noise with two SNR levels (-6 dB and -9 dB). The babble noise contained a mixture of six native Mandarin speakers (i.e. a recording of their sentence reading in Mandarin). The average volume of all stimuli was RMS normalised at about 65 dB. The same male Mandarin native speaker involved in Experiment 1 recorded video materials in a noise-cancelled booth. The speakers produced each syllable three times, from which the one with the best quality was selected as an experimental stimulus. The speaker in the video clips was only presented from the top of the head to the upper part of the neck (see Figure 2.1). The video clips were edited in Adobe Premiere Pro CC (Adobe Systems, California) at a resolution of 1280 × 720 with a rate of 29.97 frames per second. The auditory tracks were edited in Audacity (Crook, 2012) at 48 kHz with a 32-bit amplitude resolution.

2.3.1.3 Procedure

This experiment employed a same-different discrimination paradigm in which the participants were told to judge whether two syllables given in succession were the same or different in terms of lexical tones in each trial. AO, AV and VO stimuli were presented in three separate blocks that were counterbalanced for every participant. As can be seen in Figure 2.8, within each trial, the first stimulus was always presented as an auditory syllable and the second one was presented in AO, AV or VO in the corresponding block. The inter-stimulus interval (ISI) between the two was set at 500 ms, and the inter-trial interval (ITI) was set at 1,000 ms. The participants were allowed to respond to the key 3,000 ms after the onset time of the second stimulus.

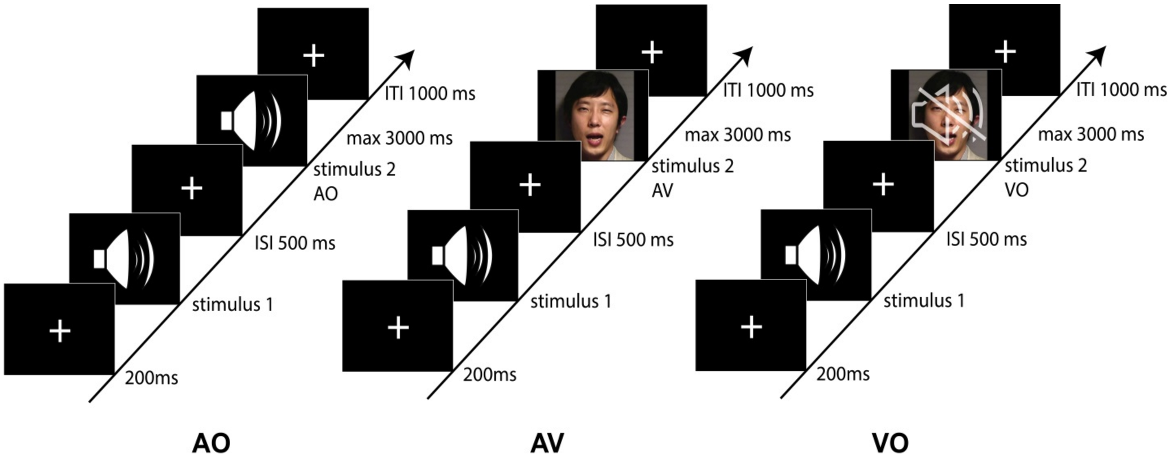


Figure 2.8 Procedure of a trial presentation in AO, AV and VO blocks. Within each trial, the first stimulus was always an auditory syllable, and the second trial was presented in different modalities according to the block.

In each condition, there were 24 lexical tone contrasts including 12 contrasts of the 'different' type (AB, BA), and 12 contrasts of the 'same' type (AA, BB). The possible tone-contrast combinations were as follows: 1. different type T1-T2, T1-T3, T1-T4, T2-T3, T2-T4 and T3-T4; 2. same type T1-T1, T2-T2, T3-T3 and T4-T4. The six 'different' types were each presented twice; the second time included the same combination with a reverse sequence of two stimuli. Every 'different' type contrast had a 'same' type contrast (AB, AA; BA, BB), so that the discrimination index (DI) as the dependent variable could be calculated. The DI was calculated using the method in Burnham et al. (2001). A DI of each condition was derived from the difference between the hit rate and the false alarm rate, divided by 4. The hit rate includes the correct rate for the 'different' and 'same' type tone contrasts, and the false alarm rate refers to the incorrect rate of the 'different' type and 'same' type tone contrasts. The range of DI was from 1 to -1. A maximal DI value of 1 indicates full discrimination ability for a correct response (perfect discrimination), while -1 refers to full discrimination ability for an incorrect response, and 0 means that discrimination is at a chance level.

In order to induce the audiovisual benefit effect, two SNR levels (-6 dB, -9 dB) of babble noise were added to the second stimulus of each trial in the noise condition. The stimuli in three listening conditions (clear, SNR-6 dB, SNR-9 dB) were presented randomly within the blocks throughout the experiment. In total, there were 216 trials (3 modalities × 3 listening conditions × 24 tone contrasts). Prior to the experiment, the participants were given 15 trials for practice and these practice trials were not used in the experiment session. The participants were not given any suggestions regarding paying attention to any specific part of the articulating faces. The whole process lasted for about 25 minutes, and two breaks in-between blocks were given.

2.3.2 Results

The statistical analysis consisted of three parts: 1. modality comparison (AO vs AV) in the clear condition; 2. AO vs AV in the noise conditions; 3. discrimination in VO. Again, a audiovisual benefit effect for lexical tone discrimination exists if the DI of AV is statistically higher than that of AO in the same condition. The analyses had Greenhouse-Geisser correction (Jennings & Wood, 1976) wherever applicable, and a post hoc comparison was adjusted using Bonferroni correction.

First, as illustrated in Table 2.4, lexical tone discrimination was best in the clear condition, and it dropped with decreasing SNR levels. The DI in VO was the worst among the modalities. The DI of AV did not seem to be better than that in AO in the noise conditions.

Table 2.4 Average DI (*N* = 20) of individual tone contrast in AO, AV and VO modalities under clear, SNR-6 dB and SNR-9 dB conditions.

Condition	Modality	T1-T2		T1-T3		T1-T4		T2-T3		T2-T4		T3-T4	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Clear	**AO**	0.98	0.01	0.94	0.03	0.96	0.03	0.89	0.05	0.98	0.01	0.97	0.03
	AV	1.00	0.00	1.00	0.00	0.98	0.03	0.95	0.03	0.98	0.03	1.00	0.00
	VO	-0.11	0.05	0.03	0.09	0.18	0.07	0.07	0.09	0.29	0.08	0.26	0.08
SNR-6 dB	**AO**	0.45	0.09	0.58	0.08	0.27	0.08	0.80	0.05	0.66	0.06	0.37	0.09
	AV	0.43	0.11	0.72	0.07	0.42	0.08	0.24	0.08	0.44	0.07	0.51	0.09
	VO	0.03	0.09	0.12	0.07	0.11	0.09	0.01	0.07	0.32	0.07	0.22	0.09
SNR-9 dB	**AO**	0.15	0.08	0.50	0.09	0.13	0.09	0.57	0.09	0.36	0.06	0.10	0.08
	AV	0.20	0.06	0.19	0.08	0.19	0.08	0.12	0.08	0.47	0.09	0.48	0.07
	VO	0.12	0.07	0.12	0.07	0.08	0.08	0.11	0.08	0.34	0.08	0.25	0.08

Clear condition

To compare AO and AV in the clear condition, a two-way repeated measures ANOVA with the factors of Modality (AO, AV) and Tone contrast (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4 and T3-T4) was subjected to the lexical tone DI. The main effect, Modality, was just significant: $F(1, 19) = 4.47$, $p = .048$. However, the main effect of the other factors was not significant. A pairwise comparison of Modalities found the DI in AV (0.98 ± 0.01) was slightly yet significantly higher than the one in AO (0.95 ± 0.01).

Noise conditions

A three-way ANOVA with the factors Modality (AO, AV) × SNR (-6 dB, -9 dB) × Tone contrast was conducted for the DI in noise conditions. The results showed that a significant main effect was found for SNR [$F(1, 19) = 26.72, p < .001$] and for Tone contrast [$F(5, 95) = 6.19, p < .001$], but not for Modality [$F(1, 19) = 0.683$, $p = .42$]. The interaction effect between Modality and Tone contrast was significant: $F(5, 95) = 15.74, p < .001$.The three-way interaction was also significant: $F(5, 95) = 4.46, p = .001$.

The comparison of Modality based on Tone contrast showed that a significant modality difference appeared in the tone contrasts T2-T3 ($p < .001$) and T3-T4 ($p = .012$). However, the direction of the modality difference was not the same. In the T2-T3 contrast, AV discrimination (0.18 ± 0.06) was significantly lower than AO discrimination (0.68 ± 0.05), while in T3-T4, AV (0.50 ± 0.07) was significantly higher than AO (0.23 ± 0.08). This means that an audiovisual benefit effect was only found in the T3-T4 contrast while the reverse effect (audiovisual inhibition effect) was found in the T2-T3 contrast (see Figure 2.9).

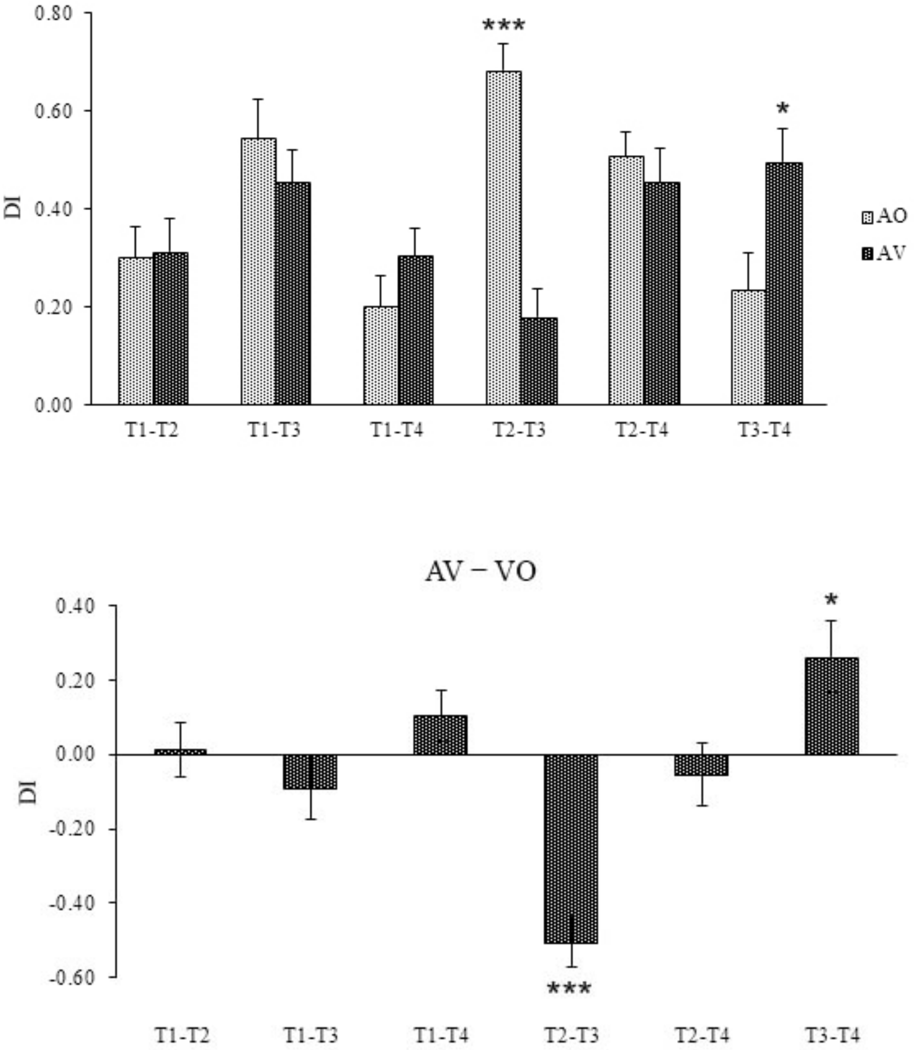


Figure 2.9 DI of individual tone contrasts in AO and AV under the noise condition (regardless of SNR level). The upper figure refers to a comparison of AO and AV modalities; the bottom figure refers to the audiovisual benefit (AV – AO) of individual tone contrasts. *: *p* < .05; ***: *p* < .001

A further pairwise comparison of Modality based on SNR and Tone contrast revealed that a significant modality difference was found in the tone contrasts T2-T3 (*p* < .001) and T2-T4 (*p* = .012) in the SNR -6 dB condition, and it was also found in the Tone contrasts T1-T3 (*p* = .019), T2-T3 (*p* = .001) and T3-T4 (*p* = .002) in the SNR -9 dB condition. However, an audiovisual benefit effect was only found in the contrast T3-T4 in SNR-9 dB, where AV was better than AO discrimination. The other tone contrast had a visual inhibition effect in which the discrimination in AV was lower than that in AO modality (see Figure 2.10).

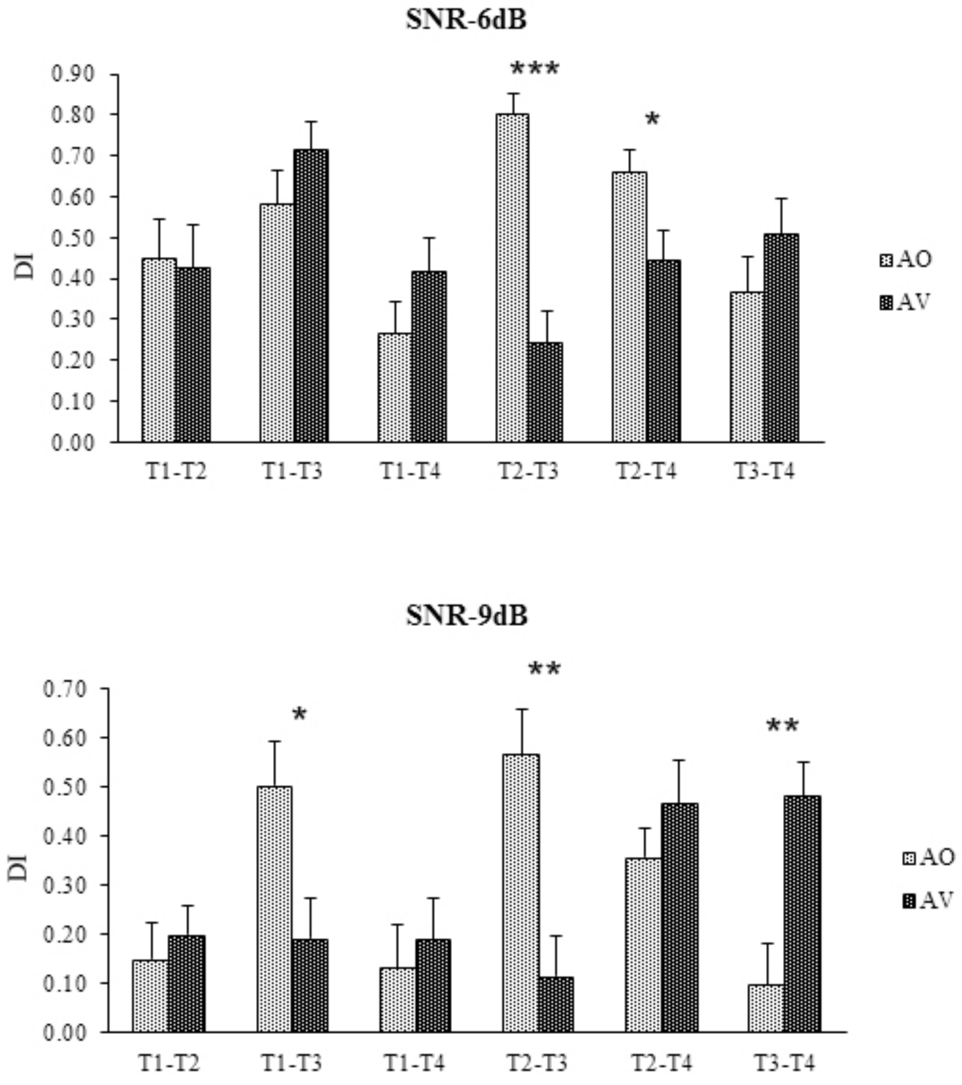


Figure 2.10 DI of individual tone contrasts in AO and AV modalities under SNR -6 dB (upper figure) and SNR -9 dB (bottom figure) conditions. *: *p* < .05; **: *p* < .01; ***: *p* < .001.

Visual-only condition

With regard to lexical tone discrimination in VO, a two-way ANOVA with Listening condition (clear, babble SNR -6 dB, babble SNR -9 dB) and Tone contrast was performed to investigate whether VO discrimination would be uneven in each tone contrast and whether adding extra noise would influence VO performance. The results showed that VO discrimination had a significant effect on Tone contrast [*F*(3.0, 57.02) = 7.09, *p* < .001], but not on Listening condition [*F*(2, 38) = .42, *p* = .66]. In addition, the interaction effect between the two was not significant: *F*(10, 190) = .79, *p* = .64. A post hoc test for Tone contrast showed that the tone contrasts T2-T4 and T3-T4 were significantly higher than the other tone contrasts, and there was no significant difference between T2-T4 and T3-T4 (see Figure 2.11). Although the overall VO performance was lowest relative to AO and AV performance, it was still significantly higher than chance level: *t*(19) = 4.32, *p* < .001.

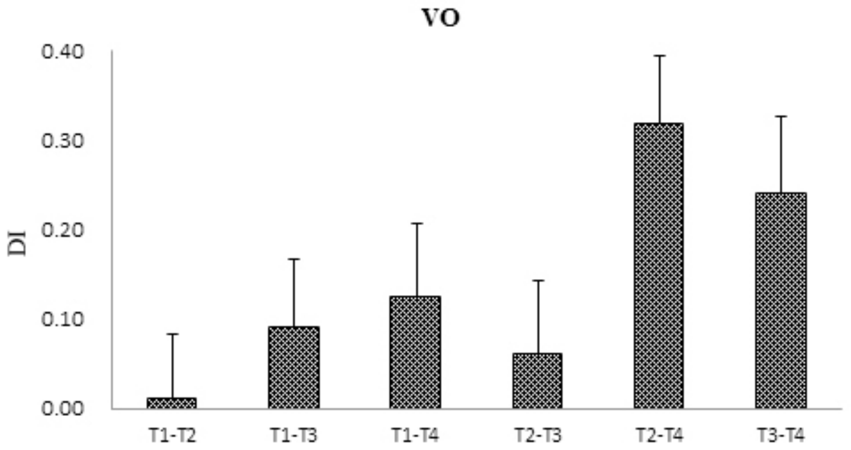


Figure 2.11 DI of the individual lexical tone contrasts in VO modality.

2.3.3 Summary of the results

The audiovisual benefit effect

In Experiment 2, an audiovisual benefit effect for general tone discrimination was found in clear conditions. In noise conditions, no significant audiovisual benefit effect was discovered. In terms of the discrimination of individual tone contrasts, the audiovisual benefit effect was significant in the tone contrast T3-T4 in the SNR -9 dB condition. Interestingly, a significant audiovisual inhibition effect (AV is poorer than AO) was also found in certain tone contrasts (e.g. T2-T3 in SNR -6 dB; T1-T3, T2-T3 in SNR -9 dB), in which presenting extra visual information decreased the discrimination relative to AO. When averaging two SNRs, the visual inhibition effect remained significantly robust in T2-T3 and the audiovisual benefit effect remained in T3-T4 (see Figure 2.9). Clearly, in the lexical tone discrimination task, adding extra visual information does not always improve but can also hinder lexical tone perception in the noise condition. This implies that mouth movement provides tone-specific visual cues that were beneficial in distinguishing T3 and T4 but simultaneously blurred the boundary between T2 and T3.

Visual-only condition

Additionally, individual tone discrimination in VO also found that T2-T4 and T3-T4 were more distinguishable than other tone contrasts, which might be due to the contrastive features between T2/T3 and T4. The better audiovisual benefit effect for T3 and T4 might indicate that the distance of the visual features between these two tones is relatively larger, while the visual features between T2 and T3 are smaller. This inference is partially supported by the pattern of VO lexical tone discrimination in which visual T3-T4 was easier to distinguish while visual T2-T3 was less distinguishable. This suggests that visual cues that contributed to the audiovisual benefit effect might engage the lip-read cues used in VO discrimination.

Task effect

Two opposite visual effects may be attributed to the potential visual cues facilitating/ inhibiting the discrimination of lexical tones, but another reason that cannot be completely ruled out is the bias caused by the same-different task paradigm. First, the task difficulty could be different in AO and AV. In the trial, given that the first stimulus was persistently auditory, the task would be easier when the second stimulus was also presented in AO. Alternatively, the task would be harder when the second one was present in AV. A discrimination task for speech perception can be conducted by comparing features and by using phonetic category information where judgement is made based on categorisation of the first stimulus (Gerrits & Schouten, 2004; Pisoni, 1973). In AO-AO trials, both approaches can be used, but in AO-AV trials, judgement depends on the memory of the tone category activated by the AO tone that matches the corresponding visual tone memory. Consequently, cross-modality trials could be more difficult to process. Second, the same-different paradigm might bring about a ‘same-trial’ bias, since individuals tend to respond to 'different-trials' only when they are very certain (Gerrits & Schouten, 2004), hence decreasing the sensitivity of discrimination. These task-specific biases might increase the visual effect on tone discrimination – i.e. tone contrasts (e.g. T2-T3) that benefit less from visual cues have a visual inhibition effect, while tone contrasts (e.g. T3-T4) that benefit more from visual cues have a substantial audiovisual benefit effect. When these two visual effects appear in the same listening condition, the visual influence on general tone perception can be neutralised as the two opposite visual effects cancel each other out when averaging all tone contrasts.

In addition to the visual effect on T2-T3 (audiovisual inhibition) and T3-T4 (audiovisual benefit), visual information also significantly influenced other tone contrasts. In order to determine whether the results could be replicated, a similar discrimination task was adopted in the following experiment (Experiment 3). Additionally, in order to reduce possible bias caused by cross-modality processing in audiovisual and VO conditions, in Experiment 3, the modality of the two stimuli within each trial remained identical (i.e. AO-AO, AV-AV, VO-VO).

2.4 Within-modality discrimination of audiovisual lexical tone perception (Experiment 3)

Similar to Experiment 2, this experiment adopts a discrimination paradigm (a same-different judgement task) to investigate the audiovisual benefit effect of lexical tones, but the experimental token and modality within trials were different from the last experiment. The previous two experiments used stimuli produced by only one speaker, to which perceivers could respond based on low-level features (e.g. speaker's idiosyncrasies) instead of speech-specific visual features of stimuli. To avoid possible bias, the syllables used within each trial were produced by two speakers in this experiment. Moreover, to avoid the task effect mentioned in the last experiment (see Section 2.3.3), the two stimuli within each trial for comparison remained consistent in modality (within-modality comparison), instead of a cross-modality comparison as in Experiment 2.

2.4.1 Method

2.4.1.1 Participants and materials

Twenty Mandarin native speakers (aged 24.7 ± 3.5 years; 16 females) participated in the study. None of them had participated in Experiment 2. All reported that they had no known hearing problems and their vision was normal or correct to normal. One of the participants was left-handed. The participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board.

The experiment employed the monosyllable /bai/ with four tones (/bāi/, /bái/, /băi/, /bài/) as stimuli presented in three modalities: AO, AV and VO. The video materials were from recordings of two native male Mandarin speakers. The stimuli properties of the two speakers can be seen in Table 2.5. Speaker 1 was the speaker in Experiments 1–2. Compared to the acoustic properties of the syllables recorded by Speaker 1, the syllables produced by Speaker 2 were slightly lower in pitch, longer in duration and weaker in T1 and T4 intensity. The auditory durations of the four tones were slightly different across the two speakers. Compared with Speaker 1, Speaker 2 produced longer T1, T2 and T3 but shorter T4. The video duration of lexical tones was measured in picture frames from mouth opening to mouth closure, and then converted to milliseconds. In addition, the T3 syllable from the new speaker had a creaking sound at the pitch turning point. The two stimuli presented in each trial were from the two speakers, respectively. The listening conditions for the presented stimuli, including clear, SNR -6 dB and SNR -9 dB conditions, were identical to those in Experiment 2.

Table 2.5 Acoustic features of the syllable /bai/ with four lexical tones from two speakers, including F0 value in Hz, duration in ms and intensity in dB. Visual duration is measured from mouth opening to mouth closure.

Syllable	Tone	Speaker	F0 (Hz)	Intensity (dB)	Audio duration (ms)	Visual duration (ms)
/bāi/	T1	1	124	66	864	1285
		2	136	65	1101	1133
/bái/	T2	1	116	69	790	1134
		2	129	64	985	1200
/băi/	T3	1	95	62	957	1301
		2	112	62	1139	1317
/bái/	T4	1	118	65	438	1084
		2	140	66	305	700

2.4.1.2 Procedure

The same paradigm (a same-different discrimination task) that was applied in Experiment 2 was also applied in this experiment. However, in this experiment, the stimuli in each trial were consistent in their modality, which means that in AO block, the stimuli to compare were presented in AO modality; in AV block, both two stimuli were presented in AV modality; in the block of VO, both stimuli were in VO modality (see Figure 2.1). The purpose was to reduce the process effect when comparing cross-modality stimuli as in Experiment 2. The procedure for the trial presentation, the number of trials and the instructions for participants remained identical to that in Experiment 2.

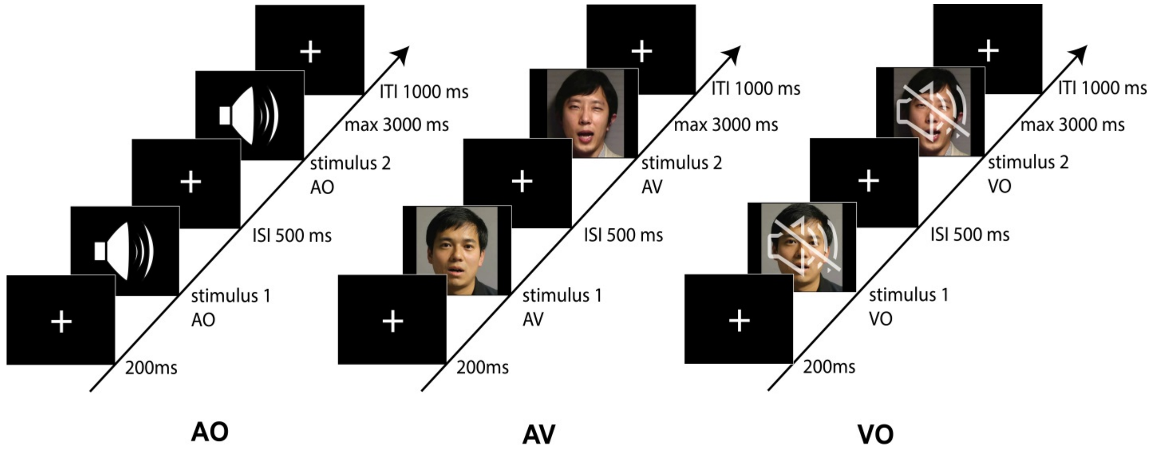


Figure 2.12 Procedure for the trial presentation in AO, AV and VO blocks. Within each trial, two different speakers produce the first stimulus and the second stimulus, and the two stimuli were always consistent in their modality.

2.4.2 Results

Due to the similarity of this experiment to Experiment 2, the statistical analysis applied the same ANOVAs as in the previous experiment. The overall data for all conditions can be seen in Table 2.6. AV discrimination did not seem to be higher than AO in any condition. VO discrimination was still the lowest one among the three modalities.

Table 2.6 Average DI (*N* = 20) of individual lexical tones in AO, AV and VO modalities under clear, SNR -6 dB and SNR -9 dB conditions.

Condition	Modality	T1-T2		T1-T3		T1-T4		T2-T3		T2-T4		T3-T4	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Clear	AO	0.97	0.03	0.85	0.04	0.94	0.03	0.63	0.09	1.00	0.00	0.88	0.04
	AV	0.87	0.05	0.90	0.04	1.00	0.00	0.62	0.08	0.95	0.03	0.85	0.05
	VO	0.04	0.06	0.09	0.07	0.19	0.09	0.00	0.11	0.10	0.10	0.07	0.10
SNR -6 dB	AO	0.51	0.08	0.58	0.07	0.23	0.04	0.89	0.05	0.31	0.07	0.11	0.06
	AV	0.38	0.10	0.63	0.08	0.24	0.07	0.27	0.05	0.34	0.10	0.43	0.09
	VO	0.13	0.07	0.05	0.10	0.14	0.09	0.14	0.08	0.11	0.09	0.25	0.08
SNR -9 dB	AO	-0.01	0.07	0.50	0.09	0.27	0.05	0.33	0.09	0.13	0.06	0.13	0.05
	AV	0.14	0.07	0.23	0.09	0.25	0.08	0.20	0.07	0.31	0.08	0.26	0.09
	VO	-0.12	0.09	-0.02	0.09	-0.12	0.09	-0.06	0.10	-0.06	0.12	-0.01	0.08

Clear condition

The DI of the clear condition was put into a two-way ANOVA (Modality × Tone contrast). The main effect, Modality, was not significant: $F(1, 19) = 0.34, p = .56$. However, Tone contrast had a significant effect: $F(1.79, 34.08) = 11.30, p < .001$. The interaction effect was only marginally significant: $F(2.95, 56.11) = 2.76, p = .051$. The comparison of Modality based on Tone contrast showed no significant effect on any tone contrasts. In the clear condition, the audiovisual benefit effect was barely observable. The DI of the individual tone contrasts had a very similar pattern in AO and AV, where the discrimination of T2-T3 was particularly poor compared to other tone contrasts (see Figure 2.13).

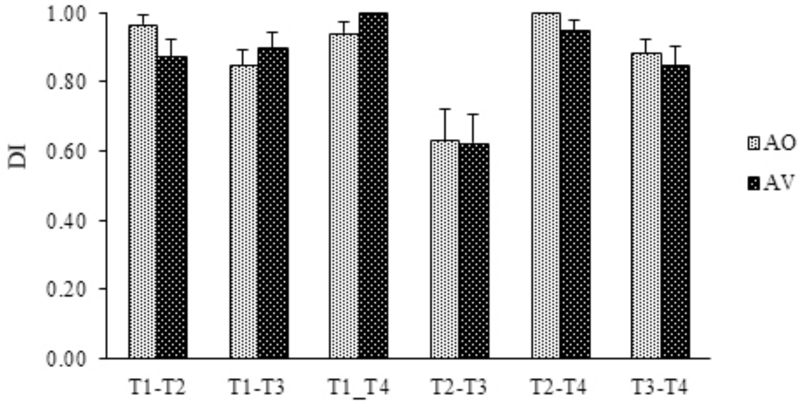


Figure 2.13 DI of the individual tone contrasts of AO and AV modalities in the clear condition.

Noise conditions

The DI in the noise condition was subjected to a three-way ANOVA (Modality × SNR × Tone contrast). The results showed that the main effects of SNR [$F(1, 19) = 31.02, p < .001$] and Tone contrast [$F(3.11, 58.99) = 7.63, p < .001$] were significant. However, Modality was not significant: $F(1, 19) = 0.44, p = .52$. The two-way interaction between Modality and SNR was not significant [$F(1, 19) = 0.63, p = .44$], but the interaction between Modality and Tone contrast was significant: $F(5, 95) = 10.43, p < .001$. The three-way interaction effect was significant as well: $F(5, 95) = 5.29, p < .001$. The comparison of Modality based on Tone contrast found significant differences in T2-T3 ($p < .001$) and T3-

T4 (*p* = .011), which was similar to the results in Experiment 2. There was an audiovisual benefit effect on T3-T4 [AO (0.12 ± 0.05) < AV (0.35 ± 0.07)], and an audiovisual inhibition effect on T2-T3 [AO (0.61 ± 0.06) > AV (0.24 ± 0.05)] (see Figure 2.14).

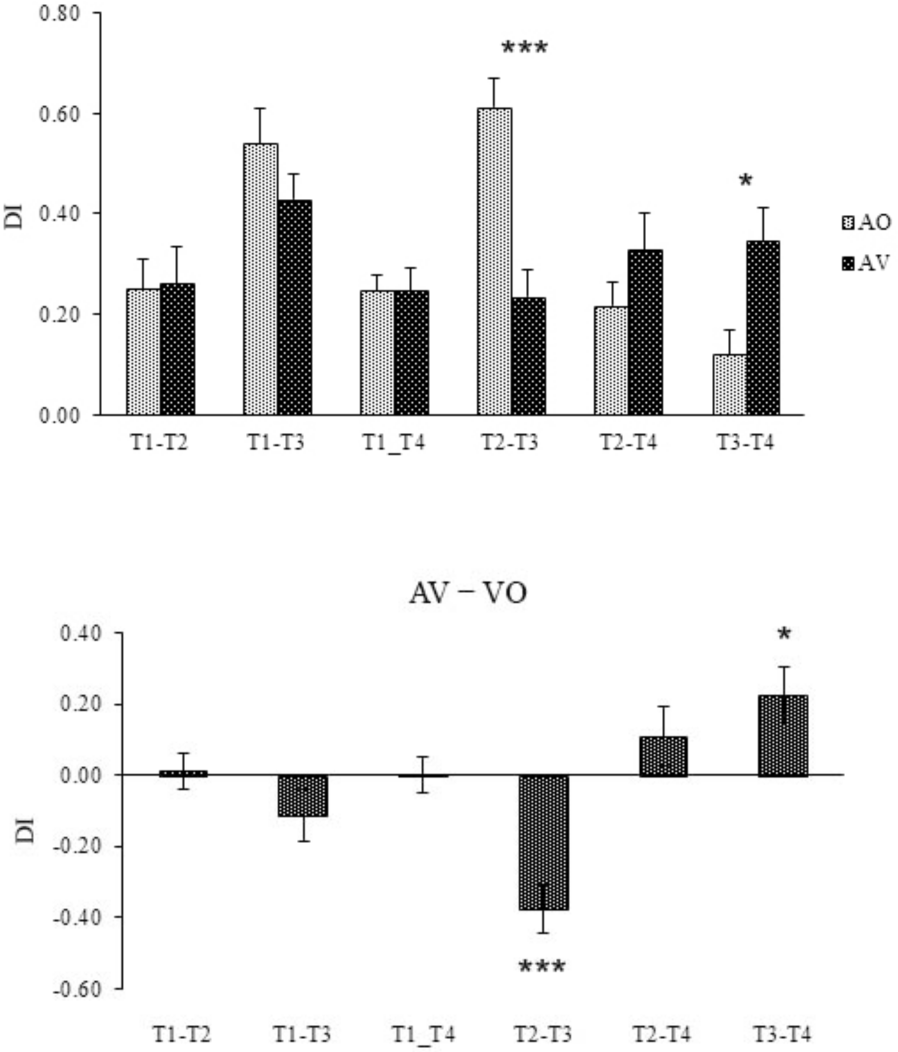


Figure 2.14 DI of individual tone contrasts in AO and AV in noise conditions (regardless of SNR level). The upper figure refers to a comparison of AO and AV modalities; the bottom figure refers to the AV benefit (AV – AO) of individual tone contrasts. *: *p* < .05; ***: *p* < .001

A further comparison of Modality based on Tone contrast and SNR found that the visual effect patterns were different in SNR -6 dB and SNR -9 dB. In SNR -6 dB, a significant audiovisual inhibition effect was found in T2-T3 (*p* < .001), and an audiovisual benefit effect was found in T3-T4 (*p* = .012), but in SNR -9 dB the audiovisual benefit effect was on T1-T2 (*p* = .039), and the audiovisual inhibition effect was on T1-T3 (*p* = .042) (see Figure 2.15).

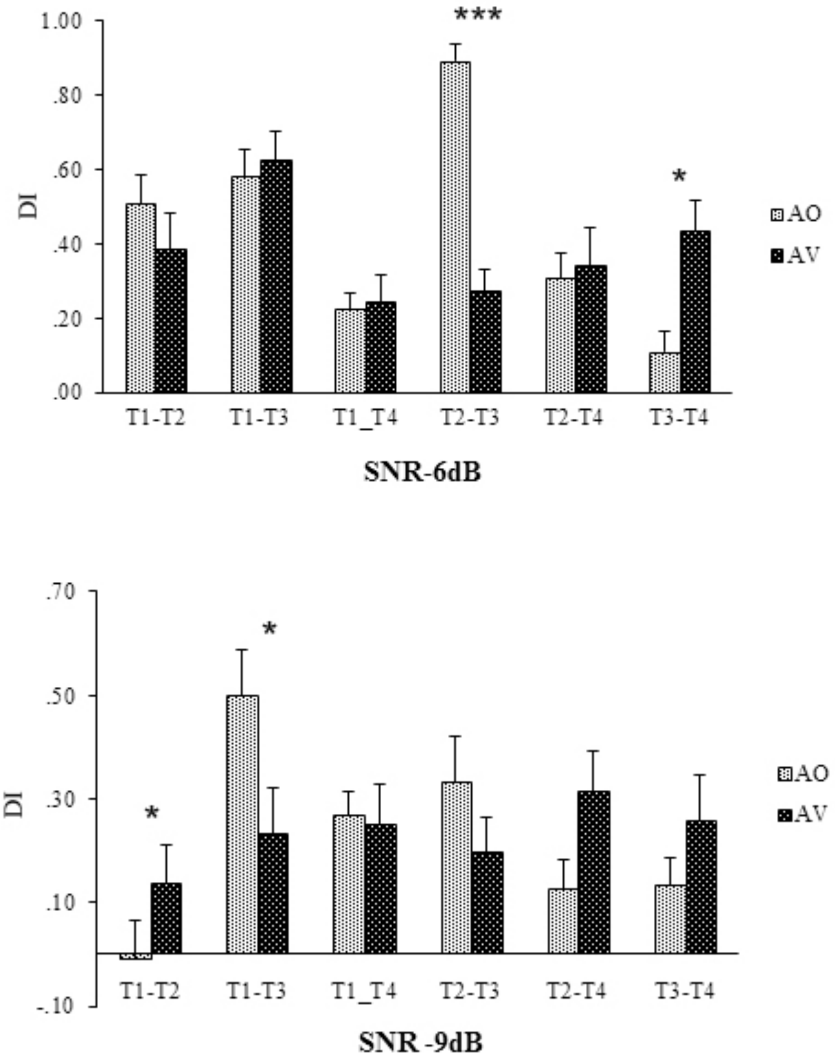


Figure 2.15 DI of individual tone contrasts in AO and AV under SNR -6 dB (upper figure) and under SNR -9 dB (bottom figure) conditions. *: $p < .05$; ***: $p < .001$

Visual-only condition

A two-way ANOVA with Listening condition (clear, SNR -6 dB, SNR -9 dB) and Tone contrast (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4) was subjected to DI in VO. The analysis found that the main effect of Listening condition was significant: *F*(2, 38) = 4.18, *p* = .023. However, the main effect of Tone contrast [*F*(3.82, 72.60) = .46, *p* = .75)] and the interaction effect were not significant [*F*(6.37, 121.02) = .55, *p* = .77]. A post hoc test of Listening condition revealed a significant difference between the SNR -6 dB and SNR -9 dB conditions. Specifically, VO performance was higher in SNR -6 dB noise than it was in SNR -9 dB noise (*p* = .005). Moreover, a t-test comparison between VO and chance level found that only VO in SNR -6 dB was significantly better than the chance level: *t*(19) = 2.70, *p* = .014 (see Figure 2.16).

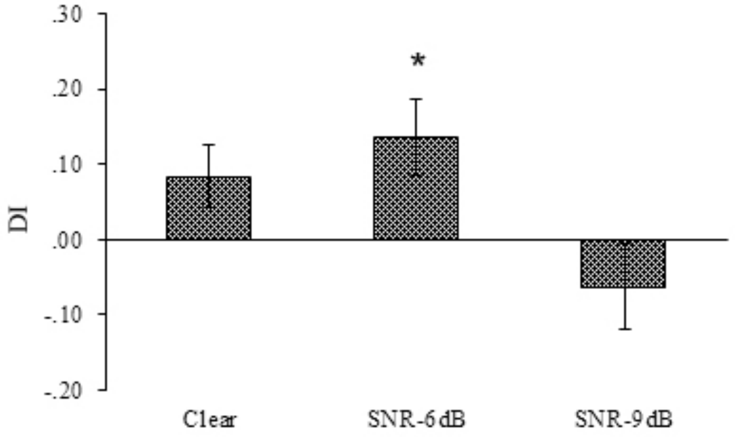


Figure 2.16 DI of general lexical tones in the VO modality in clear, SNR -6 dB and SNR -9 dB conditions. *: *p* < .05 for VO identification was better than chance level.

2.4.3 Summary of the results

The audiovisual benefit effect

Similar to the results in Experiment 2, an audiovisual benefit effect was not found in general lexical tone discrimination in any conditions, although it was found in individual lexical tones, such as T1-T2 and T3-T4 in noise. Along with an audiovisual benefit effect, a significant audiovisual inhibition effect was found in some tone discriminations, such as T1-T3 and T2-T3. When combining two SNR conditions, the tone contrasts with the strongest visual effects were found in T3-T4 (audiovisual benefit) and T2-T3 (audiovisual inhibition) (see Figure 2.14), which was consistent with the findings in Experiment 2.

Visual-only condition

The discrimination of VO in Experiment 3 was lower than that in Experiment 2. The level of noise clearly affected judgement in the VO modality. Since the noise in both experiments was identical, the different VO performances could be due to the task effect. In the VO block, both stimuli presented within each trial were in the VO modality. Therefore, the participants had to make judgements based on the lip-read result of the initial stimulus. However, the lip-read accuracy of lexical tones was very low, meaning that there was a higher chance for participants to make judgements based on unreliable recognition of the first tone in a trial. Another possible reason is related to the cross-speaker judgement in each trial, which increased the task difficulty. The tone duration of the first speaker is shorter than that of the second speaker. This might cause confusion when extracting tone duration as a visual cue; that is, visual duration could be less reliable for lip-reading lexical tones when multiple speakers are involved. This may explain why the performance in VO did not show that a particular tone contrast had better discrimination than others.

Next experiment

As the results of Experiment 1–3 have shown, an audiovisual benefit effect on lexical tone perception was found consistently. Moreover, an audiovisual inhibition effect was discovered simultaneously along with an audiovisual benefit effect. It is noteworthy that a significant audiovisual inhibition effect was only observed in same-different discrimination tasks rather than in the identification task. The audiovisual inhibition effect is possibly the result of a weak audiovisual benefit effect in the discrimination paradigm, rather than the result of genuine visual inhibition of lexical tone perception. It could be that visual cues are not as useful in T2-T3 discrimination as in T3-T4 discrimination; consequently, T2-T3 in the AV modality remains a confusable pair, and hence T2 and T3 are perceived as the same tones. That is, T2-T3 possibly has the least audiovisual benefit, while T3-T4 has the strongest audiovisual benefit.

To further test the strength of the audiovisual benefit effect on T2-T3 and T3-T4, these two tone contrasts were tested again with a two-alternative-choice identification task in the next experiment (Experiment 4). If the distinctiveness of visual features causes an audiovisual benefit effect in lexical tones, then a stronger audiovisual benefit effect would be expected in T3/T4 identification, while a weaker audiovisual benefit should be found in T2/T3 identification. If the audiovisual inhibition effect is task-dependent, then it should not appear in T2/T3 identification.

2.5 Identification of audiovisual lexical tones: Tone 2 vs Tone 3 and Tone 3 vs Tone 4

In Experiment 4, only two lexical tone contrasts – T2/T3 and T3/T4 – were tested for an audiovisual benefit effect for lexical tones with a two-alternative-choice identification task. The aim of the study is twofold: first, it is to test the existence of an audiovisual benefit effect for lexical tone perception; second, and more importantly, through comparing the degree of the audiovisual benefit effect in the two lexical tone contrasts, the distinctiveness of potential visual cues for lexical tones can be probed. In Experiments 2 and 3, adding visual information has a stronger influence on the two tone contrasts, T2-T3 and T3-T4. The audiovisual benefit effect is confirmed in T3-T4 as predicted, but the audiovisual inhibition effect is difficult to interpret. One possibility is that audiovisual inhibition is caused by a task (discrimination) effect. To rule out this possibility, this experiment re-employs the identification paradigm, but it only tests two tone-contrasts, T2/T3 and T3/T4, in separate tone identification tasks. Prediction of an audiovisual benefit effect in the T2/T3 task should be lower than that in theT3/T4 task, where the T3 benefit effect might be different in the two tasks, although T3 is physically identical across tasks. Additionally, the audiovisual inhibition effect in T2-T3 should disappear in the identification task. To allow perceivers to better capture the duration information of visual input, in this experiment, the syllables used are the monophthong syllables /a/ and /i/ instead of CVV syllables (e.g. /bai/) used in the previous experiments. This is because the duration of vowel syllables should better reflect the duration of lexical tones without the influence of varying co-articulations of preceding consonants.

2.5.1 Method

2.5.1.1 Participants

Eighteen native Mandarin speakers (aged: 26.6 ± 4.5 years; 11 females) participated in the experiment. Most of the participants were postgraduate students studying at Bournemouth University. One of them was left-handed, the rest were right-handed. None had reported a known hearing impairment, and their vision was normal or correct to normal. All participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board.

2.5.1.2 Materials

The stimuli used in this study were two monosyllables /ɑ/ and /i/ with the Mandarin lexical tones T2, T3 and T4 (/á/, /ă/, /à/, /í/, /ĭ/, /ì/). The CVV syllables used in the previous experiments were replaced with vowel-only syllables to reduce the acoustic complexity of syllables. Two male native speakers, from whom two articulations were chosen for each type of syllable, recorded video clips of the syllables. The video clips were edited using Adobe Premiere Pro CC (Adobe Systems, California) at a resolution of 1280 × 720 and a frame rate of 29.97 frames per second. The heads of the speakers (from the top of their heads to the upper part of their necks) were presented in video clips. Auditory tracks were derived from the video and edited using Adobe Audition CC (Adobe Systems, California) with a sampling rate of 44.1k Hz 32 bits saved in WAV format. The noise used was babble noise, which combined six Mandarin native speakers' sentence-reading. The SNR level was set at -9 dB.

In terms of the acoustic features of the lexical tones in the experiment, the duration, F0, intensity and were measured using Praat (Boersman & Weenink, 2013), as presented in Table 2.7. Among the three lexical tones, the tone duration of T3 is the longest and that of T4 is the shortest. Hence, visual syllable duration (from lip opening to lip closure) corresponds with the acoustic duration pattern. In terms of F0, T3 is the lowest pitch and T4 the highest. The intensity levels of the tones are not distinctively different.

Table 2.7 Corpus of Mandarin syllables in Experiment 4. Syllables /a/ and /i/ with T2, T3 and T4 with meanings, acoustic feature measurement: F0, intensity and duration, and visual mouth movement duration from opening to closure.

Syllable	Tone	Gloss	Speaker	F0 (Hz)	Intensity (dB)	Audio duration (ms)	Visual duration (ms)
/a/	T2	modal particle	1	120	69	638	1233
			2	129	69	754	1033
	T3		1	106	66	750	1300
			2	116	66	764	1033
	T4		1	133	71	386	1100
			2	110	69	533	1000
/i/	T2	modal particle/*aunt*	1	134	70	769	1133
			2	125	69	700	1000
	T3	*according to*	1	116	70	835	1167
			2	120	71	916	1067
	T4	*meaning*	1	128	71	579	967
			2	118	71	641	833

2.5.1.3 Procedure

The participants were told to identify the lexical tones of the given syllables by responding to two alternative keys on the keyboard. The experiment comprised two tasks: the two-alternative-choice identification of T2 or T3 and the two-alternative-choice identification of T3 or T4. The task sequence was counterbalanced for each participant. The syllables were presented in AO and AV modalities in the conditions of clear and babble noise at SNR -9 dB level, which were presented randomly within each block. The procedure of the trial was as follows: Frist, the participants saw a fixation cross at the centre of the monitor lasting for 1,000 ms, after which they heard or watched a syllable. They were given a maximum of 3,000 ms to respond to the correct tone of the syllable by pressing the corresponding key on the keyboard. The ITI was consistent at 1,000 ms. Each task contained three blocks, and the participants were able to take a break after each block. The modalities and listening conditions were randomised within the blocks. The total number of trials was 320 for each task, including 2 tones (T2, T3 or T3, T4) × 2 syllables (/ɑ/, /i/) × 2 tokens × 2 speakers × 2 listening conditions (clear, noise) × 2 modalities (AO, AV) × 5 repetitions. The participants were instructed to react to the stimuli as quickly as possible, and they were not informed which specific part of the articulating face they should watch. All stimuli were presented in E-prime 2.0 (Psychology Software Tools, Sharpsburg) on a PC, and auditory sounds were played via noise-cancelling headphones, Sennheiser HD 280 (Sennheiser electronic GmbH & Co. KG, Wedemark), the output volume of which remained at approximately 65 dB SPL.

2.5.2 Results

The analysis comprised three parts: the first two ANOVAs examined the audiovisual benefit effect in the T2/T3 and T3/T4 tasks, respectively, with dependent variables: identification accuracy and reaction time (RT). The audiovisual benefit in RT indicates that the RT of AV lexical tones is shorter than that of AO lexical tones. A separate ANOVA was subjected to a comparison of the audiovisual benefit difference between the two tasks.

Table 2.8 Average accuracy (%) and RT (ms) (*N* = 18) of T2/T3 and T3/T4 tasks.

	Condition	**Modality**	**T2/T3**				**T3/T4**			
			T2	**SE**	**T3**	**SE**	**T3**	**SE**	**T4**	**SE**
Accuracy	**Clear**	**AO**	97.0	0.5	97.7	0.7	97.4	0.7	96.0	0.9
		AV	96.3	0.7	94.9	1.1	96.3	1.0	96.7	1.0
	Noise	**AO**	83.3	3.0	41.5	4.0	66.5	5.5	67.3	3.6
		AV	84.1	4.1	47.3	4.7	76.1	3.7	73.6	2.8
RT	**Clear**	**AO**	898	27	895	26	860	35	811	29
		AV	853	33	885	29	799	34	732	34
	Noise	**AO**	1022	56	1169	71	1149	54	934	38
		AV	940	37	1066	50	1019	47	888	43

Identification of T2 and T3

In the analysis of identification accuracy in the task T2/T3, a three-way ANOVA with the factors Modalities (AO, AV), Listening condition (clear, noise) and Tone (T2, T3) was conducted. The results showed that the main effect of Modality was not significant: $F(1, 17) = .94, p = .35$. The main effects, Listening condition [$(F(1, 17) = 257.83, p < .001)$] and Tone [$(F(1, 17) = 52.98, p < .001)$] were significant. The interaction of Modality × Listening condition was significant: $F(1, 17) = 6.48, p = .021$. Listening condition × Tone was also significant: $F(1, 17) = 51.86, p < .001$. However, the interaction effect of Modality × Tone was not significant: $F(1, 17) = .187, p = .67$. The three-factor interaction effect was not significant either: $F(1, 17) = 1.09, p = .31$.

A pairwise comparison of Modality based on Listening condition revealed that AO accuracy (97.3 ± 0.5%) was significantly better than AV accuracy (95.6 ± 0.7%) under the clear condition, and there was a marginally significant effect whereby AV (65.7 ± 2.5%) was better than AO accuracy (62.4 ± 2.4%) under the noise condition ($p = .072$). This suggests that the audiovisual benefit effect was in fact weak in noise conditions, and there was even a trend whereby extra visual information inhibited the lexical tone identification of T2 and T3 (see Figure 2.17).

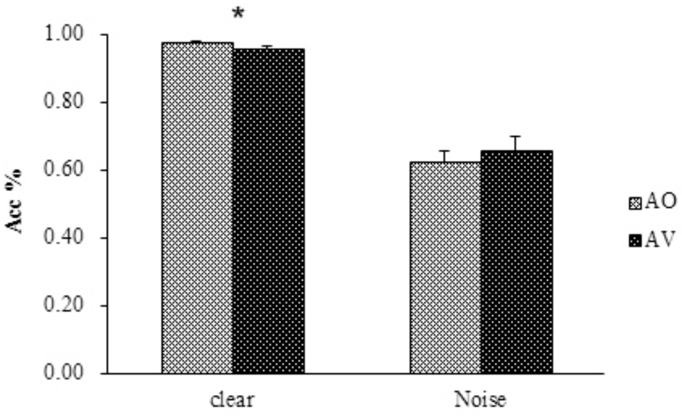


Figure 2.17 T2/T3 identification of AO and AV in clear and noise conditions. *: $p < .05$.

Even though the three-way interaction did reach a significant level, a further pairwise comparison of Tone showed that T3 primarily caused the visual inhibition effect, in which T3 in AO (97.7 ± 0.7%) was better than T3 in AV (94.9 ± 1.1%) ($p = .004$). In the noise condition, there was a slight trend towards an audiovisual benefit effect of T3 identification, despite the fact that it was not significant ($p = .13$) (see Figure 2.18). Additionally, T2 accuracy (83.7 ± 3.1%) was significantly higher than T3 accuracy (44.4 ± 4.0%) in both AO and AV modalities, which suggested that T3 was largely misperceived as T2 in noise.

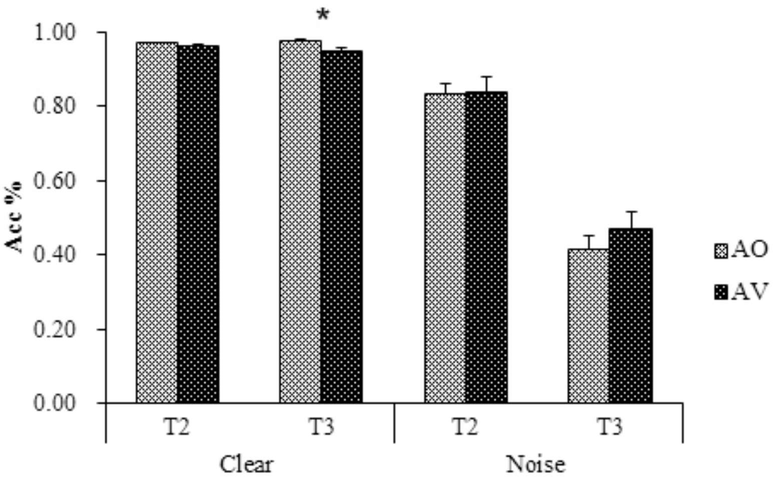


Figure 2.18 T2 and T3 identification in AO and AV modalities in clear and noise conditions. *: $p < .05$.

The RT analysis adopted the same three-way ANOVA as the accuracy analysis. The results showed the main effect of all factors was significant: Modality, $F(1, 17) = 29.62, p < .001$]; Listening condition, $F(1, 17) = 24.619$, $p < .001$; Tone, $F(1, 17) = 24.28$, $p < .001$. The two-way interaction effect of Listening condition × Tone was significant: $F(1, 17) = 15.84$, $p = .001$. Additionally, Modality × Listening condition was marginally significant: $F(1, 17) = 3.59$, $p = .075$. The other interaction effect was not significant. The pairwise of Modality further showed that the RT of AV (936 ± 34 ms) was significantly shorter than that of AO (996 ± 41 ms). This finding suggested that adding visual information speeded up the processing time in the clear and noise conditions – i.e. an audiovisual benefit effect existed. Although Modality and Listening condition demonstrated a weak interaction, the audiovisual benefit in RT was more significant in the noise condition (*p* = .003) than in the clear condition (*p* = .014) (see Figure 2.19).

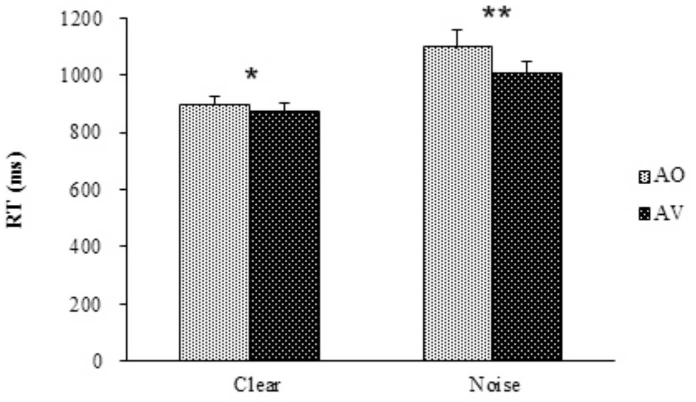


Figure 2.19 RT of T2/T3 identification of AO and AV modalities in clear and babble noise conditions. *: *p* < .05; **: *p* < .01

Identification of T3 and T4

In the analysis of identification in the T3/T4 task, the same three-way ANOVA [Modalities (AO, AV) × Listening Conditions (clear, noise) × Tones (T3, T4)] was performed. The main effects of Modality [*F*(1, 17) = 14.12, *p* = .002] and Listening condition [*F*(1, 17) = 148.51, *p* < .001)] were significant. The interaction effect between Modality and Listening condition was also significant: *F*(1, 17) = 12.43, *p* = .003. All other interaction effects were not significant. The comparisons of Modality revealed that AV accuracy (74.9 ± 2.0%) was clearly better than AO accuracy (66.9 ± 2.6%) in the noise condition (*p* = .001), but not in the clear condition (*p* = .819) (see Figure 2.20).

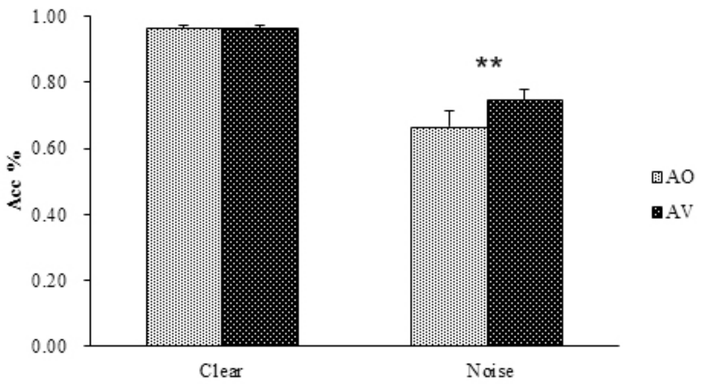


Figure 2.20 Lexical tone identification in the T3/T4 task in AO and AV modalities in clear and noise conditions. **: *p* < .01

A separate three-way ANOVA for RT in the T3/T4 task showed that the main effects of Modality [*F*(1, 17) = 21.61, *p* < .001], Listening condition [*F*(1, 17) = 96.21, *p* < .001)] and Tone [*F*(1, 17) = 94.19, *p* < .001], reached significance. The two-way interaction of Listening condition × Tone was significant: *F*(1, 17) = 17.80, *p* = .001. The three-way interaction also yielded a significant effect: *F*(1, 17) = 7.67, *p* = .013. The pairwise comparison of Modality revealed that an audiovisual benefit effect was found for both tones in both clear and noise conditions (see Figure 2.21).

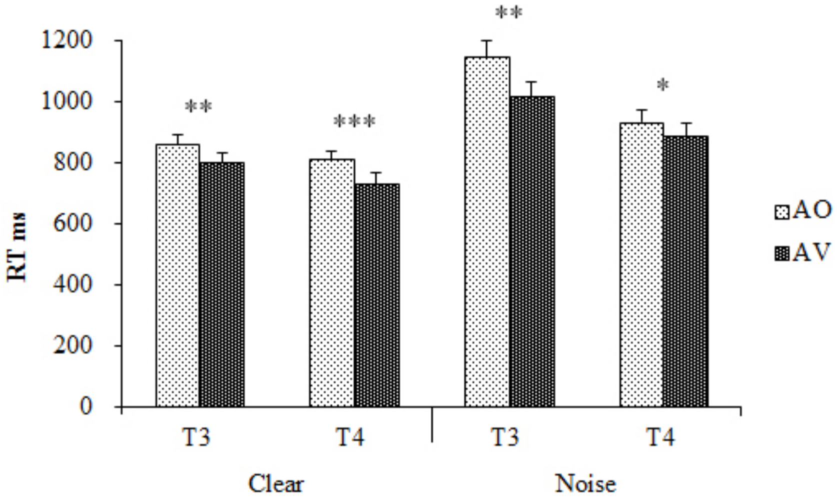


Figure 2.21 RT of T3/T4 identification of AO and AV modalities in clear and noise conditions. *: *p* < .05; **: *p* < .01; ***: *p* < .001

According to the accuracy analysis above, an audiovisual benefit effect was observed in the T3/T4 task, but not in the T2/T3 task. T3 was involved in both tasks, but the identification was not identical. The audiovisual benefit effect on T3 was found in the T3/T4 task, but it was not observed in the T2/T3 task. Additionally, in the noise condition, regardless of AO or AV, the identification rate of T3 was lower in the T2/T3 task than in the T3/T4 task. This indicates that alternative tones in the task have an influence on tone perception.

Comparison of two tasks

To further compare the audiovisual benefit effect of two tasks, a two-way repeated measures ANOVA with the factors Task (T2/3, T3/T4) and Listening condition (clear, noise) was applied to the audiovisual benefit value (AV – AO). The main effects of the two factors were significant: Task: $F(1, 17) = 6.08, p = .025$; Listening condition: $F(1, 17) = 23.91$, $p < .001$. The interaction was not significant: $F(1, 17) = 0.98$, $p = .337$. The comparisons of Task confirmed that the audiovisual benefit effect was stronger in the T3/T4 task (3.9 ± 1.1%) than that in the T2/T3 task (0.8 ± 0.8%) (see Figure 2.22).

The same three-way ANOVA was applied to RT to compare the audiovisual benefit of both tasks. The results showed that the RT audiovisual benefit effect of the two tasks was not significantly different in any of the conditions.

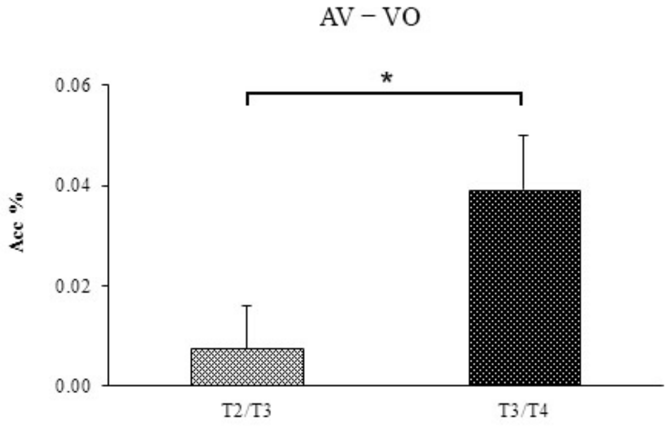


Figure 2.22 Audiovisual benefit effect (AV – AO) in the T2/T3 and T3/T4 tasks. *: $p < .05$

2.5.3 Summary of the results

Identification rate analysis found that the AV benefit effect was not identical in the T2/T3 and T3/T4 tasks. Only T3/T4 identification showed that lexical tones benefited significantly from presenting visual input in the noise condition, whereas T2/T3 identification showed that lexical tones had no significant audiovisual benefit effect in noise. A minor effect of audiovisual inhibition was found in T2/T3 identification in the clear condition but not in the noise condition. A comparison between the two tasks revealed that T3 and T4 had a stronger audiovisual benefit effect than T2 and T3.

The results for the T2/T3 task demonstrated that barely any audiovisual benefit effect was found in T2 in any of the conditions. However, visual input influenced T3 perception in two ways. In the clear condition, T3 perception had an audiovisual inhibition effect, but in the noise condition, T3 perception showed a marginal but significant audiovisual benefit effect. The T2 and T3 identification rates in the T2/T3 task were not the same. T2 was consistently much higher than T3 identification, regardless of AO or AV, in which T3 identification was only at about chance level (50%). It suggests that T3 was misperceived as T2 in the T2/T3 task, which could well be due to the confusing nature of T2 and T3. Auditory T2-T3 was recognised as a most confusing pair, sinceT2 and T3 have similar pitch contours. In the noise condition, when the acoustic feature F0 was severely degraded, the differentiation of T2 and T3 significantly decreased. By adding extra visual inputs, T3 identification was slightly improved, but the situation of T3 being misperceived as T2 was not effectively relieved. This suggests that visual information did not facilitate T3 being distinguished from T2.

In terms of the results in the T3/T4 task, both T3 and T4 had similar strengths of audiovisual benefit effect in the noise condition (see Table 2.8). Unlike the T2/T3 task, T3 and T4 identification were very close to each other, regardless of whether they were in AO or AV modality. This finding suggested that these two tones were less confusing compared to T2-T3. By adding visual input, the accuracy of both T3 and T4 was significantly improved. This suggests that visual information is useful in distinguishing T3 and T4.

T3 was identified in both tasks; however, the degree of the audiovisual benefit effect on T3 perception was different in the two tasks. The results strongly indicate that making use of visual cues for lexical tones could be more sensitive to visual feature contrast between the given two tones. In other words, the degree of the audiovisual benefit effect is greater if the visual features contrast of two presented tones is more distinctive, e.g. T3 and T4. On the other hand, the degree of the audiovisual benefit effect is weaker or even reversed if the visual feature contrast of the two tones is not distinctive, e.g. T2 and T3.

Interestingly, the RT audiovisual benefit was not different across the two tasks. The audiovisual benefit effect of RT did not seem to be affected by the response alternatives within the task. Additionally, an RT audiovisual benefit effect appeared in both clear and noise conditions, hence it was less affected by ceiling effect. The RT benefit effect could be more independent of the distinctiveness of visual information (duration information). Compared with the audiovisual benefit effect of accuracy, the RT audiovisual benefit effect might be a general characteristic of audiovisual speech perception, which may reflect the pre-linguistic processing of audiovisual speech integration.

2.6 Discussion of Experiments 1–4

Audiovisual benefit effect of general lexical tone perception

First of all, the main findings of Experiments 1–4 showed that an audiovisual benefit effect was found in both Mandarin lexical tone identification and discrimination (see Table 2.9), which supports visual information facilitating lexical tone perception, especially in the noise condition. This is also is consistent with previous studies (Burnham et al., 2001; Burnham et al., 2011; Mixdorff et al., 2005a; Mixdorff et al., 2005b; Mixdorff et al., 2006; Smith & Burnham, 2012).

In noise, the audiovisual benefit effect was greater in stronger auditory masking (noisier) conditions. For example, it was stronger in babble noise than in pink noise, and it was stronger in SNR -9dB, SNR -12 dB conditions than in SNR -6 dB condition. However, the audiovisual benefit did not increase along with decreasing SNR. There was no statistical difference between the SNR -9 and SNR -12dB conditions, and the audiovisual benefit in SNR -9 dB appeared to be slightly higher than that in SNR -12 dB in the identification rate. As Ross et al. (2007) propose, the intermediate SNR level could be a ‘special zone’ for optimal audiovisual sensory integration, and there seems to be a certain favourable noise level for maximal audiovisual benefit effect in lexical tones. In their study, SNR -12 dB was the favourable noise level. In the current experiments, the maximal audiovisual benefit for lexical tones could be around SNR -9 dB with babble noise. The varying degree of the audiovisual benefit effect across noise conditions may be due to different audiovisual cue weighting strategies the participants applied to achieve optimal perception. For lexical tones, auditory cues are more informative and reliable than visual cues. Perceivers relied heavily on auditory cues for lexical tone recognition and used fewer visual cues when auditory cues were still reliable; but, nevertheless, visual cues of tones were captured when auditory cues were considered less reliable.

Audiovisual benefit effect of individual tones

When drilling down into individual lexical tones, there is a consistent pattern for an audiovisual benefit effect for tone contrasts T2-T3 and T3-T4. In Experiments 2–4, the results showed that the audiovisual benefit effect was not even across individual tone types or tone contrasts, and the degree of the audiovisual benefit depended greatly on the responses to alternative tones in a task. Experiments 2–3 found the strongest audiovisual benefit for T3-T4 tone contrast among six tone contrasts, and Experiment 4 confirmed the audiovisual benefit was stronger in T3/T4 identification than in T2/T3 identification. This suggests that the visual feature contrast of T3 and T4 is more distinctive compared to other tone contrasts. These pattern is more consistent with the results that Mixdorff et al. (2005b) reported, where the recognition of T3 and T4 had a stronger audiovisual benefit effect in SNR -12 dB noise. They proposed that T3 and T4 had a more salient visual duration and intensity feature.

Although T3 is regarded as a more visually salient tone, T3 did not always have the same amount of audiovisual benefit across different tasks. The degree of audiovisual benefit was associated with the contrast between alternative tones in a task. The results of Experiment

4 showed that the audiovisual benefit in T2/T3 identification was lower than that in T3/T4 identification. T3 benefited more from visual information as it was perceived against T4 rather than against T2. The results in Experiments 2–3 also showed that T3-T4 discrimination had a greater audiovisual benefit than the other contrasts involved in T3 (e.g. T2-T3). This could be because T3 and T4 were more visually different, while T2 and T3 were more visually similar. The distance of the visual features between the two tones was larger, so the two tones could be more readily discriminated when visual information was available, therefore having a greater the audiovisual benefit effect (e.g. T3-T4). On the other hand, if the two tones share great similarities in visual features, a lower audiovisual benefit effect would be expected (e.g. T2-T3).

Potential visual cues of lexical tones

The perception of lexical tones is obviously affected by visual input along with auditory tones, but it is difficult to determine the exact visual cues used during perception based on the data from the current experiments. Lip-reading performance indicates that phonetic-/ tonetic-specific visual cues exist, since the identification/ discrimination of VO tones was better than chance level. However, the degree of the audiovisual benefit effect for T2-T3 and T3-T4 suggests that possible phonetic-specific visual cues could be features that are highly contrastive between T3 and T4, but much less distinguishable between T2 and T3.

Both tone duration and F0 movement seem to be the most likely potential candidates for transferring key information for lexical tones to articulating mouth movements. Compared to F0 movement, the duration is more likely to be mirrored by the period of mouth movement from lip opening to lip closure. A shorter acoustic tone (e.g. T4) also has a shorter mouth opening time. A longer tone (e.g. T3) naturally requires a longer time for mouth movement. Acoustic duration has been regarded as a secondary cue for tone perception in an adverse condition where F0 is less available (Liu & Samuel, 2004). A duration cue might be extracted from mouth movement to facilitate tone perception.

Table 2.9 Summary of the results of Experiments 1–4

Exp	Task	N	Token(s)	Speaker (s)	Syllable (s)	Noise	Main findings
1	Identification	28	1	1	/bai/, /dai/	pink noise babble noise SNR -6 dB, -9 dB and -12 dB	1. An audiovisual benefit effect for lexical tone identification was found in the following noise conditions: pink, babble, SNR -9 and SNR -12 dB; 2. The audiovisual benefit effect was not different across individual tones; 3. VO was higher than chance level.
2	Cross-modality discrimination	20	1	1	/bai/	babble noise SNR -6 dB and -9 dB	1. An audiovisual benefit of general tone discrimination was found in clear but not noise condition; 2. An audiovisual benefit effect was found in T3-T4; audiovisual inhibition was found in T2-T3; 3. VO was higher than chance level; T2-T4 and T3-T4 were higher than other contrasts.
3	Within-modality discrimination	20	1	2	/bai/	babble noise SNR -6 dB and -9 dB	1. An audiovisual benefit of general tone discrimination was not found in noise condition; 2. An audiovisual benefit effect was found in T3-T4; audiovisual inhibition was found in T2-T3; 3. VO was higher than chance level at SNR -6 dB.
4	Identification of T2/T3; identification of T3/T4	18	2	2	/a/, /i/	babble noise SNR -9 dB	1. An audiovisual benefit effect was shown in T3/T4 identification, but not in T2/T3 identification in noise condition; 2. Audiovisual inhibition was shown in T2/T3 identification in clear condition; 3. An audiovisual benefit of RT was found in all conditions for both tasks.

However, tone duration as a visual cue cannot fully explain the stronger audiovisual inhibition effect on T2-T3 in Experiments 2–3. Among the six tone contrasts, T2 and T3 were not the most similar pair in duration. T1 and T3 were even closer to each other in duration (see Table 2.5). If tone duration was the major visual cue, one would expect the tone contrast T1-T3 to have the lowest audiovisual benefit or even audiovisual inhibition effect. In Experiments 2–3, an audiovisual inhibition effect was repeatedly found in tone contrast T1-T3 in the SNR -9 dB condition, but it was not as strong as that in T2-T3. This suggests that visual duration was not the only visual cue influencing lexical tone perception; other visual cues might also be involved, such as F0 movement, but whether acoustic F0 movement can be reflected in mouth movement is still unclear.

Another role that visual information might play is to provide timing information for auditory tones, so that auditory tone information can be more effectively captured. Visual speech cues can function as temporal markers pre-cueing the auditory signal in noise condition, which allocates perceivers' attention to target speech, therefore increasing perceivers' sensitivity to acoustic features in auditory speech (Bernstein et al., 2004; Eskelund et al., 2011; Grant & Seitz, 2000; Kim & Davis, 2004; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004). According to certain non-speech studies, the ability to predict when stimuli are presented is crucial for detection. Egan et al. (1961) and Watson and Nichols (1976) reported that auditory events in noise condition were perceived better when presented regularly compared to those presented irregularly. Visual cues help timing prediction. ten Oever et al. (2014) found that detection sensibility was higher for auditory events presented irregularly with preceding visual cues compared to auditory-only events.

Mouth movements can be an indicator for the onset and offset time of acoustic tones, which not only provides tone duration information but also increases listeners' sensitivity to the acoustic tones in noise. Consequently, it enables them to better hear the tone signal in a noise environment. This can explain why T2 and T3 were greatly perceived as the same tone in the AV modality. In the case of T2 and T3, they are known to be an easily confusable pair in the auditory modality. When visual information enhances auditory detection, the confusion between T2 and T3 cannot not be relieved and might even be increased. In the AV modality, T2 and T3 could become a tone pair that does not only look alike but also sounds alike; therefore, their discrimination would benefit less from visual information or could even be biased because of it.

Additionally, the lip-reading performance of lexical tones also indicates that there could be certain visual cues other than phonetic-specific ones. The pattern of the audiovisual benefit effect across individual tones is not completely consistent with the VO performance of individual tones, which means that there are cues used in the audiovisual perception of lexical tones that are different from those (phonetic-specific cues) used in VO perception.

In sum, Experiments 1-4 found evidence that visual information does facilitate auditory lexical tone perception. Useful cues from visual information influence audiovisual lexical tone perception and could be associated with phonetic-/ tonetic-specific visual cues (e.g. tone duration, F0 movement). Furthermore, there is a possibility that non-phonetic information may also contribute to lexical tone perception (e.g. visual information as an indicator of the appearance of the auditory tone signal). Non-phonetic visual information may always be used in audiovisual speech perception, whereas tone-specific visual cues can be effectively captured when they are highly distinguishable between alternative tones like the tone contrast between T3 and T4. However, due to the limitations of the experimental design, the effect of two types of audiovisual integration could not be directly observed from the results. In order to explore further the possible processing of non-phonetic/ phonetic visual information, in the next experiment the brain activity of the audiovisual lexical tone perception is measured, where audiovisual lexical tone perception can be more directly observed in real time, which could help to address the unanswered questions in this chapter.

Chapter 3 Neural correlates of the audiovisual benefit effect in lexical tone processing

3.1 Introduction

In this chapter, the main goal is to investigate whether visual information has an influence on the early processing of auditory lexical tones in the brain. Specifically, the experiments test whether the ERP evoked by audiovisual Mandarin lexical tones demonstrate a reduction effect (smaller amplitude and shortening latency) in the early auditory components N1 and P2 compared to the ERP evoked by auditory-only lexical tones.

From the literature, presenting visual speech with auditory speech reduces and accelerates the early components N1 and P2 after auditory onset (Besle et al., 2004; Huhn et al., 2009; Klucharev et al., 2003; Knowland et al., 2014; Pilling, 2009; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005), which is interpreted as visual information that facilitates early processing of auditory speech in the brain. According to the behavioural findings of Experiments 1–4, visual information facilitates auditory tone perception, but behavioural responses are unable to reveal the time course of the visual facilitation processing. The behavioural audiovisual benefit effect of lexical tones seems to be linked to tone duration cues from visual lexical tone input, but it does not necessarily rule out cues that are non-phonetic/ tonetic (e.g. timing information) from visual input facilitating auditory speech perception. These questions can be answered by measuring the ERP responses of audiovisual lexical tones.

Methodologically, the current experiments employ an additivity model which has been applied to several audiovisual ERP studies in speech (Besle et al., 2004; Klucharev et al., 2003; Pilling, 2009; Stekelenburg & Vroomen, 2007). Specifically, if the audiovisual response is not equivalent to the sum of the response of the auditory and visual modalities (AV ≠ A+V), it indicates that audiovisual interaction occurs. Previous studies have shown that subadditivity – audiovisual potential is smaller than the sum of auditory and visual responses in the early auditory components N1 and P2 (AV < A+V) (Besle et al., 2004; Klucharev et al., 2003; Pilling, 2009; Stekelenburg & Vroomen, 2007).

Experiment 5 compared the difference in ERP responses of AV lexical tones and VO lexical tone (AV – VO) with the ERP of AO responses in terms of N1 and P2 after the onset of the auditory speech signal. If the difference response (AV – VO) is smaller in amplitude and earlier in latency compared to the AO response in N1 and P2 components (N1/P2 reduction effect in audiovisual speech), it indicates that visual information facilitates early auditory lexical tone processing. In addition to finding out whether lexical tones have an N1/P2 reduction effect, the experiments also examined whether the saliency of visual speech information was associated with the N1/P2 reduction effect; therefore, the ERP of audiovisual consonants was also measured in a separate experiment (Experiment 6). In this study, the results of lexical tones and consonants were compared in a between-subject design, where participants in the two experiments did not overlap. Consonants have a more salient visual feature (place of articulation) which occurs about 100 ms preceding the corresponding auditory signal, whereas lexical tones are less visually distinguishable, and the preceding mouth movement does not convey tone-specific information. If the N1/P2 reduction effect in audiovisual speech processing depends on the visual saliency or lip-read information from preceding mouth movement, the N1/P2 reduction effect of lexical tones should be different from that of consonants. Because the visual information of consonants is more salient than that of lexical tones, the N1/P2 reduction effect should be stronger in consonants than lexical tones. If, on the other hand, N1/P2 reduction relies on the visual prediction of the timing of auditory sound, lexical tones should be similar to consonants in the N1/P2 reduction effect.

3.2 Method

3.2.1 Participants

Thirty-nine native Mandarin speakers participated in two experiments. Twenty participants (aged 25.8 ± 4.4 years; 13 females) participated in lexical tone experiment (Experiment 5), and 19 participants (aged: 26.6 ± 5.7 years; 11 females) took part in consonant experiment (Experiment 6). All of them were right-handed. One participant from the tone experiment and two from the consonant experiment were excluded from the data analysis due to heavy artefacts (e.g. alpha brainwaves). All participants reported normal or correct to normal visual acuity and had no previous hearing impairments. The participants were compensated in accordance with a protocol approved by Bournemouth University Review Board, and every participant received a payment as a reward.

3.2.2 Materials

The stimuli were six CVV Mandarin monosyllables (/bai/, /dai/, /tai/, /bao/, /dao/ and /tao/) with four lexical tones (6 × 4 = 24 syllables), as in Table 3.1, which were presented in AO, AV and VO. These syllables were produced by two native male Mandarin speakers, respectively. The recorded videos were edited with Adobe Premiere Pro CC (Adobe Systems, California) as video clips with a resolution of 1280 × 720 and a digitisation rate of 59.94 frames per second (1 frame = 16.68 ms). The sound tracks of the videos were edited in Audacity (Crook, 2012) and Adobe Audition CC (Adobe Systems, California). All auditory tracks were digitised at 48,000 Hz, with a 32-bit amplitude resolution, and were RMS normalised to -12 dB. The duration of AV stimuli was set to 1367.76 ms (82 frames), in which the front silence-gap between video onset and audio onset was 233.52 ms (14 frames) (see Figure 3.1). Syllable durations varied, and the silent gap from audio offset to video offset consequently varied as well. The trigger was placed at the video onset time for data processing. AO and VO stimuli were derived from the AV clips and kept identical in duration.

Table 3.1 Corpus of Mandarin syllables in Experiments 5–6. Six syllables /bai/, /dai/, /tai/, /bao/, /dao/ and /tao/ with four tones with meanings for acoustic feature measurement: F0, intensity and duration, and visual mouth movement duration from opening to closure.

Syllable	Tone	Gloss	Speaker	F0 (Hz)	Intensity (dB)	Audio duration (ms)	Visual duration (ms)
/bai/	T1	*to break*	1	124	66	864	1285
			2	136	65	1101	1133
	T2	*white*	1	116	69	790	1134
			2	129	64	985	1200
	T3	*to display*	1	95	62	957	1301
			2	112	62	1139	1317
	T4	*to defeat/failed*	1	118	65	438	1084
			2	140	66	305	700
/dai/	T1	*to stay/dull*	1	128	68	689	951
			2	139	64	887	1134
	T2	*NA*	1	118	68	755	1034
			2	116	63	828	1134
	T3	*to seize/bad*	1	96	65	860	1084
			2	105	61	965	1168
	T4	*to bring/belt*	1	125	68	432	901
			2	138	65	317	901
/tai/	T1	*tyre/foetus*	1	130	67	715	901
			2	137	65	958	1051
	T2	*to lift up*	1	113	66	915	1168
			2	117	66	966	1168
	T3	*NA*	1	97	65	982	1185
			2	113	63	1122	1251
	T4	*extremely/ peaceful*	1	121	67	511	934
			2	99	65	493	1051
/bao/	T1	*Bag/to wrap*	1	122	71	682	1018
			2	128	69	824	918
	T2	*thin*	1	112	70	669	1084
			2	112	69	670	918
	T3	*full/treasure*	1	97	68	760	1101
			2	115	64	827	968
	T4	*to embrace/ to report*	1	116	69	441	968
			2	109	68	249	667
/dao/	T1	*knife*	1	126	70	666	951
			2	131	68	768	951
	T2	*to smash*	1	114	70	670	968
			2	115	67	592	1134
	T3	*island/fall*	1	99	68	781	1034
			2	119	62	867	1034
	T4	*to arrive/way*	1	117	68	420	851
			2	134	69	267	584
/tao/	T1	*big waves*	1	130	70	676	934
			2	132	67	829	1084
	T2	*to escape*	1	116	70	730	1068
			2	118	69	833	1218
	T3	*to demand*	1	98	68	783	1151
			2	105	64	987	1285
	T4	*cover/knot*	1	110	69	540	901
			2	100	67	362	851

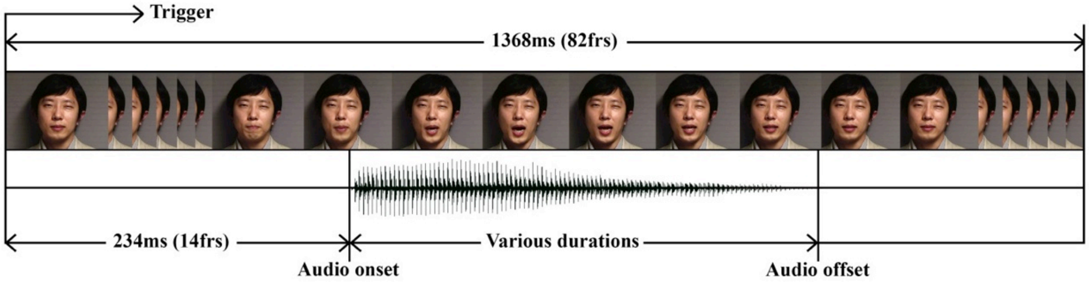


Figure 3.1 Trial structure in Experiment 5. The duration of the AV stimuli was consistently 1367.76 ms (82 frames). Two silent gaps existed at the beginning and end of the video clips. The front-gap was fixed at 233.52 ms (14 frames) after video onset, and the syllable duration and back-gap duration varied. The trigger was placed at the video onset time.

The experiment was conducted in a sound-attenuated dimly-lit chamber. The participants sat facing a 17-inch CRT monitor at a viewing distance of 70 cm. Sounds were played through two Genelec 8030A loudspeakers (Genelec Oy, Iisalmi) placed either side of the monitor. The loudness of the presented sound was approximately 65 dB SPL. The experimental stimuli were presented using E-prime 2.0 (Psychology Software Tools, Sharpsburg).

3.2.3 Procedure and EEG recording

A same-different discrimination paradigm was applied to lexical tone and consonant experiments, respectively. In the lexical tone discrimination experiment, the participants were given two syllables successively that could only differ in lexical tones. The consonant experiment followed the same procedure, except that the participants needed to respond to consonants. The syllables in both experiments remained identical. Participants were required to press the spacebar to respond only when they detected that the given two tones/ consonants within each trial were different. As can be seen in Figure 3.2, the first syllable was always presented in AO, and the second one was randomly presented in one of the modalities (AV, AO and VO) within each block. There were 540 trials presented in six blocks and trials of the 'same type' pair account for 80% of the total trial number. Only trials of the 'same type' with correct responses were analysed in the ERP results. Because participants did not need to respond to 'same type' trials, the ERPs evoked by these stimuli could have minimal interference from manual response and response preparation. The participants were instructed to respond to the task as quickly and accurately as possible and to try to refrain from blinking their eyes during presentation of the syllables. They were also instructed not to watch any particular parts of the articulating face in the video clips and instead to focus on the centre of the screen.[3]

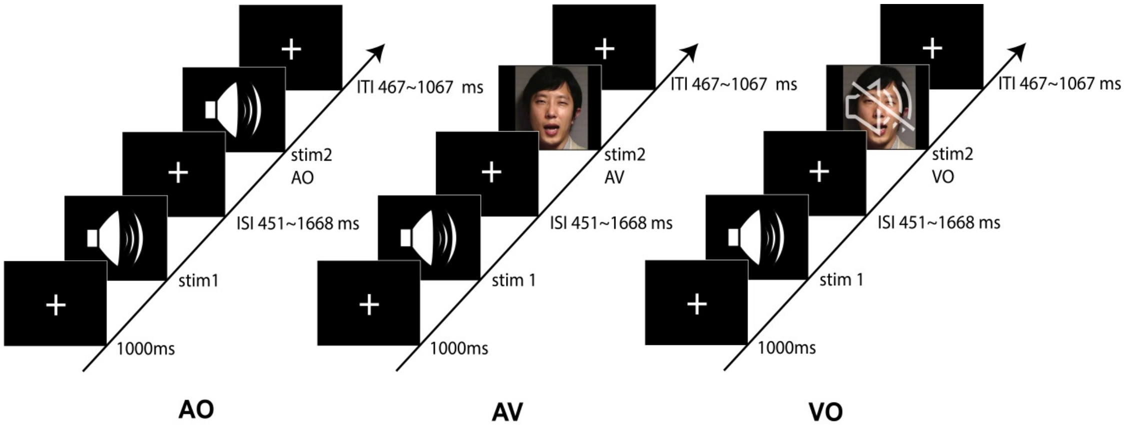


Figure 3.2 Sequential presentation of AO, AV and VO trials in Experiments 5–6. The three modalities were randomly presented within each block. Participants responded to the spacebar on the second stimulus.

EEG recordings were collected at a sampling rate of 500 Hz using a Brain-Amp DC amplifier (Brain Products GmbH, Gilching) and the Brain Vision Recorder 1.0 (Brain Products GmbH, Gilching) system with 32 Ag/AgCl electrodes mounted on a 64-channel elastic cap (actiCap, Brain Products GmbH, Gilching) arranged according to the international 10-20 system (Jasper, 1958). The EEG was recorded by 32 electrodes (Fz, Cz, Pz, FP1/2, AF3/4, F3/4, F7/8, FC1/2, FC5/6, C3/4, T7/8, TP7/8, CP1/2, P3/4, P7/8, PO7/8, O1/2 and right mastoid). The impedance of each channel was kept below 20 kΩ before

[3] A fixation cross at the screen centre appeared throughout the experiments except when AV and VO stimuli were presented. Therefore, the participants were told to look at the fixation cross whenever it was presented and to keep looking at the centre when video clips covered the fixation cross. The mouth region of AV and VO was located at the fixation cross on the screen.

recording. The electrodes were physically referenced to the left mastoid electrode and off-lined re-referenced to the average of left and right mastoids. The ground electrode was placed at AFz.

Raw data were processed via EEGLAB (Delorme & Makeig, 2004) version 14.0.0 and ERPLAB (Lopez-Calderon & Luck, 2014) toolboxes installed in Matlab R2014a (The MathWorks, Inc.). They were filtered offline with a bandpass filter 0.1–30 Hz with 48 dB/oct roll off. The continuous EEG was segmented into 1,200 ms-epochs, starting at 200 ms before the video onset and ending at 1,000 ms after video onset, and the baseline (200 ms pre-stimulus) was corrected. Trials with activities exceeding a voltage threshold of ±100 μV considered as artefacts were rejected by the method of peak-to-peak moving window of 200 ms length with a window step of 100 ms. The data were averaged for each participant.

3.3 Results

The results contain behavioural data and ERP data. The behavioural results reported the DI value (the calculation can be seen in 2.3.1.3) and RT of lexical tone and consonant discrimination tasks. The ERP analysis consists of two ANOVAs for lexical tone and consonant experiments, respectively, and a separate ANOVA for comparing the two experiments. To measure audiovisual integration processing in ERP data, an additivity model was employed for a cross-modality comparison (AO vs AV). The additive model hypothesised that the neural activity of AV stimuli is equal to the sum of A and V if unimodal auditory and visual information are processed separately, i.e. AV = A + V (Barth et al., 1995). If the sum of A and V activity is not equal to bimodal AV activity (sub-additive: AV < A + V; supra-additive: AV > A + V), it indicates there is audiovisual interaction. Therefore, an ERP comparison for the current experiments was carried out between the AO response and auditory response (AO’) evoked by AV stimuli. To attain AO’ ERP, the VO response was subtracted from the AV response. A significant difference between AO and AO’ (AV – VO) suggests audiovisual integration.

3.3.1 Behavioural results

The behavioural results, including the discrimination index (DI) (see Section 2.3.1) and the reaction time for the correct response to 'different-type' stimuli (only 'different-type' stimuli required a response), are presented in Figure 3.3. Two two-way repeated measures ANOVAs were undertaken for the DI and RT data with between-subject factor Task (tone, consonant) and within-subject factor Modality (AO, AV, VO). Greenhouse-Geisser correction (Jennings & Wood, 1976) was applied when the assumption of sphericity was violated. The post hoc comparison was corrected with the Bonferroni procedure.

The DI analysis showed that the main factors, Modality [$F(1.13, 41.91) = 1206.16, p < .001$] and Experiment [$F(1, 37) = 117.52, p < .001$], were significant. The interaction of Modality and Experiment was also significant: $F(1.13, 41.91) = 126.46, p < .001$. A comparison of Modality based on Experiment revealed that AV (0.95 ± 0.10) was higher than AO (0.93 ± 0.12) ($p = .001$) in the lexical tone experiment, while there was no significant difference between AO and AV in the consonant experiment ($p = 0.26$). For both experiments, VO discrimination was significantly the lowest among the three modalities. In terms of the Experiment comparison, the results showed no difference between lexical tones and consonants in both AO and AV modalities. However, consonant discrimination (0.52 ± 0.03) was significantly higher than lexical tone discrimination (0.06 ± 0.03) in VO ($p < .001$).

For the analysis of RT, three participants from the lexical tone task were excluded, as they had no RT recorded in the VO condition. The same ANOVA was performed, and the results showed that the main effects of Modality [$F(1.19, 40.34) = 13.89, p < .001$] and Experiment [$F(1, 34) = 6.68, p = .014$] were significant. The interaction of Modality and Experiment was significant as well: $F(1.19, 40.34) = 6.66, p = .010$. A comparison of Modality on Experiment revealed that there was no significant difference between AO and AV in RT for the lexical tone experiment ($p = .782$). However, they had a significant result in the consonant experiment ($p = .025$), where AV (1156 ± 51 ms) was significantly faster than AO (1223 ± 42 ms). The RT of VO in the lexical tone experiment (1593 ± 85 ms) was much longer than the RT of other modalities, whereas in the consonant experiment VO

(1239 ± 81 ms) was not significantly different from AO or AV. In a comparison of Experiment based on Modality, the consonant RT was much faster than the lexical tone one in VO (*p* = .005).

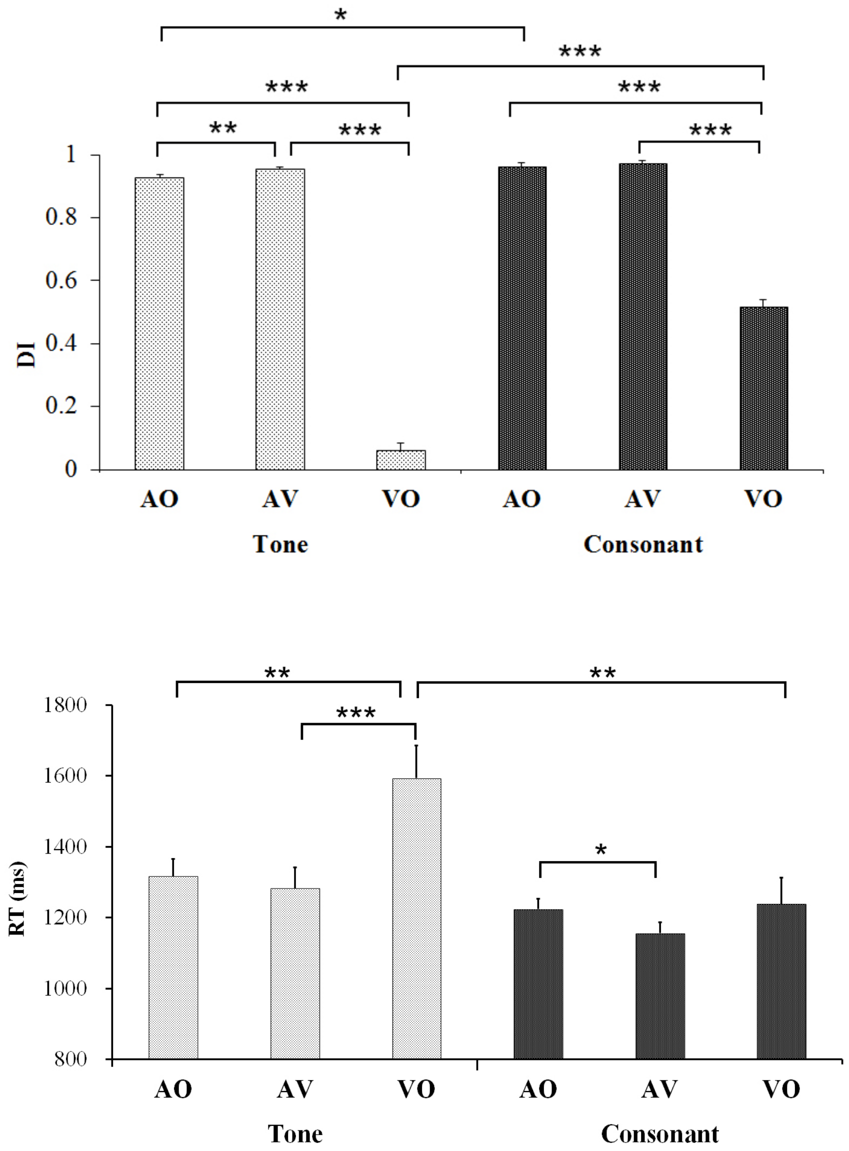


Figure 3.3 DI and RT results for behavioural data in Experiments 5–6. The upper figure presents the DI (*N* = 39) of the AO, AV and VO modalities in the lexical tone and consonant discrimination tasks. The bottom figure presents the RT (*N* = 36) for the AO, AV and VO modalities of lexical tones and consonant discrimination (correct responses to ‘different-type’ stimuli). *: *p* < .05; **: *p* < .01; ***: *p* < .001

In a comparison of the lexical tones and consonant experiments, the performance differed significantly in the VO condition. Consonants had much better discrimination than lexical tones. A certain number of consonant stimuli in VO can be reliably distinguished, and RT was not significantly slower than consonants with an auditory signal. This is because consonants are more readily lip-read with their salient visual features (place of articulation). Except for the consonant contrast between /d/ and /t/, which share a similar place of articulation (alveolar), the other contrasts were highly recognisable. On the other hand, lexical tones were less likely to be recognised with visual input only. In AO vs AV, lexical tone discrimination had an audiovisual benefit effect in the DI result., while consonant discrimination had an audiovisual benefit effect in the RT result. Although consonants were clearly better than lexical tones in VO discrimination, it is difficult to determine which speech has a better audiovisual benefit from the behavioural results in the clear condition, where the ceiling effect exerts a heavy impact.

3.3.2 ERP results

In an ERP comparison between AO and AV modalities, AV was conveyed as AO' generated from different ERPs of AV and VO responses. For both the lexical tone and consonant experiments, two statistical analyses were applied to test the difference between AO and AO' in the N1 and P2 components. A three-way repeated measures ANOVA was subjected to the mean amplitudes and peak latencies of N1 and P2, respectively, with the factors Modality (AO, AO'), Lateralisation (left, right) and Electrodes (selected electrodes vary according to the scalp distribution of the corresponding components). An interval of 156–196 ms after auditory onset time was selected in the N1 analysis and an interval of 226–266 ms after auditory onset time was selected for P2. Another approach was t-max variant of the permutation test (Blair & Karniski, 1993), which compared the responses of the AO and AO' conditions for all electrodes at each time point in selected intervals for lexical tones and consonants, respectively. The t-max variant of the permutation test was conducted using the Mass Univariate ERP Toolbox (Groppe et al., 2011) in Matlab R2014a (The MathWorks, Inc.). A point-to-point comparison of ERP can provide a spatial and temporal resolution of ERP activity, but corrections such as Bonferroni for multiple comparisons would over-reduce the statistical power and lead to a type II error. The advantage of a permutation test is that it can avoid spurious results yet maintain a certain level of statistical power in massive multiple comparisons.

3.3.2.1 Lexical tones (Experiment 5)

Figure 3.4 presents the ERPs evoked by non-response 'same-type' stimuli in the AO, AV and VO modalities from 434 ms before auditory signal onset up to 764 ms after onset at Cz. Before the auditory onset time (0 ms), only video stimuli (e.g. articulating face) were presented.[4] Therefore, the ERPs of AV and VO had very similar visually evoked waveforms before the auditory onset time, and the ERP of AO was at the baseline level, since no video stimuli were presented in this condition. After the auditory onset time, ERPs deviated under different conditions. The AO waveform showed a typical auditory component N1 maximising at about 156 ms after auditory onset, followed by P2 peaking at about 245 ms. In the same time window, a clear auditory N1 can be seen in AV ERP. For the VO waveform, although no auditory stimuli were presented after 0 ms, it seemed to have activity that was evoked at about 196 ms.

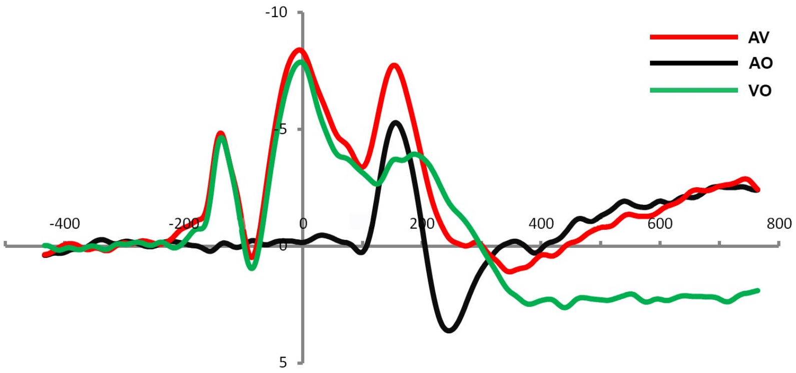


Figure 3.4 Grand average (N = 17) of waveforms recorded at Cz from lexical tone stimuli in the AO, AV and VO modalities.

[4] Because the experiment focused on the auditory components N1 and P2, for clarity and demonstration purposes, the time point of 0 ms in the ERP waveforms presented in the following figures is based on the onset of auditory signals.

In a mean amplitude analysis of the auditory N1 component in the interval 156–196 ms, a three-way repeated measures ANOVA with the factors Modality (AO, AO'), Lateralisation (left, right) and Electrode (F3/4, FC1/2, FC5/6, C3/4, T7/8, CP1/2) was conducted. Twelve electrodes were selected to represent the strong activity of N1 (see Figure 3.5). The symmetry of the selected electrodes contributed to the analysis of lateralisation. Greenhouse-Geisser correction (Jennings & Wood, 1976) was applied when the assumption of sphericity was violated.

The analysis showed that the main effects of Modality [$F(1, 16)$ =16.87, p = .001] and Lateralization [$F(1, 16)$ = 9.21, p = .008] were significant. The main effect of Electrode was also significant: $F(1.60, 25.67)$ = 8.91, p = .002. Any interactions between the factors was not significant. A comparison of Modality showed that the mean amplitude of N1 evoked by AO (-3.48 ± 0.37 μV) was significantly higher than that evoked by AO' (-2.02 ± 0.40 μV). A comparison of Lateralisation showed that N1 activity in the left hemisphere (-2.97 μV ± 0.39 μV) was significantly stronger than that in the right hemisphere (-2.53 ± 0.31 μV) in both AO and AO' (see Table 3.2 and Figure 3.5).

The same ANOVA was applied to the N1 latency data for lexical tones. The results showed that a significant effect of Modality was found: $F(1, 16)$ = 12.70, p = .003. Electrode was also significant: $F(2.30, 36.82)$ = 8.91, p < .001. A comparison of Modality confirmed that the N1 peak latency of AO' (152.8 ± 2.8 ms) was 8.5 ms faster than that of the AO response (161.3 ± 2.1 ms) (see Table 3.2 and Figure 3.5).

For a mean amplitude analysis of P2 component, the time window of 226–266 ms was selected. Since P2 distribution maximised in the central area (see topography in Figure 3.5), five electrodes representing strong positive activity were selected: FC1, FC2, Cz, CP1 and CP2. A two-way repeated measures ANOVA was subjected to a P2 amplitude comparison of AO and AO' with the factors Modality (AO, AO') and Electrode (FC1, FC2, Cz, CP1, CP2). The results showed that the main effect of Modality was significant: $F(1, 16)$ = 11.75, p = .003; the main effect of Electrodes was also significant: $F(4, 64)$ = 14.60, p < .001. The interaction between Modality and Electrode was marginally significant: $F(2.83, 45.33)$ = 2.74, p = .077. A modality comparison revealed that the P2 amplitude of AO' tone response (1.23 ± 0.63 μV) was smaller than that of the AO response (2.55 ± 0.46 μV) (see Table 3.2 and Figure 3.5).

The same ANOVA was performed for P2 latencies, and a marginally significant effect was found for Modality: $F(1, 16) = 3.80$, $p = .069$. The rest of the effects failed to reach significance. A modality comparison confirmed that the P2 peak latency of AO' (240.7 ± 6.0 ms) was 7.3 ms faster than that of AO (248.1 ± 4.6 ms) (see Table 3.2 and Figure 3.5).

Table 3.2 Average ERP amplitude and latency ($N = 17$) in AO and AO' (AV – VO) modalities in the lexical tone experiment.

Task	Modality	N1				P2			
		Amplitude (μV)	SE	Latency (ms)	SE	Amplitude (μV)	SE	Latency (ms)	SE
Lexical Tone	AO	-3.48	0.37	161.3	2.1	2.55	0.46	248.1	4.5
	AO'	-2.02	0.40	152.8	2.8	1.23	0.63	240.7	6.0
	AO-AO'	-1.46	0.35	8.5	2.4	1.33	0.39	7.3	3.8

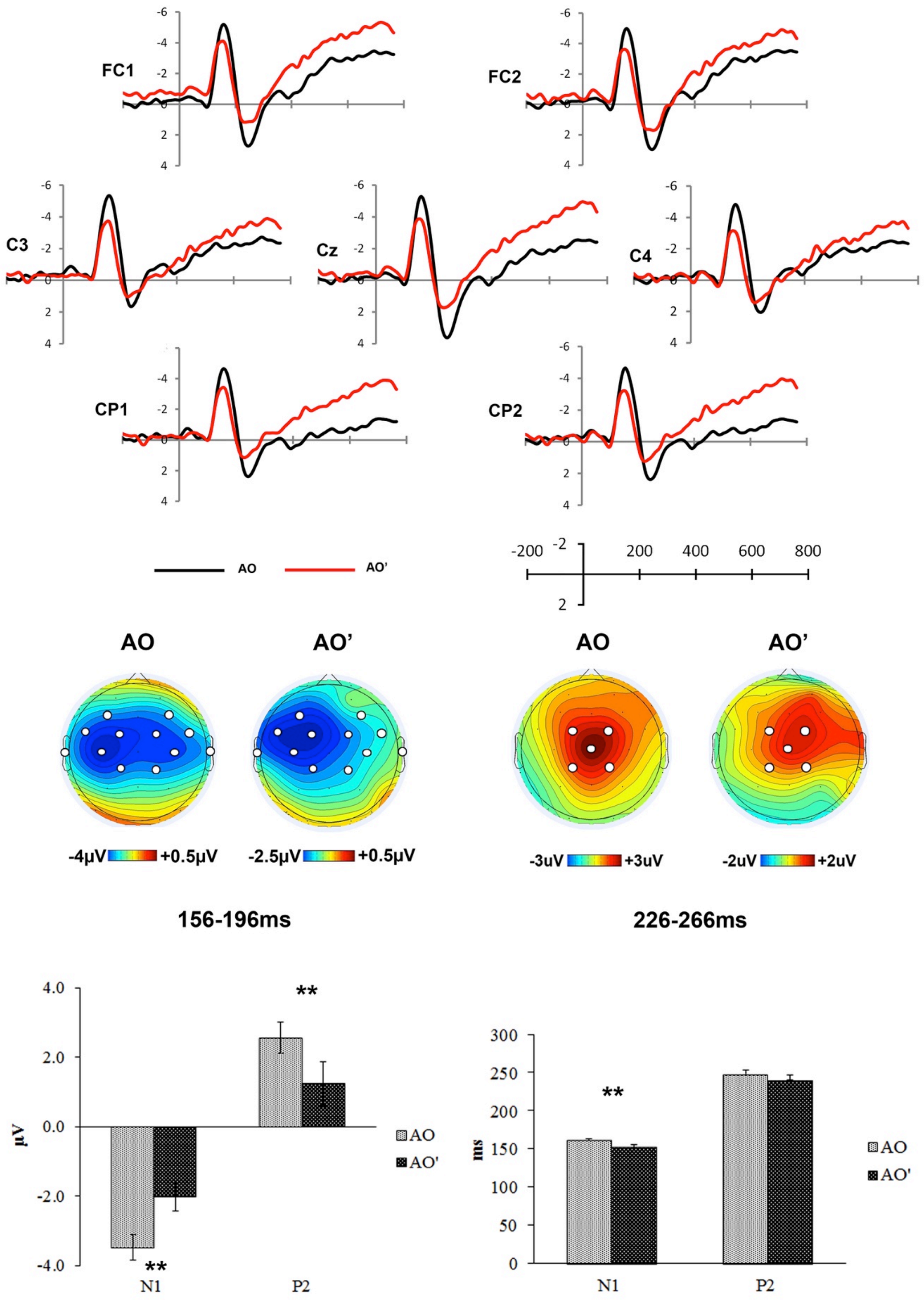


Figure 3.5 Average ERPs of seven electrodes (FC1/2, C3/4, Cz, CP1/2) of AO and AO’ in the lexical tone experiment. The topographies represent the voltage distribution of the time windows 156–196 ms (N1) and 226–266 ms (P2). The electrodes used in the analysis were marked as white dots. The bottom bar charts represent a comparison between AO and AO’ in mean amplitude (left) and peak latency (right) for N1 and P2, respectively. **: $p < .01$.

A t-max permutation test for a point-to-point comparison between AO with AO’ ERP was conducted for every electrode within the same time intervals 156–196 ms (N1) and 226–266 ms (P2), respectively. As can be seen in Figure 3.6, the negativity of AO – AO’ was significant at C3, F4, FC6, FC2, C4, T8 and TP8 within the N1 time window. The electrodes where comparisons had a significant effect were distributed in the typical auditory cortex area of the scalp, and most of them were located in the right hemisphere. Regarding the result for the P2 interval, significant positivity for AO – AO’ appeared at Cz and FC1.

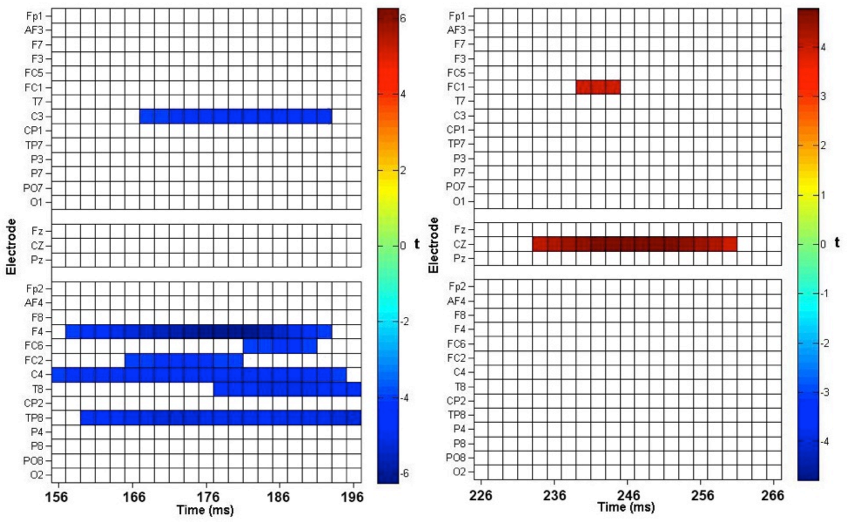


Figure 3.6 Lexical tone results of t-max permutation tests for the difference between AO’ and AO in the intervals 156–196 ms (N1) and 226–266 ms (P2) in lexical tone discrimination. The time-points having a significant effect are colour-coded to represent a negative direction or a positive direction ($p < .05$).

Based on the results for the lexical tones above, a significant N1 reduction effect in amplitude (within 156–196 ms) and in peak latency was found in the fronto-central area of the scalp. A significant P2 reduction effect was also found in amplitude (within 226–266 ms) in the central area, but not in peak latency (see Figure 3.5). A t-max permutation test also showed that the N1 reduction effect of lexical tones tended to be right-lateralised (see Figure 3.6).

3.3.2.2 Consonant (Experiment 6)

The ERPs elicited by AO, AV and VO stimuli in the consonant experiment are very similar to those recorded in the lexical tone task. For example, with regard to the ERPs at Cz (see Figure 3.7), before the auditory onset time (0 ms), the waveforms of AV and VO nearly overlapped. After auditory onset, the AV consonant evoked an auditory N1 peaking at about 154 ms, while the AO consonant evoked an N1 maximising at about 159 ms. For the VO consonant, the ERP had a smaller negative deflection that peaked at about 184 ms. Because of the similarities between and comparability of lexical tone and consonant ERPs, the electrodes and time intervals selected for N1 and P2 in the consonant analysis remained consistent with those in the lexical tone analysis.

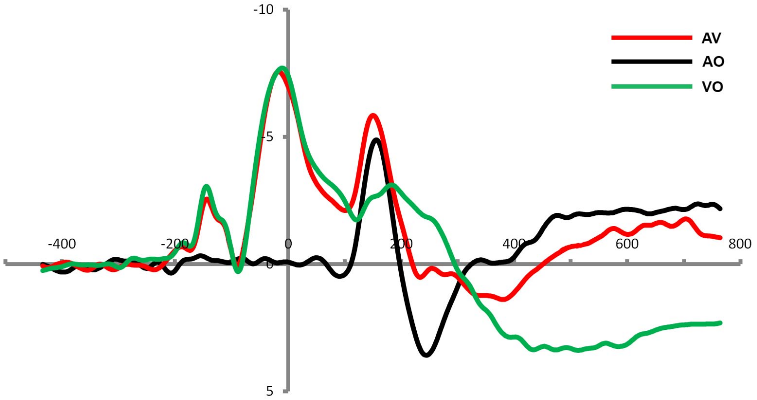


Figure 3.7 Grand average (N = 18) of waveforms recorded at Cz in the consonant experiment in the AO, AV and VO modalities.

Similarly, for the N1 mean amplitude in the consonant experiment, the same ANOVA with the factors Modality (AO, AO'), Lateralisation (left, right) and Electrode (F3/4, FC1/2, FC5/6, C3/4, T7/8, CP1/2) was conducted. The results showed that the main effect of all three factors was significant: Modality $F(1, 17) = 24.13$, $p < .001$; Lateralisation was marginally significant), $F(1, 17) = 3.93$, $p = .060$; Electrode, $F(2.93, 49.76) = 4.04$, $p = .013$. The interaction of Modality with Lateralisation was significant: $F(1, 17) = 5.68$, $p = .029$. The interaction of Lateralisation with Electrode was also significant: $F(3.17, 53.86) = 3.42$, $p = .022$. The other interaction failed to reach significance. The comparison of Modality showed that the N1 amplitude of consonant AO (-2.92 ± 0.40 μV) was larger than N1 of

AO’ (-1.10 ± 0.33 μV). A comparison of Modalities based on Lateralisation showed that AO was significantly larger than AO’ in both the left (p < .001) and right (p = .001) hemispheres. However, a comparison of Lateralization based on Modality revealed that asymmetric lateralisation was significant in the AO response, in which N1 activity in the left hemisphere (-3.12 ± 0.38 μV) was larger than that in the right one (-2.72 ± 0.42 μV) (p = .022). In the AO’ condition, there was no significant difference between the two hemispheres (p = .84) (see Table 3.3. and Figure 3.8).

N1 peak latency data were put into the same ANOVA and the results revealed that the main effects of Modality [F(1, 17) = 14.02, p = .002] and Electrode [F(2.39, 40.65) = 5.71, p = .004] were significant, but the main effect of Lateralisation was not significant, F(1, 17) = .203, p = .66. A comparison of Modality showed that AO’ latency (150.6 ± 1.6 ms) was 7.9 ms (± 2.1 ms) earlier compared to AO latency (158.6 ± 1.7 ms) (see Table 3.3 and Figure 3.8).

In the analysis of P2 mean amplitude, a two-way repeated measures ANOVA with the factors Modality (AO, AO’) and Electrode (FC1, FC2, Cz, CP1, CP2) showed that the main effect of Modality was not significant: F(1, 17) = 2.74, p = .116. Electrode was significant: F(3.20, 54.43) = 12.35, p < .001. The interaction of the two factors was not significant: F(2.20, 37.47) = 1.03, p = .24.

The results for P2 latency showed that a marginally significant effect was found for Modality, F(1, 17) = 3.71, p = .071. The rest of the tests failed to reach significance.

Table 3.3 Average ERP amplitude and latency (N = 18) in AO and AO’ (AV – VO) modalities in the consonant experiment.

Task	Modality	N1				P2			
		Amplitude (μV)	SE	Latency (ms)	SE	Amplitude (μV)	SE	Latency (ms)	SE
Consonant	**AO**	-2.92	0.40	158.6	1.7	2.6	0.30	246.3	3.1
	AO'	-1.10	0.33	150.6	1.6	1.76	0.40	237.3	5.2
	AO-AO'	-1.82	0.37	8.0	2.1	0.84	0.49	9.0	4.7

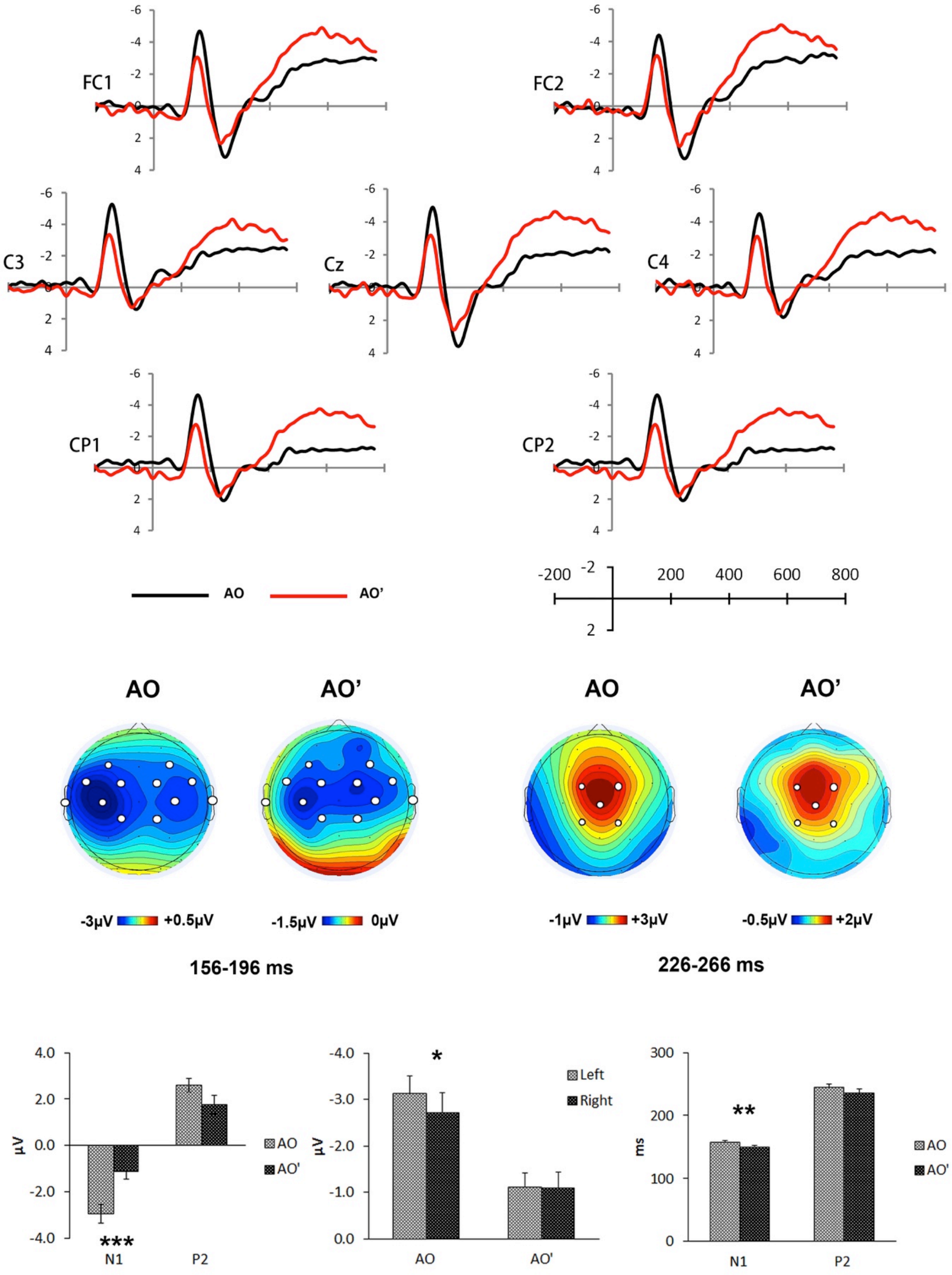


Figure 3.8 Average ERPs of seven electrodes (FC1/2, C3/4, Cz, CP1/2) of AO and AO' in the consonant experiment. Topographies represent the voltage distribution in the time windows 156–196 ms (N1) and 226–266 ms (P2) in AO and AO'. The electrodes involved in the analysis were marked as white dots. The bar charts represent comparisons between AO and AO' in mean amplitude (left), lateralisation (middle) and peak latency (right). *: *p* < .05; **: *p* < .01; *** *p* < .001

A t-max permutation test of the difference between AO and AO’ within the same time windows of 159–196 ms (N1) and 226–266 ms (P2) was performed. In the N1 interval, a large number of electrodes distributed evenly over the scalp reflected significant negative activity, including F3, FC5, FC1, T7, C3 CP1 and P7 in the left hemisphere, Fz, Cz and Pz in the middle vertex, and F8, F4, FC6, C4, T8, CP2, TP8, P4, P8 and PO8 in the right hemisphere (see Figure 3.9). However, in regard to the P2 interval, the comparisons were not significant for any electrodes.

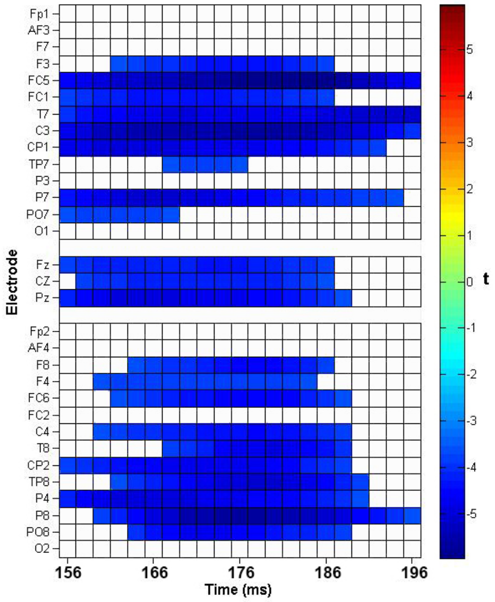


Figure 3.9 Consonant results of t-max permutation test for the difference between AO and AO’ within the interval 156–196 ms in the consonant task. Time-points that have a significant effect are colour-coded to represent a negative direction or a positive direction ($p < .05$).

An analysis of the ERP results for consonants showed that an N1 reduction effect was found in the amplitude (within 159–196 ms) and peak latency in the fronto-central electrodes, but no P2 reduction was found (see Figure 3.8). The results of t-max permutation test showed that the N1 reduction effect of consonants tended to be bi-lateralised (see Figure 3.9).

3.3.2.3 Comparison of lexical tones and consonants

According to the analyses above, a reduction effect in N1 was found in both the lexical tone and consonant experiments. A reduction in P2 (amplitude) was shown in the lexical tone experiment, but it was absent in the consonant experiment. The reduction in N1 was not identical in lexical tone and the consonant experiments. Based on the results of the t-max permutation test, the N1 reduction or audiovisual integration activity (AO – AO') distribution of the lexical tones and consonants appeared to differ in hemispheric patterns. In the lexical tone results, a clear right lateralisation was observed, while in the consonant results the hemisphere pattern tended to be more bi-lateralised. This gives rise to different activities of audiovisual interaction in the left and right hemispheres between the two experiments.

To further compare the audiovisual integration activity (AO – AO') of lexical tones and consonants, an additional ANOVA with a between-subject factor Experiment (lexical tone, consonant) and the within-subject factors Lateralisation (left, right) and Electrode (AF3/4, F3/4, FC5/6, FC1/2, C3/C4, T7/T8, CP1/CP2) was subjected to the mean amplitude and peak latency of the difference wave of AO and AO' within the time window 156–196 ms.[5]

The results for amplitude showed that the interaction of Experiment × Lateralisation was significant: $F(1, 33) = 7.29, p = .011$. A comparison of Lateralisation based on Experiment found that a significant lateral effect occurred in the lexical tone experiment, where the activity in the right hemisphere (-1.59 ± 0.37 μV) was stronger than that in the left hemisphere (-1.13 ± 0.37 μV) (p = .034), but the lateral effect was not significant in the consonant experiment (p = .119). In a comparison of Experiment based on Lateralisation, there was no significant difference between the lexical tones and consonants in either hemisphere.

[5] Only the ERPs of AO – AO' within the N1 time range were analysed for a comparison between lexical tones and consonants, because no reduction in P2 was found in the consonant experiment.

The ANOVA for peak latency showed that the between-subject factor Experiment was significant: $F(1, 33) = 14.\ 17$, $p = .001$. The consonant latency (174. 3 ± 2.5 ms) was 13.3 ms earlier than the lexical tone response (187.6 ± 2.5 ms) (see Figure 3.10).

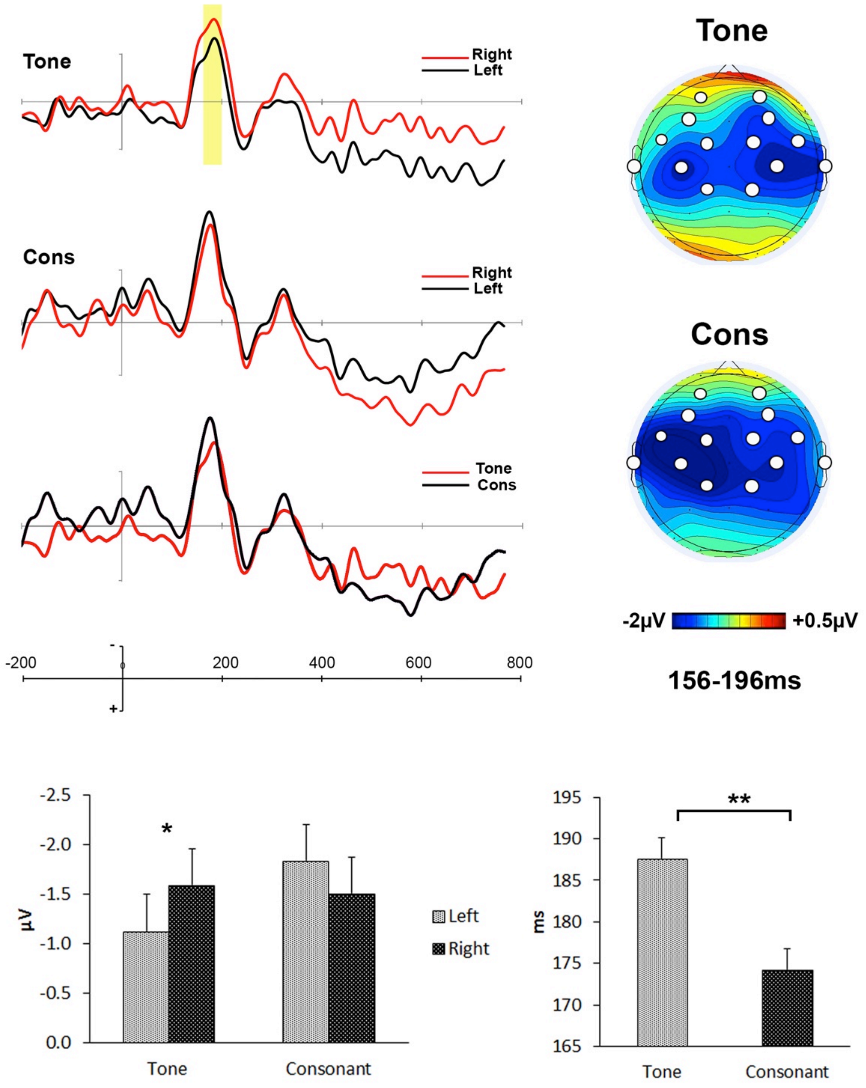


Figure 3.10 ERP waveforms represent the AO – AO' waves of selected channels (AF3/4, F3/4, FC5/6, FC1/2, C3/C4, T7/T8, CP1/CP2) from the lexical tone and consonant experiments, and they are marked as white dots on the topographies. The first two waveforms represent a hemispheric comparison for each experiment. The bottom waveforms represent a comparison between lexical tones and consonants. The statistical significance of the amplitude in the time window is highlighted in yellow. The bar charts are comparisons of hemispheres in amplitude (left) and peak latency (right) of the two experiments *: $p < .05$; **: $p < .01$

In the comparison of lexical tones and consonants in the audiovisual interaction ERPs, the amplitude was not statistically different, but lateralisation of lexical tones and consonant responses had contradictory patterns. Lexical tones were right-lateralised, whereas consonants tended to be left-hemisphere dominant, albeit not statistically significant. The latency between the two experiments was different as well. Consonantal responses were about 13 ms earlier than lexical tone responses.

3.4 Summary of the results and discussion

The results of the current experiments showed a few major electrophysiological features of audiovisual lexical tone integration. First, the auditory processing of lexical tones was both reduced and accelerated by visual inputs in the auditory N1 time range, which further supports an audiovisual benefit effect on lexical tone perception. Second, in the time course, audiovisual lexical tone integration maximised later and lasted longer (over N1 to P2) than audiovisual consonant integration. Third, lexical tone interaction processing showed right-lateralised activation.

N1 and P2 reductions in the amplitude and latency of lexical tones and consonants

Based on the analyses mentioned above, a reduction effect in amplitude and latency was found in both lexical tone and consonant processing during the early auditory component time range. Under the influence of visual speech, the auditory process was clearly smaller and faster in the N1 interval. In the AV modality, the response of both lexical tones and consonants was reduced and speeded up in the auditory N1 component. The results suggest that visual information can significantly affect the process of auditory lexical tones like the visual information effect on auditory consonants. Auditory N1/P2 is known to reflect activities involving sensory processing that are sensitive to physical variations in stimuli (Näätänen & Winkler, 1999). Therefore, the reduction in the lexical tone task within this time range (particularly in the N1 time interval) supports the audiovisual interaction of lexical tones that began in sensory processing (pre-linguistic processing), similar to the audiovisual integration of consonants in this time range.

Compared to consonants, the lack of salient visual features (place of articulation) in lexical tones did not restrain the reduction effect during audiovisual speech processing. As can be seen in the behavioural data for lexical tones in VO, lip-reading lexical tones was much more difficult than lip-reading consonants, as mouth movement-related visual cues provide limited information for distinguishing lexical tones. Nevertheless, lexical tones had a similar reduction effect to consonants. This suggests that the reduction effect found in lexical tones does not only depend on how much phonetic information visual input can convey. Instead, it is mainly due to non-phonetic information from visual input.

However, the reduction effects between lexical tones and consonants are not completely identical. First, based on t-max analyses of the N1 and P2 time windows, the difference between AO and AO' (reduction effect) of lexical tones was weaker, but lasted longer, compared to the integration activity in consonants. For both lexical tones and consonants, the reduction effect was present in N1 but almost absent in P2. In the lexical tone analysis, auditory N1 amplitude was reduced and latency was accelerated in the AV condition, but fewer electrodes showed a reduction in P2 amplitude and the reduction latency in P2 was not significant. For consonants, the reduction effect was significant over a large number of electrodes in N1, but not significant in P2. The audiovisual integration process was maximised within the auditory N1 time range (156–196 ms) and attenuated within the P2 time range (226–266 ms). This suggests that in the N1-P2 stage of audiovisual integration processing, lexical tones seemed to have weaker integration processing, but a longer latency compared to the process for consonants.

The comparison between lexical tones and consonants in N1 reduction confirmed that audiovisual integration activity was different in latency. The latency of consonants was about 13 ms earlier than that of lexical tones, which suggests that the audiovisual integration process of consonants was faster than that of lexical tones. In terms of amplitude in N1, no significant difference between the two experiments was found, but t-max permutation test and the following ANOVA showed that consonant response tended to be slightly stronger than that of lexical tones.

From previous studies in the literature, there are two explanations for the N1 reduction effect in audiovisual speech processing. First, N1 reduction (in latency) is determined by the phonetic saliency or predictiveness of visual information for the phonetic content of auditory speech (van Wassenhove et al., 2005). Second, N1 reduction (in amplitude/ latency) reflects non-speech-specific visual information predicting the timing of upcoming auditory signals (Stekelenburg & Vroomen, 2007). In the current experiments, because the lexical tone and consonant stimuli are physically identical, any difference in the reduction effect between them should be due to their different processing of audiovisual speech integration. The phonetic saliency of visual information could affect audiovisual integration in this stage. There is no doubt that visual consonants have much stronger predictability for ensuing auditory consonants compared to lexical tones. One can anticipate a consonant from a distinctive place of articulation (e.g. bilabial) from the preceding mouth movement hundreds of milliseconds prior to the auditory signal. However, leading lip movements are less likely to predict lexical tones. Because the visual cues of lexical tones may rely on other cues such as duration (Smith & Burnham, 2012), head movement (Burnham et al., 2006) and laryngeal movement (Chen & Massaro, 2008), but not the place of articulation, the judgement of lexical tones may not be completed until the middle of syllable duration. Therefore, onset mouth movement is much less predictable for lexical tones. The difference in the reduction effect between the two speech units suggests that the reduction effect or audiovisual integration processing may be involved the processing of visual features in terms of phonetic saliency for consonants in this early time stage (mainly in N1).

The N1 reduction effect possibly reflects two types of audiovisual speech integration processing. Non-speech/ phonetic visual information (e.g. timing information) and phonetic visual information (e.g. place of articulation) interact with auditory speech processing. The difference between lexical tone and consonant N1 reduction is stronger in latency than in amplitude. This suggests that reducing the amplitude might reflect a non-phonetic type of audiovisual integration, while shortening the latency is more sensitive to a phonetic-specific type of audiovisual integration.

Lateralisation of the N1 reduction effect

For both experiments, the topographies of the difference wave (AO – AO') strongly resembled common auditory N1 scalp distribution maximised at the fronto-central electrodes, and their lateralisation was asymmetric across the hemispheres in the lexical tone and consonant responses. As illustrated in Figure 3.10, the two speech units had opposite patterns in lateralisation in N1, in which lexical tone response was right-hemisphere dominant, but consonant response was bi-lateralised but with a tendency towards being left-lateralised. The different lateralisation in the N1reduction effect further indicates that the audiovisual integration processing of lexical tones differs from that of consonants.

The hemispheric dominance of audiovisual integration seems to vary across studies; hence it is difficult to interpret lateralisation of the reduction effect in two experiments. One possible explanation is that how lateralisation of the N1 reduction effect differs across lexical tones and consonants might be determined by whether the processing is linguistic-specific or not. In auditory speech studies, the lateralisation of lexical tones depends on how lexical tones are perceived. A characteristic of lexical tones is to form phonemic contrasts through pitch variation. Therefore, lexical tones can be perceived as speech units as well as pitch variations. It is well known that the left hemisphere dominates the speech process (Broca, 1861; Wernicke, 1874), such as phonemes and words (Kimura, 1973; Shankweiler & Studdert-Kennedy, 1967; Studdert-Kennedy & Shankweiler, 1970), while melody and prosodic signals, such as musical pitch, are predominantly processed in the right hemisphere (Bryden, 1982; Curry, 1967; Kimura, 1973). There is evidence that the processing of lexical tones is more left-hemisphere dominant if they are perceived as linguistic information (lexical tone categories); otherwise, processing is more right-lateralised if they are treated as non-linguistic specific information (pitch variations) (Jongman et al., 2006). Left-lateralisation was observed in the responses of Mandarin native speakers in a dichotic identification task, but there was no such lateralisation found in English native speakers (Wang et al., 2001). However, Lou et al. (2006) compared lexical tones and consonant processing at the pre-attentive level in an MMN approach and found that lexical tone response was more lateralised to the right hemisphere, while consonant response dominated in the left hemisphere. A later study (Shuai & Gong, 2014) found that the bottom-up processing (acoustic feature processing) of lexical tones was right-hemispheric dominance, while the top-down processing (semantic processing) of lexical tones was left-hemispheric dominance. This characteristic of language-biased lateralisation might also be reflected in the processing of audiovisual integration. The right lateralisation of lexical tone reduction effect suggests that the audiovisual integration of lexical tones in the N1 time stage is non-phonetic processing. In contrast, bi-lateralisation (with a tendency towards left lateralisation) of the N1 reduction effect of consonants possibly reflects the phonetic processing of consonants.

In sum, the ERP experiments provide direct evidence that visual information influences auditory lexical tone processing in the auditory cortex by reducing and accelerating early responses within the N1 time range. The results indicate that audiovisual lexical tone integration is non-phonetic sensory processing in the early stage (about 188 ms) of speech processing.

Chapter 4 Incongruent audiovisual lexical tones perception

The previous experiments investigated the audiovisual benefit effect where visual information facilitates the perception of auditory lexical tones in an audiovisual congruent situation. In this chapter, the experiments aim to examine the effect of incongruent visual information on the perception of auditory lexical tones by comparing Mandarin lexical tones perception between congruent and incongruent conditions. The chapter includes a behavioural experiment on lexical tone identification (Experiment 7) in Section 4.1 and an ERP experiment (Experiment 8) in Section 4.2.

4.1 Incongruent visual information modifies auditory lexical tone perception (Experiment 7)

4.1.1 Introduction

Different from the previous experiments, Experiment 7 focuses on exploring whether incongruent visual information has an influence on auditory tone perception in an incongruent audiovisual lexical tone syllable like the classic McGurk effect in certain consonants. In the experiment, incongruent and congruent audiovisual lexical tones that only differ in their visual input were compared in a two-alternative-choice identification task. Specifically, syllable /a/ with two lexical tones T3 and T4 was selected as the experimental material. An auditory T3 syllable was paired a the visual T4 syllable to combine and form an incongruent lexical tone syllable $A_{T3}V_{T4}$ (e.g. A/ǎ/V/à/), and auditory T4 was paired with visual T3 to create an incongruent lexical tone syllable $A_{T4}V_{T3}$ (e.g. A/à/V/ǎ/). The reason for choosing T3 and T4 is due to their distinctive visual feature of duration, which exhibits a stronger influence on auditory tone perception, as shown in the results of Experiments 2–4. Using visually salient lexical tones can maximise the strength of the visual effect on auditory tone perception in incongruent syllables. Because the articulatory movements of lexical tones are not radically different from each other (i.e. looks similar), incongruent visual information cannot change the perception of auditory lexical tones in the place of articulation like consonants, but it might change the duration perception of an auditory tone. As mentioned previously, tone duration is a secondary cue when F0 information as a primary cue is available. A change in the perception of tone duration might consequently bias the identification of lexical tones. To enhance the effect of incongruent visual information, noise was added to weaken the auditory signal.

In addition, the identification of incongruent audiovisual vowels was added in a separate task. The syllables /ɑ/ and /i/ were used for incongruent vowel combinations (A/i/V/a/, A/a/V/i/), and all syllables were consistently in T3. The purpose of comparing lexical tones and vowels in incongruent perception was to compare with the McGurk effect due to the saliency difference in visual information. Unlike lexical tones, vowels can be easily recognised in vision through mouth movement (e.g. lip roundedness). In an incongruent vowel, the conflict between auditory and visual input should consequently be easily attended to (Summerfield & McGrath, 1984). One would expect an incongruent visual vowel to have a strong influence on auditory vowel identification. To enhance the incongruent visual speech effect on auditory tones in an incongruent condition, noise was employed to decrease the strength of the auditory signal. The participants were required to respond only to auditory speech as fast as possible (ignoring visual speech) in all conditions.

The prediction for the results of the lexical tone task is that, in clear condition, the accuracy of incongruent and congruent lexical tones may be the same. This is because, first, tone judgement heavily relies on intact F0 cues from auditory tones, which remain identical across two conditions; and second, the discrepancies between auditory and visual inputs in incongruent audiovisual lexical tones should be difficult to notice because lexical tones have very similar articulatory movement, therefore the mismatch between auditory and visual inputs should not affect the identification rate. In noise condition, the identification of the incongruent lexical tones is expected to be lower than that of the congruent ones. This is because when F0 is less reliable, perceivers tend to judge lexical tones by using duration cues, which could be perceptually changed by incongruent visual information (e.g. in $A_{T4}V_{T3}$ syllable, V_{T3} could lengthen A_{T4} duration in perception), consequently, tone identification may be biased (e.g. $A_{T4}V_{T3}$ could be illusorily perceived as T3). In terms of RT results, incongruent and congruent lexical tones should be different in both clear and noise conditions. The RT of $A_{T3}V_{T4}$ is expected to be faster compared to the RT of congruent T3, because visual T4 should shorten the duration of auditory T3 in perception. In contrast, the RT of $A_{T4}V_{T3}$ should be longer than the RT of congruent T4, as visual T3 could lengthen the duration of the auditory T4 in perception. The RT difference between congruent and incongruent tones should be greater in noise condition than in clear condition.

As for the prediction of results for the vowel task, incongruent vowels should be greatly different from congruent vowels in both clear and noise conditions. Incongruent vowels should be significantly worse than congruent vowels in their identification rates. Additionally, incongruent vowels are expected to be much slower than congruent vowels in RT results. This is due to the large discrepancy between auditory and visual vowels in the incongruent condition, which is very likely to reduce correct identification.

4.1.2 Method

4.1.2.1 Participants

Twenty-one native Mandarin speakers (aged: 26.7 ± 6.1 years; 13 females) were recruited from Bournemouth University. Most of the participants did not participate in the previous experiments. None had reported any diagnosed hearing impairments, they had normal or correct to normal vision and all were right-handed. The participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board.

4.1.2.2 Materials

The audiovisual syllables used in the lexical tone task were the congruent syllables /ǎ/ (T3) and /à/ (T4) (both auditory and visual syllable /ɑ/ in T3 or T4) and the incongruent syllables A/ǎ/V/à/ ($A_{T3}V_{T4}$) and A/à/V/ǎ/ ($A_{T4}V_{T3}$) (auditory syllable /ɑ/ in T3 was paired with visual /ɑ/ in T4; auditory /ɑ/ in T4 was paired with visual /ɑ/ in T3). In the vowel task, the audiovisual syllables also contained the congruent syllables /ɑ// (the same as in the lexical tone task), /i/ and the incongruent syllables A/a/V/i/ (auditory /ɑ/ was paired with visual /i/)

and A/i/V/a/ (auditory /i/ paired with visual /ɑ/). All vowel syllables were consistently in T3 (see Table 4.1).

Table 4.1 Audiovisual stimuli duration and corresponding auditory and visual component durations and onset times (in ms).

Task	Congruence	Syllable	Token	Stimulus duration	Lip onset	Audio onset	Lip movement duration	Audio duration
Tone	Congruence	/ă/	T3	1869	267	767	1535	830
	Congruence	/à/	T4	1869	250	767	1268	480
	Incongruence	A/ă/V/à/	$A_{T3}V_{T4}$	1869	250	767	1268	830
	Incongruence	A/à/V/ă/	$A_{T4}V_{T3}$	1869	267	767	1535	480
Vowel	Congruence	/ă/	a	1869	267	767	1535	830
	Congruence	/ĭ/	i	1869	684	767	1034	903
	Incongruence	A/ă/V/ĭ/	AaVi	1869	684	767	1034	903
	Incongruence	A/ĭ/V/ă/	AiVa	1869	267	767	1535	830

The video materials were recordings of a native male Mandarin speaker (24 years old). The video recordings were edited in Adobe Premiere Pro CC (Adobe Systems, California) as pixels of 1280 × 720 clips with a digitisation rate of 29.97 frames per second (1 frame = 33.37 ms). In the video, only the lower half of the speaker's face was presented (no eyes shown). The soundtracks of the videos were edited in Adobe Audition CC (Adobe Systems, California). All auditory tracks were digitised at 48 kHz with 32-bit amplitude resolution and were RMS normalised in amplitude to -9 dB. The noise condition was babble noise which contained mixed voices from six Mandarin native speakers (3 females) reading sentences in Mandarin. The duration of audiovisual stimuli was set at 1,868.54 ms (56 frames), in which the front silence-gap between stimulus onset and auditory onset remained at 767.43 ms (23 frames) (see Figure 4.1). In natural speech, mouth movement always occurs before sound is produced (approximately 100 ms ahead). For the syllables in the current experiment, lip-movement onset time was about 298 ms on average earlier than auditory onset. The syllable durations varied; thus, the silence gap from audio offset to video offset also varied (see Table 4.1).

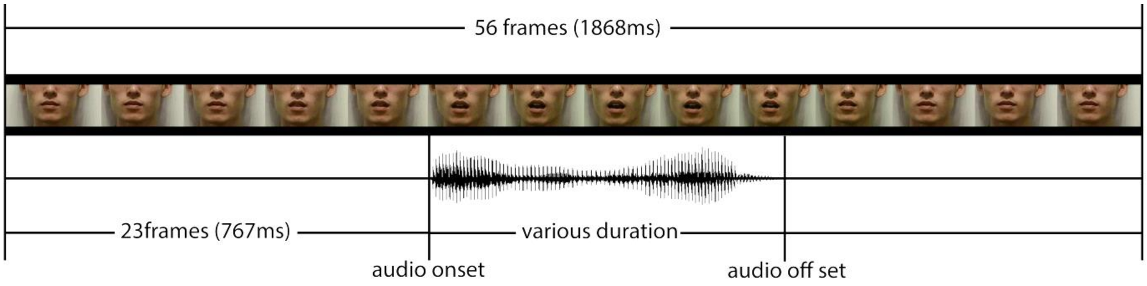


Figure 4.1 Trial structure in Experiments 7–8. The duration of the audiovisual stimuli was 1,868.54 ms (56 frames). Two silent gaps occurred before and after the audiovisual syllable. The period from the beginning of the clip to auditory onset was fixed at 767.43 ms (23 frames), and auditory syllable duration and the period from auditory offset to the end of the video clip varied.

4.1.2.3 Procedure

The experiment took place in a soundproof booth. All participants took part in two two-alternative-choice identification tasks: a lexical tone identification task and a vowel identification task. In the lexical tone task, participants were told to watch video clips presented on a monitor, and they were instructed to only respond to auditory lexical tones (T3 or T4) by pressing a key on a keyboard. In the vowel task, the participants were asked to react to auditory vowels (/a/ or /i/). For each trial, a fixation cross was displayed at the centre of the screen for 1,000 ms, and a video clip followed it. From video offset, an extra 3,000 ms was allowed to respond, and the ITI remained at 1,000 ms throughout the trials. For each task, each stimulus contained two tokens, and each syllable was repeated 18 times in clear and noise conditions. The total trial number of each task was as follows: 2 congruent syllables × 2 incongruent syllables × 2 tokens × 2 listening conditions × 18 repetitions = 288 trials. All conditions were randomised. Each task consisted of six blocks and lasted for about 25 minutes. The lexical tone task and the vowel task were counterbalanced for each participant. Before the experimental sessions, the participants were given a practice session to familiarise themselves with the tasks, and they were encouraged to make their judgements as quickly as possible. As for the data-collecting apparatus, the participants sat in a soundproof booth and faced a 19-inch monitor at a distance of 60 cm. Sounds were played through noise-cancelling headphones: Sennheiser HD 280 (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany). The loudness of the presented sound was approximately 65 dB SPL. The experimental stimuli were presented through E-prime 2.0 (Psychology Software Tools, Sharpsburg, USA) on a controlled PC.

4.1.3 Results

Results for lexical tones

For analysis of the lexical tone results, a three-way repeated measures ANOVA with the factors Congruence (congruence, incongruence), Tone (T3, T4) (T3 and T4 refer to audiovisual syllables whose auditory inputs were T3 and T4, respectively, regardless of their congruence) and Listening condition (clear, noise) was conducted for accuracy.

Table 4.2 Mean accuracy (%) and RT (ms) (N = 21) of the lexical tone task in all conditions.

	Condition	Congruent T3		Incongruent $A_{T3}V_{T4}$		Congruent T4		Incongruent $A_{T4}V_{T3}$	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
Acc (%)	Clear	97.5	0.88	97.2	0.68	98.1	0.81	97.5	0.75
	Noise	95.4	1.16	94.5	1.63	88.4	2.13	90.0	1.27
RT (ms)	Clear	1315	16	1308	15	1239	11	1242	13
	Noise	1400	20	1387	20	1377	25	1387	28

The analysis showed that the main effect of Listening condition was significant: $F(1, 20) = 36.23$, $p < .001$; Tone was marginally significant: $F(1, 20) = 3.47$, $p = .077$; Congruence was not significant: $F(1, 20) = .004$, $p = .95$. The two-way interaction of Listening condition with Tone was significant [$F(1, 20) = 9.65$, $p = .006$], but the other interactions failed to reach significance. The tone accuracy in the clear condition exhibited different patterns from that in noise. T3 (95.0 ± 1.3%) was significantly higher than T4 (89.2 ± 1.6%) in the noise condition ($p = .024$) (see Figure 4.2). The incongruence effect of lexical tones was not found in any conditions.

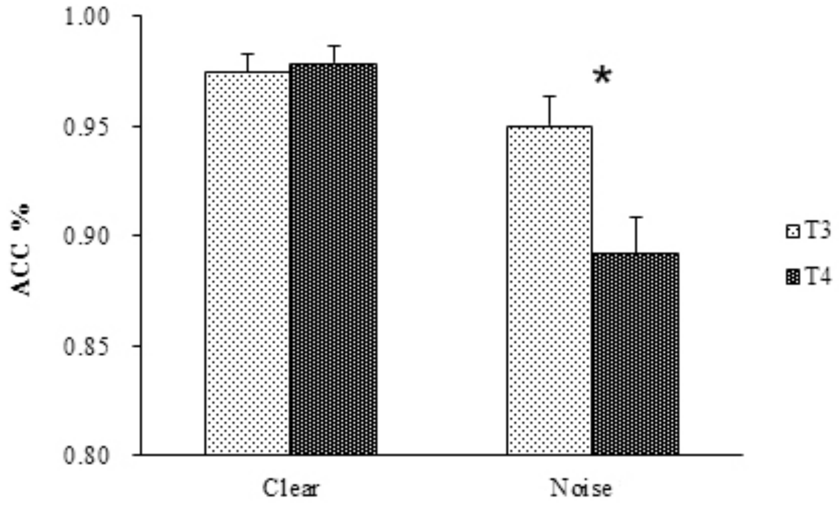


Figure 4.2 Accuracy of auditory T3 and T4 in clear and noise conditions. T3 and T4 refer to the audiovisual syllables whose auditory components were T3 and T4, regardless of their congruence. *: $p < .05$.

The same ANOVA was performed for the RT data, and the results revealed that the main effect of Listening condition was significant: $F(1, 20) = 54.80$, $p < .001$. The main effect of Tone was also significant: $F(1, 20) = 13.01$, $p = .002$. The two-way interaction of Listening condition with Tone was significant: $F(1, 20) = 13.72$, $p = .001$, and the interaction of Congruence with Tone was also significant: $F(1, 20) = 6.01$, $p = .024$. The interaction of the three factors was not significant: $F(1, 20) = 1.11$, $p = .31$. Post hoc comparisons of Congruence based on Tone showed that the RT of the congruent syllable T3 (1358 ± 17 ms) was significantly slower than the RT of the incongruent syllable $A_{T3}V_{T4}$ (1347 ± 17 ms) ($p = .001$), but the comparison of congruent T4 (1308 ± 16 ms) and incongruent $A_{T4}V_{T3}$ (1315 ± 19 ms) was not significantly different ($p = .23$) (see Figure 4.3).

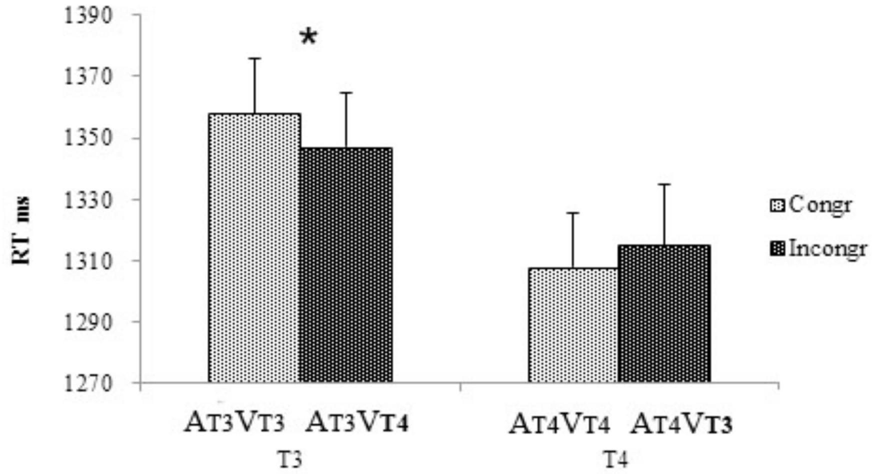


Figure 4.3 RT comparison between congruent and incongruent audiovisual syllables in the lexical tone task. T3 and T4 refer to the auditory tones of the audiovisual syllables. *: $p < .05$.

The analysis revealed that the incongruent syllable $A_{T3}V_{T4}$ was faster than the congruent syllable T3 in response time, and there was a trend that $A_{T4}V_{T3}$ was slower than T4 in response time, although the incongruent effect was not significant. However, $A_{T3}V_{T4}$ was not illusorily perceived as T4 and $A_{T4}V_{T3}$ was not misperceived as T3 either in the clear or noise conditions.

Results for vowels

For the vowel task, a similar three-way repeated measures ANOVA with the factors of Congruence (congruence, incongruence), Vowel (/ɑ/, /i/) and Listening condition (clear, noise) was performed for accuracy.

Table 4.3 Mean accuracy (%) and RT (ms) (*N* = 20[6]) for the vowel task in all conditions.

	Condition	Congruent AaVa		Incongruent AaV**i**		Congruent AiVi		Incongruent AiV**a**	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
Acc (%)	Clear	95.3	2.39	94.7	2.30	96.5	2.39	93.7	2.34
	Noise	96.0	2.37	72.1	6.25	96.0	2.36	74.3	5.68
RT (ms)	Clear	1241	20	1355	23	1321	18	1308	15
	Noise	1270	19	1424	21	1352	15	1426	21

The results showed that the main effect of Congruence was significant: *F*(1, 19) = 16.75, *p* = .001. The main effect of Listening condition was also significant: *F*(1, 19) = 16.33, *p* = .001. Moreover, a significant two-way interaction of Congruence with Listening condition was found: *F*(1, 19) = 14.71, *p* = .001. A post hoc test revealed that the accuracy of the incongruent syllable (75.0 ± 4.9%) was significantly lower than that of congruent syllables (95.8 ± 2.5%) in the noise condition (*p* = .001). In the clear condition, the difference between the accuracy of incongruence (95.9 ± 2.5%) and congruence (94.3 ± 2.4%) only reached marginal significance (*p* = .094) (see Figure 4.4). This suggests that the discrepancy between auditory and visual vowels greatly affected perception in the noise condition.

[6] One participant's results in the vowel task were removed from the analysis since the participant had missing RT data in an incongruent condition.

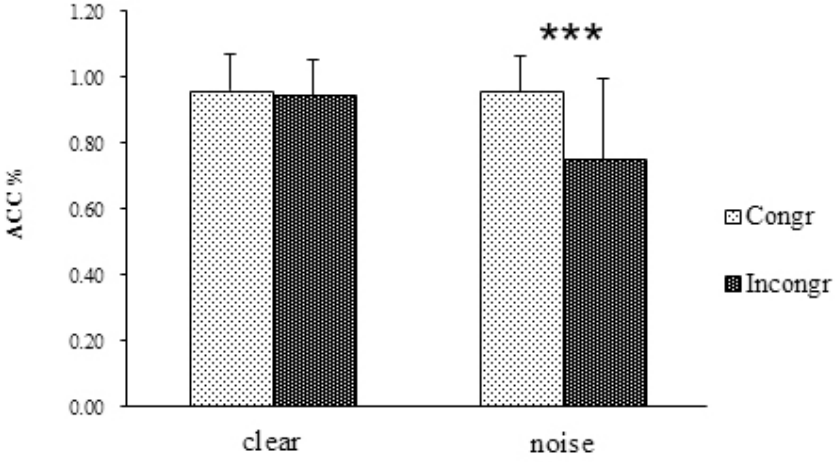


Figure 4.4 Accuracy of congruent audiovisual vowels (/a/ and /i/) and incongruent audiovisual vowel stimuli (AaVi and AiVa). *: $p < .05$

A similar three-way ANOVA for RT data showed that three main effects were significant [Congruence: $F(1, 19) = 89.72$, $p < .001$; Vowel: $F(1, 19) = 12.81$, $p = .002$; Listening Condition: $F(1, 19) = 19.82$, $p < .001$)]. A significant effect of the interaction of Congruence with Listening Condition was found: $F(1, 19) = 18.95$, $p < .001$. A significant effect of the interaction of Congruent with Vowel was also found: $F(1, 19) = 65.93$, $p < .001$. The interaction of Listening condition with Vowel was barely significant: $F(1, 19) = 4.33$, $p = .051$. The three-way interaction effect was significant: $F(1, 19) = 6.66$, $p = .018$. Post hoc comparisons of Congruence difference of vowels revealed that the RT of incongruent vowels was significantly slower than the RT of congruent vowels in all conditions ($p < .001$), except for the comparison between /i/ and AiVa in the clear condition ($p = .106$) (see Figure 4.5).

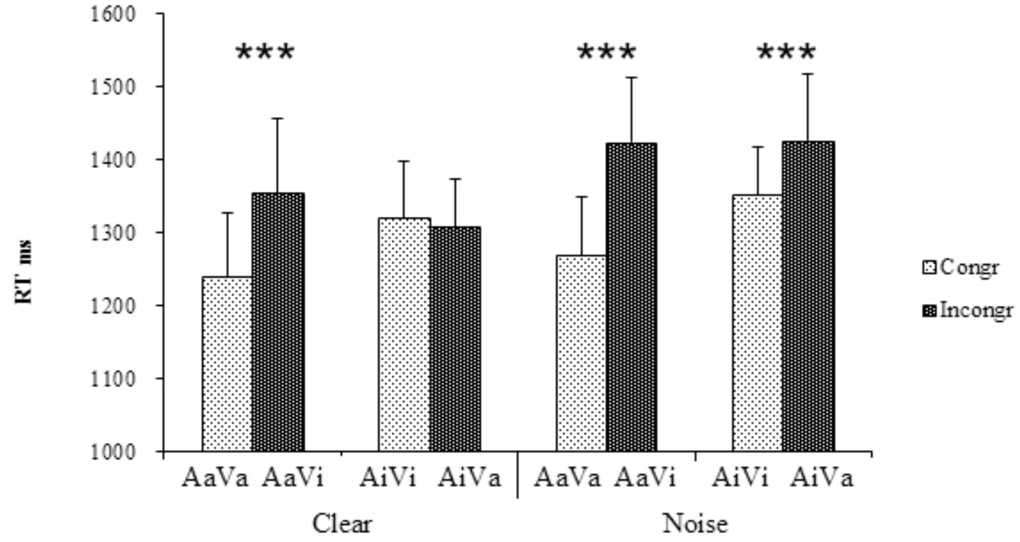


Figure 4.5 RT of congruent audiovisual vowels (/a/, /i/) and incongruent audiovisual vowel stimuli (AaVi and AiVa). *: *p* < .05

The RT analysis of the vowels produced two findings: 1. The response time of incongruent vowels was much slower compared to that of congruent vowels, which suggests that vowel perception was greatly hindered by audiovisual discrepancies. 2. The response time for incongruent visual vowels was also influenced by the timing of visual cues. This can be seen from the RT difference between congruent /ɑ/ and /i/. In comparisons of vowel types (/ɑ/ vs /i/) in the congruent condition, regardless of whether it was a clear or noise condition, the RT of /ɑ/ was shorter than that of /i/ (*p* < .001). This indicates that congruent /ɑ/ was persistently faster than congruent /i/. This could be due to the onset of their mouth movement, where the onset of visual /ɑ/ (267 ms) was earlier than that of visual /i/ (684 ms). The visual cues (lip roundedness) of vowels /ɑ/ and /i/ can be effectively identified from the very beginning of movement, and the earlier onset appears, the faster identification can be made. The timing of visual cue onset influences incongruent vowels. In a comparison between /i/ and AiVa, the RT of the incongruent vowel AiVa was not slower than the RT of congruent vowel /i/ (*p* = .106), suggesting that a mismatched visual /ɑ/ speeds up the response time of vowel perception, despite the large discrepancy between auditory /i/ ans visual /ɑ/, because the lip movement onset of /ɑ/ in AiVa appears earlier.

4.1.4 Summary of the results

Accuracy

Between the two tasks, lexical tone identification was clearly less influenced by mismatched visual information relative to vowel identification. Incongruent visual information hardly altered tone identification in either the clear or the noise condition, which indicates that visual information has a weaker effect on the auditory tone process compared to vowels. Even in the noise condition, where auditory cues (F0) were impaired, the effect of mismatched visual information for lexical tones was unable to reduce identification. Lexical tone perception was, therefore, more acoustically dominant, even in the noise condition. However, the possibility that a ceiling effect led to a negative result cannot be completely ruled out. As shown in Figure 4.2, lexical tone accuracy in the noise condition remained high (close to 90%), and the auditory lexical tone cue F0 might still be readily available for identification.

In contrast, the incongruence effect seriously inhibited vowel identification in the noise condition, which demonstrates that mismatched visual information has a strong effect on auditory vowel perception. The incongruent vowel effect could be mainly because a large discrepancy between auditory and visual inputs in an incongruent vowel was clearly apparent, therefore impeding correct identification in a speedy judgement. When the participants were instructed to react to stimuli as fast as possible, the visual input might have activated the representation corresponding to auditory vowels, which was inconsistent with actual dubbed auditory vowels. To achieve optimal identification, the activation by visual vowels needed to be suppressed, which can be explained by the slower RT. Although there was no solid evidence from the current data that incongruent vowels generated fused perception of a novel vowel, there is a possibility that the fused vowel /ə/, phonetically in the middle of /ɑ/ and /i/, caused incorrect identification, particularly in the noise situation.

Reaction time

In terms of the RT results in lexical tone identification, incongruent visual information shortened or lengthened the processing time of judgements of auditory tones. The response time for the incongruent syllable $A_{T3}V_{T4}$ was speeded up under the effect of incongruent visual T4 compared to the response time for congruent T3. The response time for incongruent $A_{T4}V_{T3}$ was extended due to the influence of incongruent visual T3 compared to the response time for congruent T4 (see Figure 4.3). The variation in RT in the incongruent condition was systematically consistent with the duration of visual input and not with the incongruence status (incongruent syllables tend to have longer RTs). The results suggest that incongruent visual information alters the perception of auditory lexical tones in terms of duration, as predicted.

The RT of a lexical tone relates to the duration perception of a lexical tone. The long tone T3 had a slower RT, and the short tone T4 had a faster RT, regardless of whether the condition was clear or noise (see Figure 4.3). The F0 movement of T3 is more complicated and needs longer time to articulate; consequently, a longer processing time would be required to capture the variation in F0. Processing T4 does not require a long RT as the F0 movement of T4 varies rapidly. For incongruent lexical tones, the perception of $A_{T3}V_{T4}$ had a shorter RT in contrast with congruent T3 RT, which indicates that $A_{T3}V_{T4}$ was perceived as a shorter T3 than congruent T3, while $A_{T4}V_{T3}$ was perceived as a longer T4 than congruent T4.

However, auditory and visual information did not seem to contribute equally to the duration integration of incongruent lexical tones. The auditory signal dominated in integration. No matter how much incongruent visual information shortened or extended the tone duration perceived, the response times for audiovisual syllables containing auditory T3 were persistently longer than the response times for those containing auditory T4. For example, in terms of response time, $A_{T3}V_{T4}$ was longer than $A_{T4}V_{T3}$, and $A_{T4}V_{T3}$ was shorter than $A_{T3}V_{T4}$ (see Figure 4.3). This suggests that lexical tone audiovisual perception in either incongruent or congruent condition still depends on the auditory signal, although visual information has some effect.

With regard to the results for the vowel task, as explained earlier, the response time for incongruent vowel perception was strongly affected by the conflict between auditory and visual information. Additionally, the response time for incongruent vowel perception was also affected by how early visual information was presented. The mouth movement for /ɑ/ was about 417 ms earlier than that for /i/ (see Table 4.1), and mouth movements were already effective as visual cues for identifying vowels. Consequently, the response time for /ɑ/ was faster than that of /i/ RT. In contrast, the RT of incongruent vowels was persistently slower than the RT of congruent vowels (except for AiVa) because of the apparent conflict between auditory and visual signals. However, the RT difference between congruence and incongruence was not always identical across the conditions of vowel types. As can be seen in Figure 4.5, the RT difference between AaVa and AaVi was consistently larger than the RT difference between AiVi and AiVa (in both clear and noise conditions). It appears that the mismatched visual vowel /ɑ/ could speed up the response time of the auditory vowel, regardless of the congruence status. It suggests that the timing when visual vowels (mouth movement) are presented influences the response time of auditory vowels, even though preceding visual vowels are inconsistent with ensuing auditory vowels (e.g. visual /ɑ/ preceding auditory /i/). This implies that the process for incongruent audiovisual vowels is not necessarily limited to the activation of phonetic-specific audiovisual integration; instead, it could reflect, in part, a non-phonetic audiovisual integration process.

In sum, the results of the experiment imply that incongruent visual information alters the perception of lexical tones in terms of duration. However, the results are unable to further demonstrate that incongruent visual information can decrease tone identification, which suggests that incongruent visual information modifies lexical tone perception on a non-phonetic (duration feature) level of processing.

4.2 Mismatched negativity of incongruent audiovisual lexical tones (Experiment 8)

Experiment 7 found that for the incongruent lexical tone $A_{T3}V_{T4}$, the incongruent visual information of T4 shortens the processing time of auditory T3. This provides evidence that this incongruent lexical tone could be perceived as the same lexical tone but with a shorter duration compared to the duration of congruent T3 without noticing the discrepancy between auditory and visual inputs. However, the effect of incongruent visual information was not successfully observed in the accuracy results, which could be because incongruent audiovisual lexical tone interaction is not processed on the phonetic (categorical) level but on the level of physical feature (duration) processing. To further test incongruent audiovisual lexical tone processing in the brain, the approach of eliciting MMN component, an index of pre-attentive detection of an infrequent stimulus from a frequently presenting stimulus in brain response (Näätänen et al., 2007), was employed in the present experiment.

4.2.1 Introduction

In the literature, the MMN evoked by McGurk stimuli can be obtained with a passive oddball paradigm, where incongruent syllables are presented infrequently (deviant) among congruent syllables that are frequently present (standard) in a stimulus sequence. The deviant stimuli elicit a larger negativity in-between the P2 and N2 time windows at the frontal or fronto-central electrodes (Colin et al., 2004; Colin et al., 2002b; Möttönen et al., 2002; Saint-Amour et al., 2007; Sams et al., 1991). In those studies, the auditory speech of an incongruent syllable is identical to that of a congruent syllable, yet the visual speech information differs (e.g. incongruent: A1V2; congruent: A1V1); the incongruent syllable is set as a McGurk syllable. In this situation, the auditory ERPs evoked by deviant and standard stimuli should be the same, as the auditory inputs in the two conditions are physically identical. The MMN (measured after auditory signal onset) evoked by deviant stimuli indicates the results of incongruent visual speech modifying auditory speech perception at a pre-attention level of processing.

The current experiment applied this method to observe the effect of incongruent visual information on the processing of auditory tones. The incongruent lexical tone A/ǎ/V/à/ ($A_{T3}V_{T4}$) was presented as the deviant stimulus, and the congruent lexical tone T3 was the standard stimulus. The target stimuli were audiovisual stimuli that were identical to either deviant or standard stimuli, but they were presented in black-and-white video. The participants were required to respond to the target stimuli only. The stimuli in different conditions were quasi-randomly presented in a sequence according to certain proportions (see Method in Section 4.2.2.3). The task was to detect the target stimuli (by pressing a key on a keyboard) which occasionally appeared in the sequence, which could detract the participants' attention from incongruent and congruent stimuli. MMN was expected to be found in the auditory N2 time range at the frontal area of the scalp. Additionally, like the vowel task in Experiment 7, the second part of this experiment measured the MMN evoked by the audiovisual vowels in order to compare lexical tones with speech that has more salient visual features. The incongruent vowel syllables A/a/V/i/ (auditory /ɑ/ paired with visual /i/) as deviant stimuli were presented within a sequence of repetitive congruent vowels /ɑ/.

From the RT results for the incongruent tone $A_{T3}V_{T4}$ in Experiment 7, incongruent visual information altered the perception of auditory tone duration. Therefore, such visual modification processing is expected to be observed in the current experiment. The prediction of the lexical tone response is that the MMN evoked by incongruent lexical tone $A_{T3}V_{T4}$ will be found after auditory signal onset. Specifically, the negativity evoked by the incongruent lexical tone will be larger than that evoked by the congruent lexical tone in the N2 time window, maximising at the frontal electrodes. The MMN will be similar to the MMN driven by McGurk illusion stimuli, as reported in the literature. Moreover, the result for accuracy in Experiment 7 suggests that the effect of incongruent visual information on the tone perception is rather weak, and this visual influence could possibly occur only on the level of physical feature (duration) processing. In auditory MMN studies of Mandarin lexical tones, evidence shows that MMN evoked through detecting the acoustic features of lexical tones (e.g. pitch variation) is small and tends to be right-lateralised (Li & Chen, 2015; Luo et al., 2006; Xi et al., 2010). Therefore, the expected MMN elicited by incongruent audiovisual lexical tones could be rather modest, and it could be right-hemispheric dominance.

In contrast, the MMN evoked by incongruent audiovisual vowel A/a/V/i/ could be radically different from that of incongruent lexical tones. Incongruent visual vowel information has a stronger effect on auditory vowel processing due to its highly distinguishable lip movements. From the results of the last experiment, the discrepancies between auditory /a/ and visual /i/ greatly hindered audiovisual vowel perception. Therefore, the prediction of the MMN evoked by the incongruent audiovisual vowel might have two directions. First, it could be similar to the McGurk MMN reported in the literature, because an incongruent visual vowel fused with a dubbed auditory vowel forms a novel vowel percept. Second, due to the large audiovisual discrepancies, MMN could be more like the visual MMN reported in Files et al. (2013), which is activated in the posterior temporal or the MMN evoked by non-audiovisual-illusion stimuli reported in Besle et al. (2005), which is activated in the bilateral occipital area.

4.2.2 Method

4.2.2.1 Participants

Fourteen native Mandarin speakers (aged: 21.1 ± 1.8 years; 4 females) from the Psychology and Education Department of Shenzhen University in China participated in the experiment. The data of four participants were deleted in the final analysis due to excessive artefacts. All of them were right-handed. All participants reported normal or correct to normal visual acuity and had no previous hearing impairments. The participants were compensated in accordance with a protocol approved by the Bournemouth University Review Board and Shenzhen University.

4.2.2.2 Materials

Three audiovisual syllables were used in the experiment: the congruent audiovisual syllable /ǎ/ (both auditory and visual inputs are /ɑ/ in T3), the incongruent audiovisual lexical tone A/ǎ/V/à/ ($A_{T3}V_{T4}$) and the incongruent audiovisual vowel A/a/V/i/. The stimuli materials were the same as those used in the last experiment. When making the incongruent syllables, a dubbed auditory track replaced the original auditory track of the congruent audiovisual syllable, and it aligned with the position where the replaced track was located so that it was able to match the mouth movement in the video track. In the lexical tone oddball sequence, congruent /ǎ/ (T3) as a standard stimulus and incongruent syllable A/ǎ/V/à/ ($A_{T3}V_{T4}$) as a deviant stimulus were presented in a quasi-random order. In the vowel oddball sequence, congruent /ɑ/ was the standard stimulus and incongruent A/a/V/i/ was the deviant one. The target stimulus, interspersed in sequences, was a black-and-white version of either the congruent or incongruent stimuli.[7] For both the lexical tone and vowel sequences, the auditory input remained the same while the visual input was different across conditions. A male native speaker of Mandarin produced two exemplars for each syllable. This can reduce the low-level feature processing of the speaker's idiosyncrasies; therefore, participants pay more attention to specific features of the audiovisual syllable, such as acoustic differences or visual information differences.

The length of the audiovisual stimuli was consistent at 56 frames (1868.54 ms), in which the front silence-gap (only visual input) before auditory onset remained at 23 frames (767.43 ms) (see Figure 4.1 in Section 4.1.2.2.). In natural speech, mouth movements always occur before sound is produced. For the syllables used in the current study, the lip-movement onset time was about 83 ms to 517 ms earlier than the auditory signal. The duration of the visual input (from mouth opening to mouth closure) was from 1,034 ms to 1,535 ms (see Table 4.4). The video clips were edited in Adobe Premiere Pro CC (Adobe Systems, California) as pixels of 1280 × 720 clips at 29.97 frames per second (1 frame = 33.37 ms) and the soundtracks were edited in Adobe Audition CC (Adobe Systems, California). All auditory tracks were digitised at 48 kHz with 32-bit amplitude resolution and were RMS normalised in amplitude to -12 dB.

Table 4.4 Audiovisual stimuli durations, corresponding auditory and visual component durations and onset time (in ms).

Session	**Condition**	**Congruence**	**Token**	**Stimulus length**	**Lip onset**	**Audio onset**	**Visual syllable duration**	**Auditory syllable duration**
Tone	standard	Congruence	T3	1869	267	767	1535	830
	deviant	Incongruence	$A_{T3}V_{T4}$	1869	250	767	1268	830
Vowel	standard	Congruence	/ɑ/	1869	267	767	1535	830
	deviant	Incongruence	AaVi	1869	684	767	1034	830

[7] Black-and-white audiovisual syllables as the target stimuli consist of both standard and deviant black-and-white video, which aimed to prevent participants from using the strategy of identifying the target by relying on the type of task-irrelevant stimuli (congruent/incongruent syllables).

4.2.2.3 Procedure

The experiment consists of two discriminations: lexical tones and vowels. The task was to respond to audiovisual syllables in the black-and-white video and to ignore those in the coloured video. The probabilities of standards, deviants and targets for each session were 70%, 16% and 14%, respectively. The syllables were presented in a quasi-random order so that deviants and black-and-white stimuli would not appear in the first few trials and to avoid presenting consecutive deviants or targets. The ITI duration ranged from 1,000 to 1,400 ms. Each rest block contained 50 trials. In total, there were 600 trials in each session, and the whole experiment lasted for approximately 60 minutes. Only task-irrelevant trials (coloured video) with correct responses were included in the data analysis. The participants were instructed to respond to the task as quickly and accurately as possible and to avoid blinks as best they could during the syllable presentation. Five-minute practice trials were given before the experimental trials. The experiment was conducted in a sound-attenuated chamber. The participants sat facing a 17-inch LCD monitor at a viewing distance of about 60 cm. Sound was played through Etymotic ER-4P earphones (Etymotic Research, Elk Grove Village). The loudness of the presented sound was approximately 65 dB SPL. The experimental stimuli were presented through E-prime 2.0 (Psychology Software Tools, Sharpsburg).

4.2.2.4 EEG Recording

EEG signal was recorded at a sampling rate of 500 Hz with a Brain-Amp-MR amplifier (Brain Products GmbH, Gilching) and the Brain Vision Recorder 1.0 system (Brain Products GmbH, Gilching). There were 64 Ag/AgCl electrodes mounted on a 64-channel elastic cap (actiCap, Brain Products GmbH) arranged according to the international 10-20 system (Jasper, 1958). An extra electrode was placed under the left eye to monitor the electrooculograms (EOGs) that resulted from eye-blinking. The impedance of each channel was kept below 20 kΩ before recording. The ground electrode was set to AFz. Raw data were processed in EEGLAB (Delorme & Makeig, 2004) version 14.0.0 and ERPLAB

(Lopez-Calderon & Luck, 2014) tool boxes installed in Matlab R2014a (The MathWorks, Inc.). They were filtered offline with a bandpass filter from 0.1–30 Hz with a roll-off slope of 48 dB/octave. In addition, they were re-referenced off-line with common average reference (average of the whole head electrodes),[8] excluding EOG. Continuous EEG was segmented into 1,600 ms -epochs, starting at 200 ms before stimulus (video clip) onset and ending at 1,600 ms after stimulus onset, and the baseline [200 ms pre-stimulus to 768 ms (auditory onset)] was corrected. Trials with activities exceeding a voltage threshold of ± 100 μV (artefacts) were rejected using peak-to-peak moving window method, which was 200 ms in length with a window step of 100 ms. Approximately 80–85% of the recording was preserved. Individual averages with correct responses only were calculated.

Data Analysis

The analysis measured the MMN component by comparing the activities of incongruent responses and congruent responses. The auditory input in standard and deviant conditions remained identical, while the visual input was different in lexical tones (T3 or T4) and vowels (/ɑ/ or /i/). If mismatched visual information affected the auditory process, the auditory component of the incongruent syllable would be different from that of the congruent stimulus. Only auditory responses (waveforms after 767 ms) from the incongruent and congruent conditions were included in the analysis. With application of the passive oddball paradigm, MMN is calculated by subtracting the congruent response from the incongruent response, and MMN should maximise at the frontal area of the scalp at about 150-300 ms (P2-N2 time window) after auditory signal onset. Therefore, the analysis included the frontal electrodes. According to the literature on auditory MMN, it is reversed in polarity at the right and left mastoids (Alho, 1995). However, studies of McGurk MMN have not found inverted polarity at the mastoids, suggesting that the MMN evoked by a fused percept and typical auditory MMN could be different in terms of the generator in the brain (Colin et al., 2004). In the analysis, the electrodes of two mastoids were included. Because the visual input in congruent and incongruent conditions was not identical, it is possible that MMN would be evoked by visual input differences rather than

[8] EEG recorded from the electrodes in the left and right mastoids was included in the data analysis, therefore all the electrodes were referenced to their average voltage.

by the audiovisual integration process. On that point, the occipital electrodes were also included in the analysis. Hence the regions of interest (ROIs) were the frontal electrodes (Fz, F1, F2), the occipital electrodes (Oz, O1, O2) and the electrodes in left and right mastoid regions (TP9, TP10).

4.2.3 Results

4.2.3.1 Results for lexical tones

The time window for mean amplitude analysis of lexical tones was from 180 to 280 ms, covering the time range from P2 peak to N2 peak of the auditory responses. Two-way ANOVAs (Congruence × Electrode) were submitted to different ROIs, and subsequent pair-wise comparisons were corrected with Bonferroni approach. As shown in Table 4.5 and Figure 4.6, a significant difference was found in the main effect of Congruence at the frontal electrodes [$F(1, 9) = 5.776$, $p = .040$], but there was no significant effect found in the occipital [$F(1, 9) = 0.280$, $p = .659$] or mastoid areas [$F(1, 9) = 0.724$, $p = .417$]. This suggests that lexical tone incongruent response was significantly larger than the congruent response in negativity in the frontal region. Further pairwise comparisons of Congruence based on each electrode in the frontal region showed significant negativity at F1 ($p = .019$) and F2 ($p = .045$), but not at Fz ($p = .280$). Individual electrodes on the occipital and mastoid areas had no significant effect on Congruence. An extra t-test was run to compare F1 and F2 in difference voltages between incongruent and congruent responses, but it did not show a significant result: $t(10) = 1.313$, $p = .222$, which suggests that the negativity evoked by incongruent lexical tones in the frontal electrodes was more bilateral.

Table 4.5 Mean amplitude (μV) of lexical tone response (N = 10) in standard (congruent) and deviant (incongruent) conditions at selected electrodes within a time interval of 180-280 ms after auditory onset.

ROI	**Electrode**	**Mean Amplitude (μV)**				**Main effect**	
		Std	**SE**	**Dev**	**SE**	$\boldsymbol{F(1,9)}$	***p***
Frontal	Fz	-0.02	±0.16	-0.25	±0.25	5.776	0.040
	F1	1.30	±0.93	1.02	±0.90		
	F2	1.45	±1.12	0.90	±1.03		
	Avg.	0.91	±0.66	0.56	±0.65		
Occipital	OZ	0.16	±0.56	0.21	±0.80	0.208	0.659
	O1	0.80	±0.52	0.72	±0.69		
	O2	-0.01	±0.56	0.37	±0.66		
	Avg.	0.32	±0.50	0.44	±0.69		
Mastoid	TP9	-0.83	±0.49	-0.99	±0.63	0.724	0.417
	TP10	-0.53	±0.40	-0.78	±0.51		
	Avg.	-0.68	±0.44	-0.89	±0.54		

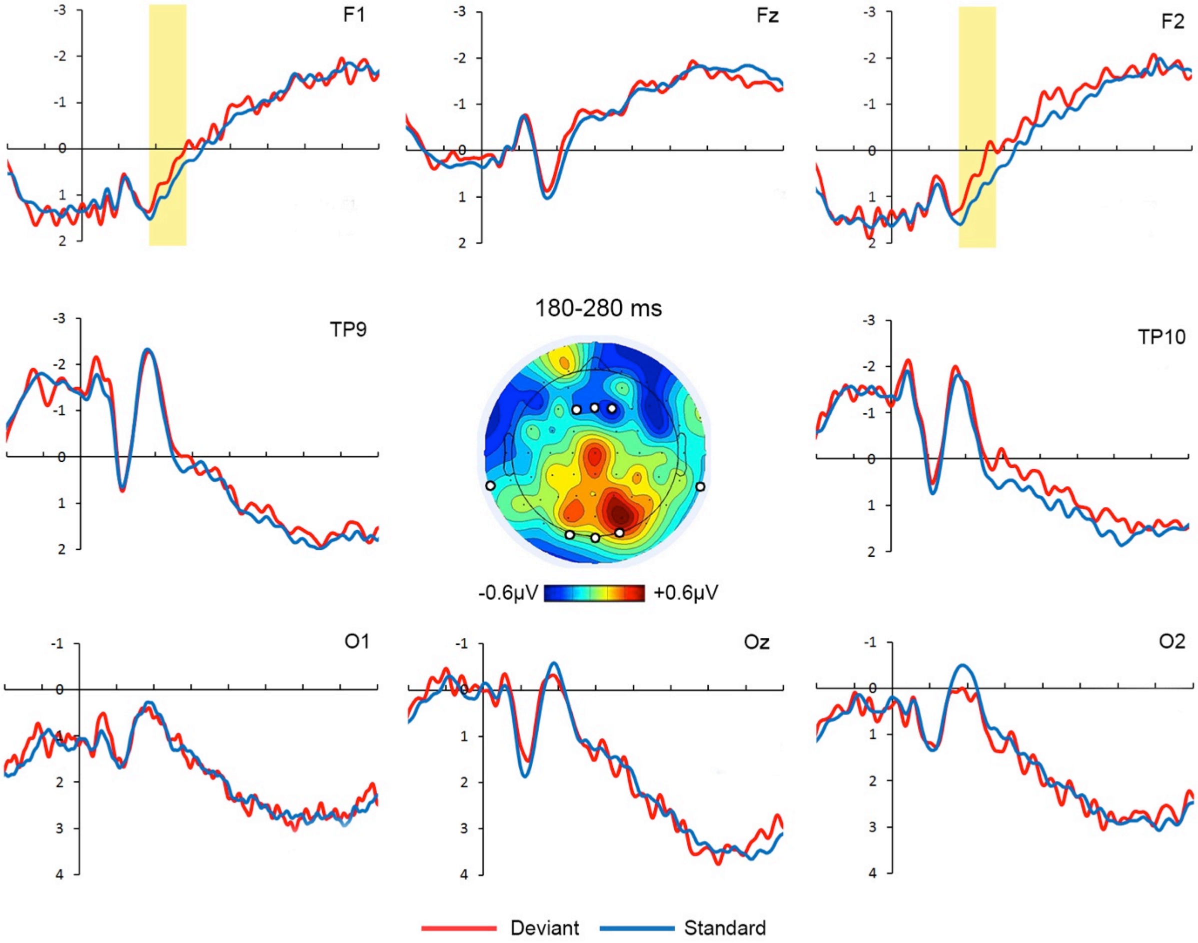


Figure 4.6 Grand average of ERPs of lexical tone response (*N* = 10) to standard (congruent) and deviant (incongruent) audiovisual syllables. 0 ms time point refers to auditory signal onset. The topography represents the different activity of deviant and standard responses within the time interval (180–280 ms). The analysed electrodes are marked as white dots on the topography. The yellow shaded area indicates the time interval during which the comparison was significant (*p* < .05).

4.2.3.2 Results for vowels

For the vowel ERPs, a majority of the electrodes showed that the ERPs for congruent and incongruent conditions started with a large deviation from auditory P2 and lasted for a long period of time (about 400–500 ms). Therefore, the time window of vowels for the statistical analysis was set at 180–480 ms after auditory onset.

Similarly, two-way ANOVAs were applied to the mean amplitudes in the three ROIs. As shown in Table 4.6, the main effect of Congruence was significant at the occipital [$F(1, 9) = 14.363, p = .004$] and mastoid electrodes [$F(1, 9) = 7.153, p = .025$], but not at the frontal electrodes, $F(1, 9) = 0.13, p = .727$. In other words, there was significant negativity from the difference activity between incongruent response and congruent response at the occipital and mastoid electrodes, but not in the frontal area. Specifically, pairwise comparisons showed that all the selected electrodes in the occipital area had significant negativity (Oz: p = .003; O1: p = .046; O2: p = .010). The negativity in the mastoid area was significant at the electrode on the left mastoid TP9 (p = .049), but it was only marginally significant at the right mastoid TP10 (p = .053) (see Figure 4.7).

Table 4.6 Mean amplitude (μV) of vowel response in standard and deviant conditions at selected electrodes within the time interval 180–480 ms after auditory onset.

ROI	**Electrode**	**Mean Amplitude (μV)**				**Main effect**	
		Std	**±SE**	**Dev**	**±SE**	**$F(1, 9)$**	***p***
Frontal	Fz	-0.60	±0.19	-0.49	±0.23	0.13	.727
	F1	-0.15	±0.48	-0.15	±0.48		
	F2	-0.35	±0.37	-0.20	±0.44		
	Avg.	-0.37	±0.32	-0.28	±0.32		
Occipital	OZ	1.09	±0.64	0.20	±0.66	14.363	.004
	O1	0.50	±1.11	-0.08	±0.97		
	O2	0.66	±0.76	-0.15	±0.74		
	Avg.	0.75	±0.76	-0.01	±0.71		
Mastoid	TP9	0.62	±0.58	-0.20	±0.62	7.153	.025
	TP10	0.92	±0.58	-0.01	±0.66		
	Avg.	0.77	±0.56	-0.11	±0.61		

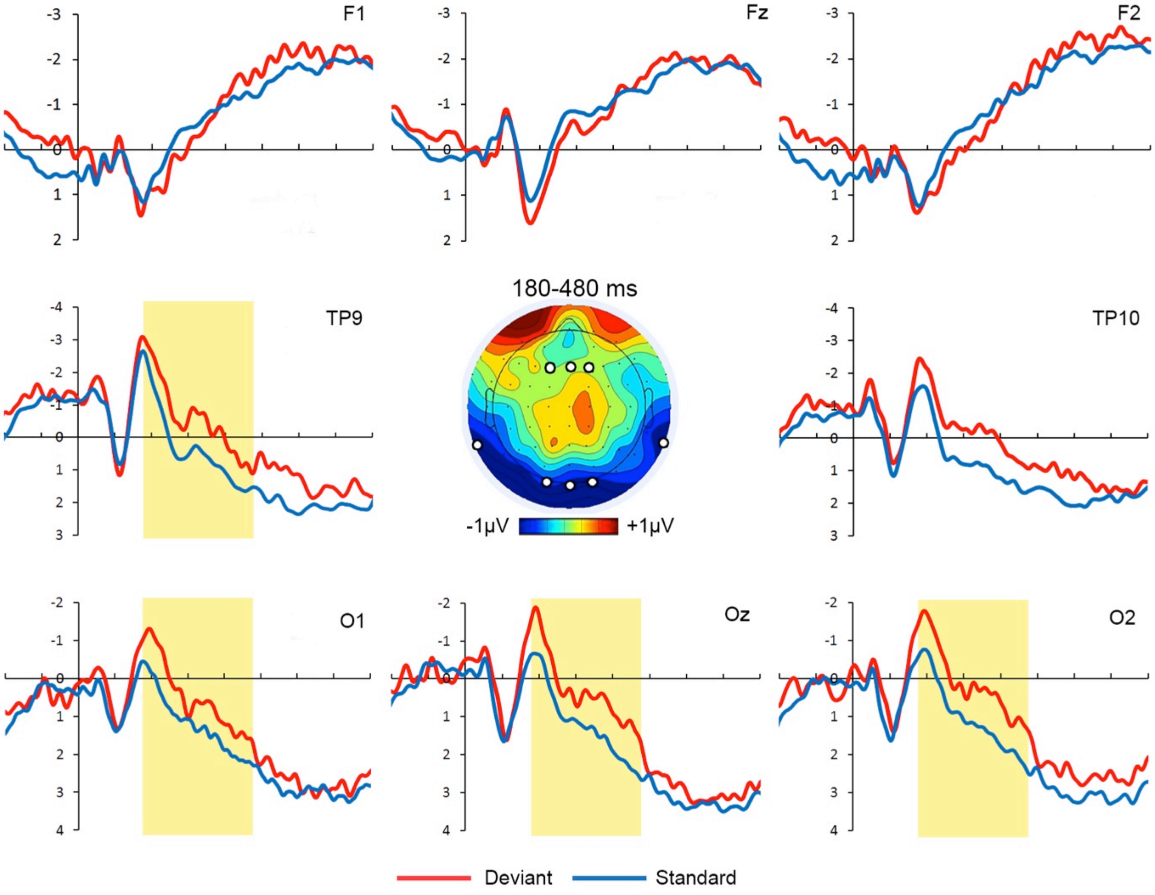


Figure 4.7 Grand average ERPs of vowel response ($N = 10$) to standard (congruent) and deviant (incongruent) audiovisual syllables. 0 ms time point refers to auditory signal onset. The topography represents the different activity of deviant and standard responses within the time interval (180–480 ms). The analysed electrodes were marked as white dots on the topography. The yellow shaded area indicates the time interval during which the comparison was significant ($p < .05$).

4.2.3.3 Summary of results

First, deviant-evoked negativity was found in both speech conditions, but it was very different in the scalp distribution, strength and duration of activity. The negativity of the lexical tone response evoked by $A_{T3}V_{T4}$ was analogous to McGurk MMN, and it was distributed over the anterior frontal electrodes from the auditory P2 peak, lasting about 100–200 ms (see Figure 4.6). In contrast, the vowel negativity evoked by AaVi was similar to visual MMN, which was maximised at the occipital area and spread over the bilateral posterior temporal sites (see Figure 4.7), starting at about 180 ms and ending at about 380–650 ms. The results suggest that the audiovisual integration of incongruent lexical tones and vowels could represent two different processes, which should be closely associated with the way that incongruent visual information affects auditory speech processing.

Lexical tone MMN

Unlike traditional auditory MMN, which maximises at Fz and two mastoids or at central midline sites (Luck, 2005), lexical tone MMN was distributed over the anterior frontal and bilateral temporal sites, and the MMN was not significant and not polarity-reversed at either mastoid region. MMN with a distribution of supratemporal and frontal activities suggests that the source of auditory MMN could be at the locus of the supratemporal cortices (Colin et al., 2004; Colin et al., 2002b). Therefore, the current observation suggests that the MMN of lexical tones was possibly generated in a brain area different from the primary auditory cortex.

The MMN of lexical tones was much weaker in amplitude and shorter in latency duration compared to vowel MMN. Based on the literature, (auditory) MMN can indicate a behavioural discrimination threshold (Kujala et al., 2007; Näätänen et al., 2007; Sams et al., 1985). A larger difference between a standard stimulus and a deviant stimulus evokes stronger MMN, while a smaller difference leads to weaker MMN. Moreover, a weak McGurk effect can reduce the elicitation of McGurk-MMN (Eskelund et al., 2015). Therefore, the small MMN in lexical tone results suggests that the effect of incongruent visual information on auditory lexical tone processing is rather weak. The auditory difference between the incongruent lexical tone $A_{T3}V_{T4}$ and the congruent lexical tone T3 could be small. However, as the auditory tones (T3) were identical in incongruent $A_{T3}V_{T4}$ and congruent T3 stimuli, the deflection evoked by incongruent $A_{T3}V_{T4}$ should come from the interaction between visual T4 and auditory T3. In other words, the effect of incongruent visual information on auditory lexical tone processing is modest but does exist.

Vowel MMN

MMN was found in the vowel response, but the scalp distribution was very different from traditional auditory MMN or McGurk MMN. The negativity was spread over the posterior electrodes, particularly at the occipital and bilateral posterior temporal electrodes on the scalp, which resembles the distribution of visual MMN. This result indicates that vowel MMN might not reflect incongruent audiovisual vowel integration (detection of an illusory auditory vowel) but represents the activity of detecting visual information difference between incongruent and congruent vowels. This could be because incongruent auditory and visual vowels have less interaction before MMN occurs. Besle et al. (2005) reported that the MMN evoked by an audiovisual deviant was sensory-specific when the combination of auditory and visual inputs did not lead to an illusory percept in a non-speech audiovisual event (e.g. a circle simultaneously presented with a tone sound). Specifically, an audiovisual deviant differing in visual input from an audiovisual standard generated an MMN that maximised over the bilateral occipital area at a latency of 192 ms, which was very similar to visual MMN. MMN driven by visual speech stimuli seems difficult to achieve. In some reports, visual MMN failed to be observed in visual-speech-only conditions (Colin et al., 2004; Colin et al., 2002b; Ponton et al., 2009; Saint-Amour et al., 2007; Sams et al., 1991), but Files et al. (2013) found clear visual MMN elicited by visual CV syllables in bilateral posterior temporal areas. Additionally, visual MMN of non-speech stimuli has frequently been recorded over both the occipital and posterior temporal areas of the cortex (Czigler, 2014; Pazo-Alvarez et al., 2003). The results of Besle et al. (2005) suggest that auditory input and visual input were processed independently in the absence of an audiovisual illusion; consequently, audiovisual MMN was more specific to the processing of a deviant modality. Thus, it is possible that vowel MMN mainly represents the process of detecting visual vowel deviants.

Another possibility is that vowel MMN could reflect a process of detecting audiovisual discrepancies in incongruent vowels. Unlike an artificial audiovisual event in a non-speech stimulus, the perception of the incongruent vowel (AaVi) in the current experiment could lead to an illusory vowel percept /ə/.[9] At the same time, a clear discrepancy would

[9] Although the experiment did not directly test the illusory vowel generated by the incongruent audiovisual vowel, most of the participants reported that they heard /ə/ or the vowel in the middle of the auditory and visual vowel inputs when the incongruent vowel (AaVi) was presented.

inevitably be noticed due to the significant distinction between auditory and visual features (Summerfield & McGrath, 1984). Kushnerenko et al. (2008) reported a similar scalp distribution from the difference activity between the syllable AbVg (fusion McGurk) and the syllable AgVb (combination McGurk) recorded from infants. AgVb does not generate a unified percept but rather a consonant cluster /bg/. AgVb elicited a negative deflection, mainly at the posterior temporal site on the left hemisphere and the temporal site on the right hemisphere within 290–590 ms. The authors explained that negativity could reflect detecting a conflict between auditory and visual inputs. MMN evoked by an incongruent audiovisual vowel could also reflect a process of detecting a mismatch between the auditory signal /ɑ/ and mouth movement of /i/ more than a process of the audiovisual integration of a fused vowel.

4.3 Discussion

This chapter focuses on incongruent audiovisual lexical tone perception in behavioural and brain activity (MMN) with incongruent audiovisual vowel perception as a comparison condition. Behavioural and MMN data provided clear results that the perception of audiovisual lexical tones was not identical in incongruent and congruent conditions which differed in visual input. This indicates that incongruent visual information modulates the perception of auditory lexical tones. Experiment 7 found that incongruent visual lexical tones changed the duration perception of lexical tones. Experiment 8 found that incongruent lexical tone deviance elicited MMN. The existence of the incongruent effect of audiovisual lexical tones further supports incongruent visual information altering lexical tone perception.

These experiment results cannot directly answer the question regarding specific visual information that contributes to changes in auditory processing. However, the area of the face containing visual features (such as articulating movement) is restricted to the mouth region, as all visual stimuli presented in the experiments were in the speaker's lower face only (without showing the upper head or neck). Thus, eyebrow, laryngeal muscle and even head movements are unlikely to be visual cues contributing to the incongruent effect for lexical tones. According to the results of Experiment 7, visual duration (the time from mouth opening to mouth closure) could be a key feature for changing auditory perception (that shortens perceived duration in $A_{T3}V_{T4}$). However, visual duration only affected the processing time of auditory tones but not lexical tone identification, which suggests that incongruent visual information (duration) does not interact with auditory tones on the level of phonetic processing.

In Experiment 8, MMN could be evoked by the incongruent lexical tone $A_{T3}V_{T4}$, which further provides evidence that auditory T3 modified by visual T4 can be discriminated from original T3. MMN was elicited during an early stage of auditory processing, from about 174 ms (within auditory P2 time range) to about 350 ms after auditory onset. MMN represents the pre-attentive activity of automatic detection of a deviant auditory signal over a standard auditory signal. Because the auditory tone remained identical across conditions, MMN at the frontal electrodes after auditory signal onset indicates that auditory tone perception was modulated by incongruent visual information before MMN was evoked. This suggests that audiovisual integration of incongruent lexical tones occurred in the pre-attentive stage of processing, or even earlier.

Limitations

There are limitations associated with the experimental design of the two experiments discussed above. In Experiment 7, the lack of an effect of incongruent visual information on lexical tones in the identification rate is possibly because the noise mask was not strong enough to efficiently reduce the identification close to the perception threshold between T3 and T4. Because the perception of lexical tones greatly depended on F0 cue rather than visual cues, adding a stronger noise mask might potentially have further limited the use of acoustic cues and enhanced the influence of visual information on auditory tone identification.

In Experiment 8, the distraction task might not have been able to completely shift attention away from task irrelevant stimuli to the targets. The task was colour identification (responding to black-and-white video clips), which was an easy task and did not require close attention to conduct. The deviant stimuli in vowels could still involuntarily catch perceivers' attention due to their large discrepancy. This may explain why vowel MMN was possibly influenced more by the visual vowel /i/, so that MMN was strongly activated at the occipital and bilateral temporal sites as visual MMN distribution. In a future study, the distraction task could be more engaging, requiring perceivers to direct their full attention to the task stimuli. Moreover, the modality of the stimuli for the distraction task needs to be auditory-based (e.g. beep sound identification) or a modality other than audiovisual modality (e.g. a tactile identification task) so that visual deviation across standard and deviant conditions can be less attended to.

In sum, the behavioural and MMN results found evidence that incongruent visual information influences auditory tone processing. The mismatched visual information of T4 shortened the duration perception of auditory tone T3. MMN result confirms that incongruent visual information interacts with the early (pre-attentive) processing of auditory lexical tones.

Chapter 5 General discussion and conclusion

The studies (Experiments 1-8) in the previous chapters found empirical evidence that, in audiovisual lexical tone perception, visual information from articulatory movements affects the perception of auditory lexical tones. Furthermore, the studies demonstrated that audiovisual lexical tone perception comprises two levels (non-phonetic and phonetic) of audiovisual integration processing. In the following sections, the main findings specific to these two processes in audiovisual lexical tone perception will be highlighted and their theoretical implications for audiovisual speech integration will be discussed.

5.1 Main findings

This dissertation focuses on the audiovisual perception of Mandarin lexical tones in terms of the audiovisual benefit effect and the McGurk effect. That is, the studies in this dissertation aimed to address two main issues: 1. Can visual information facilitate the perception of Mandarin lexical tones? 2. Can incongruent visual information change the perception of Mandarin lexical tones?

These questions cannot be fully answered without touching on an underlying process of audiovisual speech integration: how the information that visual speech provides interacts with the processing of auditory speech in Mandarin lexical tones. Based on relevant empirical research and reviews, visual speech provides two types of information. First, phonetic-specific information that is extracted from the articulatory movements of the mouth, lips, teeth and jaw (conveying information such as place of articulation, lip roundedness/ opening, duration and loudness) and from the movement of eyebrows and head (conveying F0 information) complements speech perception. Second, non-phonetic information signalled by the onset of visual speech movement predicts the timing of ensuing auditory speech, hence increasing the sensitivity to detecting auditory speech. These two types of visual information suggest that there are two processes of audiovisual speech integration. Non-phonetic audiovisual integration is general perceptual processing that reflects the temporal relationship between auditory and visual inputs, whereas phonetic audiovisual integration is linguistic processing that requires the mapping of speech-specific articulatory movement onto phonetic/ phonological representation in long-term memory.

In the audiovisual perception of Mandarin lexical tones, the behavioural and electrophysiological results of the studies in Chapters 2–4 suggest two main findings. First, visual information facilitates lexical tone perception in congruent audiovisual speech and visual information changes the durational perception of lexical tones in incongruent audiovisual speech. Second, there are two types of visual information that influence lexical tone perception: non-phonetic and phonetic visual information, suggesting two levels of audiovisual integration processing. These findings for audiovisual lexical tones support the hypothesis of multiple-stage processing in audiovisual speech integration. Before discussing the theoretical implications in more details, the main findings concerning the two types of visual information reported in Experiments 1–8 will be summarised.

5.1.1 Audiovisual benefit effect of lexical tone perception

First, the behavioural results in Experiments 1-4 showed clear evidence that the perception of audiovisual lexical tones were better than that of auditory lexical tones in terms of identification accuracy and discrimination in noise conditions, suggesting that visual information for lexical tones is used effectively to facilitate lexical tone perception. Based on the results of Experiments 2–4, the strength of the audiovisual benefit effect was different across individual tones or tone contrasts. In particular, it was stronger in the tone contrast between T3 and T4, but weaker in the tone contrast between T2 and T3. This suggests that the audiovisual benefit effect is also influenced by the phonetic visual features of lexical tones. Regarding the acoustic features of duration or F0 contour, they are contrastive between T3 and T4, but similar between T2 and T3. Therefore, it is reasonable to suggest that duration or F0 contour could have stronger visual correlates in audiovisual lexical tones perception.

The most substantial differences in the acoustic features of T3 and T4 were tone duration and F0 contour. T3 has the longest tone duration, while T4 has the shortest. T3 has a falling-rising contour but T4 has a falling contour. In terms of tone production, pitch contour is closely related to tone duration. The complex contour of T3 involves more complicated muscle excitation and hence requires longer duration for implementation (Zhang, 2002). Because a certain amount of time is needed for the transition between high and low pitches, the duration of the transition affects the pitch contour (Xu, 1998). In auditory tone perception, a longer tone duration tends to enhance the perception of pitch contour, but when the duration is shorter than 90 ms, pitch contour is not reliably perceivable (Greenberg & Zee, 1979). Moreover, it takes longer to produce a rising contour tone compared to a falling one (Sundberg, 1973), which explains why T3 is much longer than T4. In terms of the acoustic features of T2 and T3, there are many similarities in pitch contour and duration. The pitch contours of both tones fall at the beginning, before the turning point, and then gradually rise towards the end (see Figure 2.2). The temporal location of the turning point serves as a crucial cue to perceptually distinguish T2 and T3 (Moore & Jongman, 1997). T3 pitch transition at the turning point is lower and larger than T2, thus T3 duration is slightly longer than T2. A longer T2 is easily perceivable as T3 (Blicher et al., 1990). These similar features of T2 and T3 account for the ambiguity between T2 and T3 in their perception.

Second, the findings for lexical tones in lip-reading performance imply that phonetic visual information (duration or F0 contour) was used in audiovisual lexical tone perception. Lexical tones were able to be identified or discriminated when the auditory signal was removed, which indicates that phonetic-/ tonetic-specific visual cues are available for Mandarin lexical tones. Although lexical tones are difficult to distinguish visually, the results of Experiments 1 and 2 demonstrated that lexical tone lip-reading performance in the VO condition was slightly better than chance. Additionally, the results in Experiment 2 showed that tone contrasts that involved T4 (especially T2-T4 and T3-T4) in visual discrimination were better than other contrasts, while visual discrimination in the contrast of T2-T3 was rather poor. The results suggest that the visual feature of T4 stands out from T2 or T3. As mentioned in the paragraph above, T2 and T3 are similar in pitch contour (falling-rising), which is contrastive to the pitch contour of T4 (falling). In terms of duration, T4 is much shorter compared to T2 and T3. This suggests that F0 contour and duration could be transmittable to tone articulatory movement as useful visual cues in recognising or discriminating visual tones.

Thus, the audiovisual benefit effect seems to be related to lip-reading performance in individual lexical tones. The discrimination of T3-T4 had the strongest benefit effect among other contrasts, and T3-T4 also achieved better lip-reading performance in the VO condition. The discrimination of T2-T3 had the weakest audiovisual benefit effect, it was also worse in the VO condition. This suggests that the visual cues used for lexical tone lip-leading may also be used to contribute to the audiovisual lexical tone benefit effect. That is, the visual cues for audiovisual lexical tone perception are phonetic-specific (e.g. duration) and convey tonetic information complementing acoustic cues.

However, when it comes to explaining the weak audiovisual benefit effect or even the reverse effect (audiovisual inhibition) in the contrast of T2-T3, visual duration cues cannot fully explain why T2-T3 had a poor audiovisual benefit effect. The discrimination of T2-T3 became worse when adding visual input rather than not including visual input. The visual durations of these two tones are indeed less contrastive; thus, visual duration cue was not helpful for their discrimination. However, they were not the only pair that was close in duration. The durations of T1 and T3 were even closer. There might be other visual cues that caused the low/reverse audiovisual benefit effect in T2-T3. One possible explanation is that the visual information used in T2-T3 discrimination could be a non-phonetic visual cue. As explained earlier, visual information can enhance perceivers' sensitivity to the auditory signal without actually provide phonetic information. A non-phonetic visual cue improves the sensitivity to auditory T2 and T3. However, T2 and T3 are the most confusable tones in auditory perception owing to their similar pitch contours. Even if the visual cue enhances perceivers' sensitivity to auditory tone signals, T2 and T3 would still be difficult to distinguish.

In general, the behavioural findings confirm that visual information facilitates lexical tone perception. The stronger audiovisual benefit effect in T3-T4 and lip-reading performance suggest that visual information that contribute to lexical tone perception is related to phonetic visual information (e.g. duration). Moreover, the weaker audiovisual benefit or audiovisual inhibition in T2-T3 suggests that non-phonetic visual cues might be involved in audiovisual lexical tone perception. Phonetic visual cues (e.g. tone duration) can be more effectively captured when they are highly distinguishable between alternative tones like the tone contrast between T3 and T4. Non-phonetic visual information (timing cue) may always be used in audiovisual lexical perception.

5.1.2 Neural correlates of audiovisual lexical tone processing

The ERP study (Experiment 5) found direct evidence that visual information affects the sensory processing of auditory lexical tones at about 188 ms after auditory signal onset. This early audiovisual processing of lexical tones involved non-phonetic audiovisual integration. In Experiment 5, the auditory N1 evoked by audiovisual lexical tones was significantly reduced in amplitude and shortened in latency compared to the N1 evoked by auditory-only lexical tones. The results suggest that visual information alleviates and accelerates the sensory processing of auditory lexical tones in the auditory cortex.

The N1 reduction effect on audiovisual speech processing has been attributed to visual information anticipating upcoming auditory sound (Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). However, whether visual information predicts the content (phonetic information) or timing of ensuing auditory speech is controversial. van Wassenhove et al. (2005) proposed that preceding visual speech inputs activate possible visemic representations so that they predict the phonetic information of upcoming auditory speech. In their study, they suggested that more salient visual speech (e.g. bilabial /p/) has stronger predictability, and therefore results in a larger reduction effect in latency. In contrast, Stekelenburg and Vroomen (2007) argued that visual speech anticipates the timing of the upcoming auditory signal rather than its content. They found that a N1 reduction effect existed in congruent and incongruent audiovisual speech and non-speech audiovisual events; they also found that the N1 reduction effect disappeared when there was no anticipatory movement from visual input. Hence, their results indicate that the N1 reduction effect depends on whether visual input can predict the time when auditory signals will appear, but it does not depend on whether visual input can anticipate the content (phonetic information) of upcoming auditory signals. In the case of audiovisual lexical tones, articulatory movement physically precedes the auditory signal, but the preceding visual feature is unlikely to predict upcoming auditory tone information because the articulation of tone variation depends on variation in vocal fold vibration. Therefore, it is unlikely that the reduction effect of lexical tones is due to visual input that predicts tone-specific information preceding auditory tones. That is to say, N1 reduction in lexical tones is not due to a phonetic-specific visual information integration. The results for lexical tones seem to support the argument in Stekelenburg and Vroomen (2007) that the N1 reduction effect reflects non-phonetic audiovisual integration, instead of phonetic-specific audiovisual processing.

Furthermore, the findings revealed the time course when visual information interacts with auditory tone processing. The visual influence on auditory tones started at least as early as about 150 ms, maximising at about 188 ms (within the auditory N1 time range). In the ERP data, the time interval from stimulus onset to 200 ms is commonly regarded as sensory-specific processing, where activity in the N1 time range does not involve auditory feature extraction (Hillyard et al., 1998). Therefore, the reduction in N1 indicates that visual influence starts before auditory phonetic or phonemic processing. This suggests that the audiovisual integration of lexical tones is pre-linguistic (i.e. non-phonetic) processing. It further supports that audiovisual speech integration can be independent of the processing of saliency of visual speech features.

However, the results of the comparison of lexical tones and consonant responses in N1 reduction suggest that whether N1 reduction is phonetically related depends on whether speech has distinctive phonetic-specific visual information provided by preceding articulation movement (movement that occurs before the auditory signal). In Experiment 6, N1 reduction in amplitude and latency was found in the audiovisual consonant response. After comparing the two experiments, the results showed that N1 reduction was different in lexical tones and consonants in terms of latency and lateralisation. Specifically, the lexical tones were about 13 ms later than consonants in the latency of N1 reduction activity. If the N1 reduction effect does not involve a phonetic-specific process, then lexical tones and consonants should be the same in their reduction effects. Because the stimuli used in the two experiments were physically identical, non-phonetic visual information (for predicting the timing of the auditory signal) should be the same. Previous studies have found that the reduction in N1/P2 in amplitude is insensitive to the congruency of speech stimuli, whereas audiovisual shortening of the component latency is congruence-dependent

(Knowland et al., 2014; van Wassenhove et al., 2005). This suggests that the reduction in the amplitude and the shortening in latency reflect two different audiovisual integration mechanisms. Latency shortening may be due to preceding visual speech input (place of articulation) that initiates the phonetic-specific process. As van Wassenhove et al. (2005) proposed, visual features start being extracted to form an abstract representation, which is compatible with auditory input at a later stage. In contrast, amplitude reduction could reflect non-phonetic processing, such as the temporal relationship between auditory and visual inputs. Therefore, the latency difference between lexical tones and consonants in the reduction effect in N1 should be due to the saliency of visual information.

In addition, the different topographies of lateralisation in lexical tones and consonants seem to support the hypothesis of non-phonetic audiovisual integration. The N1 reduction in lexical tones was right-lateralised while consonant reduction was bi-lateralised (with a tendency to be left-lateralised). Left-hemispheric dominance is well known for the processing of speech (Broca, 1861; Wernicke, 1874). The processing of lexical tones is left-lateralised when lexical tones are perceived as speech information; otherwise, the processing is right-lateralised when lexical tones are perceived as acoustic features (e.g. pitch variation) (Gandour et al., 2004; Li & Chen, 2015; Luo et al., 2006; Shuai & Gong, 2014; Wang et al., 2001; Xi et al., 2010). Following this logic, lateralisation difference could suggest that the N1 reduction of consonants reflects phonetic-specific audiovisual integration, while N1 reduction of lexical tones reflects non-phonetic audiovisual integration.

Two important findings from ERP Experiments 5–6 can be highlighted for audiovisual integration processing in lexical tones. First, visual information influences or facilitates the processing of auditory lexical tones in the N1 time range (about 188 ms) before the stage when phonetic information starts to be processed. Second, the audiovisual integration (visual facilitation) of lexical tones at this stage could be non-phonetic processing.

5.1.3 Incongruent audiovisual lexical tones perception

The studies on the perception of incongruent audiovisual lexical tones (Experiments 7–8) found evidence that incongruent visual information is able to change auditory tone perception, which resembles the McGurk effect triggered by certain incongruent consonants. The findings of Experiment 7 indicate that incongruent visual information altered the duration perception of auditory tones but did not change the categorical perception of auditory lexical tones. Specifically, incongruent $A_{T3}V_{T4}$ syllables were reacted to faster than congruent T3 syllables, suggesting that the shorter visual tone (T4) speeded up the response time of the auditory tone (T3). The response time was related to the perception of tone duration. Perceiving a longer tone (T3) required a longer response time, and vice versa. Therefore, the shorter response time in perceiving $A_{T3}V_{T4}$ suggests $A_{T3}V_{T4}$ was perceived as shorter T3 due to the influence of visual T4. This is due to the auditory and visual inputs integrate in terms of duration feature (Green & Miller, 1985). That is, the auditory duration of T3 integrated with the visual duration of T4, resulting in perceiving a shorter T3 compared to the duration of the original T3. However, this incongruent visual duration did not influence the phonetic processing of auditory tones, as it did not alter the tone identification accuracy (change T3 to T4 in perception), even when the acoustic signal was degraded by noise. It suggests that the audiovisual integration of incongruent lexical tones may be independent of the phonetic process and instead be constrained to lower-level feature (duration) integration.

In the ERP study (Experiment 8), the MMN component was elicited by incongruent lexical tone $A_{T3}V_{T4}$ syllables within a 180–280 ms time range after auditory onset at the anterior frontal electrodes. The MMN of the incongruent lexical tone indicates that mismatched visual information modified the processing auditory tone in the pre-attentive stage or earlier. Moreover, the result for MMN suggests that visual modification could not reflect phonetic processing but rather the processing of lower features (e.g. duration). MMN is not exclusively sensitive to phonemic change; it is also sensitive to variations in acoustic features, such as frequency, duration and intensity (Näätänen & Winkler, 1999). In the literature on auditory MMN studies of Mandarin lexical tones, MMN response tends to be larger and left-lateralised if lexical tones are perceived as speech information, otherwise it tends be smaller and more right-lateralised if lexical tones are perceived as pitch variation (Li & Chen, 2015; Luo et al., 2006; Xi et al., 2010). In the distribution of MMN evoked by incongruent lexical tones, there was no clear pattern of lateralisation and the activity was rather small, suggesting that the processing might not be speech-related.

Generally, in audiovisual lexical tone perception, visual speech of Mandarin lexical tones certainly has an impact on auditory lexical tone perception. Visual information improves auditory tone identification and discrimination behaviourally, and it facilitates sensory auditory tone processing. Incongruent visual information is able to modify the perception of auditory lexical tones at the lower level of feature processing (e.g. duration) in the pre-attentive stage, at about 174 ms. Given the findings above, there are two types of visual information that contribute to the visual effect in Mandarin lexical tone perception. The audiovisual benefit effect in behavioural results (Experiments 2–4) supports the availability of a visual cue (duration) that is specific to lexical tone features, while the results of Experiments 5–8 indicate the existence of a visual cue that is not necessarily related to the phonetic identity of lexical tones. These empirical findings for audiovisual lexical tone perception favour the hypothesis that audiovisual speech integration has multiple stages of processing.

5.2 Theoretical implications of audiovisual lexical tone perception

Early and late audiovisual speech integration

In terms of theories of audiovisual speech integration, the main topic is the processing stage (early or late) when auditory and visual sensory information integrate (Massaro & Jesse, 2007; Schwartz et al., 1998). Early integration indicates that the integration of auditory and visual inputs has occurred prior to their sensory information being phonetically categorised. In contrast, late integration indicates that integration takes place after sensory information has engaged in phonetic processing (been mapped onto a phonetic/ phonemic prototype) (Green, 1998). The main question seems to concern whether audiovisual integration involves any language-specific processing. If integration is language-dependent, it may favour the late integration hypothesis; otherwise, it would be considered as early integration (Schwartz et al., 1998).

Schwartz et al. (1998) presented four audiovisual fusion models associated with speech perception that depict the processing levels of audiovisual integration. The first model is the direct identification model, which proposes that sensory-specific information (e.g. acoustic spectra and face parameters) can be directly transmitted to audiovisual representation by simplifying the processing from the sensory information stage to the representation forming stage. The second model is the dominant recoding model, where the auditory modality is dominant and visual input is recoded in auditory form prior to integrating with auditory information. The characteristics of the source (e.g. voiced, nasal) are only extracted from the auditory input. The third model is the motor recoding model. This model is based on the Motor Theory of Speech Perception that Liberman & Mattingly (1985) proposed, in which auditory input is processed through articulatory gestures of speech. The two sources are an amodal form of encoding (neither auditory nor visual modality), and they are integrated through a motoric configuration. The fourth model is the separate identification model, which is based on the Fuzzy Logical Model of Perception proposed in Massaro (1987). Auditory and visual inputs are independently processed and compared to phonetic prototypes; thereafter, the phonetic information from each modality is integrated. The first three models can be considered early integration models, while the last model (separate identification) accounts for a late integration model. These models appear to capture processes at a certain stage of audiovisual integration, but they do not explain that situation where both early and late integrations can happen during audiovisual speech processing.

Multiple-stage audiovisual speech integration

In contrast with the models of early or late audiovisual speech integration, audiovisual speech integration has been proposed to occur in multiple stages containing non-phonetic audiovisual integration and phonetic-specific audiovisual integration (Eskelund et al., 2011; Kim & Davis, 2014; Klucharev et al., 2003; Lalonde & Holt, 2016; Peelle & Sommers, 2015; Schwartz et al., 2004; Soto-Faraco & Alsius, 2009; Stekelenburg & Vroomen, 2007).

During audiovisual speech perception, auditory integration can be characterised as non-speech-specific or general perceptual processing. For example, preceding visual speech signals prepare or predict the timing of the ensuing auditory, which reduces uncertainty and increase the sensitivity of auditory signal detection (Bernstein et al., 2004; Grant & Seitz, 2000; Kim & Davis, 2004; Schwartz et al., 2004); the order between auditory and visual speech signals can be detected while experiencing the McGurk illusion (Munhall et al., 1996; Soto-Faraco & Alsius, 2007, 2009); visual rate/ visual speaking rate information affects the VOT perception of plosive consonants (Green & Miller, 1985). In phonetic-specific audiovisual integration, phonetic-specific information extracted from articulatory movement is a representation of phonetic/ visemic information retrieved from long-term memory. Visual phonetic information complements the acoustic information in speech. For example, the visual place of articulation conveys the second formant frequency of acoustic information in stop consonants, which is vulnerable in noise (e.g. /b/ vs. /d/). Therefore, phonetic visual information greatly benefits word recognition in noise (e.g. Sumby & Pollack, 1954). The McGurk illusion in consonants (e.g. AbVg is illusorily perceived as /d/) is also due to visual phonetic information integrating with auditory speech information.

Both levels of processing are important to audiovisual speech perception and clearly they are not mutually exclusive. Klucharev et al. (2003) observed a reduction effect (audiovisual integration) on vowels maximised as early as 85 ms after auditory onset, regardless of stimulus congruency; however, the difference response between congruent and incongruent stimuli started at later latency of 155 ms. The authors proposed that the earlier processing was non-phonetic and reflected a temporal-spatial property in audiovisual integration, as it was congruence-independent, while the later processing was phonetic-related audiovisual integration as it was sensitive to the congruency of the stimuli. Similarly, Stekelenburg and Vroomen (2007) found that the ERP response of audiovisual speech integration in N1 was independent of audiovisual congruency and was related to the period when the visual input precedes the upcoming auditory signal (i.e. the temporal relation between auditory and visual inputs), which suggests the reduction in N1 was non-phonetic audiovisual integration. The authors also found that the reduction in the later component P2 was sensitive to the congruency of audiovisual speech stimuli, which indicates that phonetic-specific integration processing occurs in a later time course. Soto-Faraco and Alsius (2007) reported that asynchronicity between incongruent auditory and visual speech signals was clearly detected, but at the same time audiovisual signals still integrated as a unified percept (in McGurk stimuli). Similarly, in the McGurk effect for audiovisual vowels, the discrepancy between auditory and visual vowels can still be detected yet simultaneously an illusory fusion vowel is also perceived (Summerfield & McGrath, 1984). These results suggest that both visual non-phonetic information processing (audiovisual temporal synchronicity) and phonetic information processing exist in audiovisual speech integration.

Multiple-stage audiovisual integration in lexical tones

In the perception of audiovisual lexical tones, both non-phonetic and phonetic visual information take part in the integrating process. The findings of the current studies support the hypothesis of multiple-stage integration which contains two levels of audiovisual speech integration at different stages.

First, the results for the stronger audiovisual benefit effect in T3-T4 contrast and the availability of visual cues in lexical tone lip-reading showed evidence that audiovisual integration was related to phonetic/ tone-specific information (e.g. tone duration) processing. Duration information can be captured from the period from mouth opening to mouth closure, which complements duration or F0 contour information for auditory lexical tones.

Moreover, the ERP studies of the N1 reduction effect demonstrated that audiovisual integration in an early stage (at about 188 ms) of auditory lexical tone processing was non-phonetic. For lexical tones, the N1 reduction effect or audiovisual integration in N1 was independent of the saliency or predictability of visual phonetic information (place of articulation). As for audiovisual lexical tone perception, visual information about place of articulation was not critical, but visual duration was. However, visual duration information cannot be attained from the onset of mouth movement; consequently, auditory tone information is not predictable only from the onset of mouth movement. The N1 reduction effect of lexical tones could be related to audiovisual processing on a non-phonetic level.

Additionally, in incongruent audiovisual lexical tones, incongruent visual information integrates with auditory tone perception on a non-phonetic level. The findings of the behavioural study revealed that incongruent visual information modulated the duration perception of auditory lexical tones, generating a percept of the same tone as the auditory input but with different duration. It suggests that this audiovisual lexical tone was not processed on a phonetic level but on a non-phonetic level. The ERP study found that incongruent lexical tones activated MMN in the frontal area within 180–280 ms after auditory signal onset. This indicates that incongruent audiovisual lexical tones integrate in the pre-attentive stage of processing.

Therefore, the findings of the studies in this project suggest that two levels (non-phonetic and phonetic) of audiovisual processing may take place in different stages during audiovisual lexical tone perception. In the ERP results, non-phonetic audiovisual integration was found in the early auditory component N1, which represents sensory processing before linguistic or other cognitive processing. The processing of phonetic visual information (tone duration cues) could occur in a stage later than non-phonetic, because one needs to perceive the time from onset to offset of mouth movement to capture the tone duration information, which takes longer to process. Thus, as demonstrated in Figure 5.1, the first stage is non-phonetic audiovisual integration where non-phonetic visual information (timing cue) predicts the upcoming auditory tone signal. The second stage is phonetic audiovisual integration where auditory tone information and visual tone information (e.g. duration information) integrate.

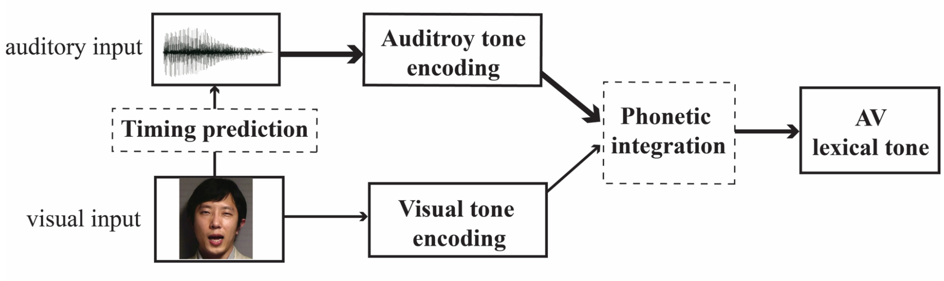


Figure 5.1 Schematic of the two stages of audiovisual lexical tone integration.

Between the two types of visual information, phonetic visual information may be less frequently used in audiovisual lexical tone perception. Phonetic visual information (duration cue) is not always beneficial to lexical tone perception. It is useful to distinguish tones that have a salient duration feature, but less helpful for tones that are not durational distinctive. Therefore, phonetic visual information might play less of a role in audiovisual lexical tone perception. Non-phonetic audiovisual integration is not determined by the saliency of visual duration information but relies on the temporal relation between auditory and visual inputs. Thus, a timing cue should always be used in audiovisual perception.

5.3 Practical implications

From the findings of the present studies, practical implications can be drawn for Mandarin lexical tone learning and teaching. The audiovisual benefit studies in this dissertation have shown that lexical tone perception is improved when presenting the speaker's face. Visual cues (e.g. duration) are useful in distinguishing certain lexical tones, particularly T3 and T4. Applying an approach of audiovisual feedback with auditory tones and videos of articulating movements could assist native and non-native speakers in Mandarin lexical tone learning.

Training in audiovisual has been proven to greatly improve the perception of a second language (e.g. Li, 2016). The perception and production of Mandarin lexical tones can be effectively improved by training with audiovisual feedback. So (2003) used an audiovisual feedback approach to train non-native speakers in Mandarin lexical tone production and perception. The training provided auditory sounds, pitch contours on graphs and tone descriptions in text. They found that the improvements in production and perception were better in an audiovisual feedback approach than in a simple feedback approach (correct/incorrect). So (2006) found that Mandarin lexical tone learning was improved by using auditory sounds and pitch contours in graphic animations as feedback information in training Cantonese, Japanese and English learners. Chen and Massaro (2008) trained Mandarin native speakers to use visual information by providing natural audiovisual lexical tone recordings (auditory sounds along with talking faces), and visual lexical tone recognition was greatly improved after training.

5.4 Limitations and further work

The studies in the dissertation have found that visual information has an impact on auditory tone perception in both audiovisual congruent and incongruent conditions, and the studies also found evidence that the visual effect was caused by both phonetic and non-phonetic visual cues. However, there are some improvements that can be made to further justify the two different types of audiovisual integration processes in future studies. In the studies in this dissertation, the two types of visual cues were not manipulated individually, hence the two corresponding processes of visual cues were difficult to separate. For example, in the experiments in Chapter 2, the behavioural findings for the audiovisual benefit effect could not give direct evidence that non-phonetic visual cues clearly played a part in visual facilitation, because the lexical tones were still lip-readable even though the lip-reading performance was low, suggesting a certain degree of residual phonetic visual information was still preserved. In future studies, to verify non-phonetic visual facilitation for lexical tones, visual input should be strictly controlled by removing lip-read (phonetic) information from the visual input. For example, the audiovisual benefit effect can be tested when visual articulation (especial lip movements) cannot be visualised clearly or when visual articulation is incongruent to dubbed auditory tones in audiovisual stimuli. In terms of phonetic-specific visual cues for lexical tones, although the evidence from Experiments 2–3 pointed to tone duration possibly being a crucial cue to improve tone perception in noise, especially in recognition/ discrimination between T3 and T4, tone duration was not directly controlled so as to be discrete from other possible visual features; therefore, it is difficult to draw a solid conclusion that a duration cue is the most essential phonetic visual cue for lexical tones. In future studies, visual duration can be further manipulated as an independent variable in experimental designs.

The studies in this dissertation comprised of sets of experiments on the audiovisual benefit effect in a congruent condition and the McGurk effect in an incongruent condition, respectively; consequently, it is unclear whether the incongruent lexical tons have the same audiovisual benefit effect as congruent lexical tones. Congruency could be set as a condition in studies of the audiovisual benefit effect. This design can be applied to ERP experiments in future research. According to Stekelenburg and Vroomen (2007), the N1 reduction effect is a congruence-independent process, and the reduction in later components (e.g. P2) is congruent-sensitive, which suggests two processes along with speech perception. Knowland et al. (2014) also found that reductions in N1/P2 in amplitude and latency responded differently to the congruency of speech stimuli. The reduction in component amplitude was insensitive to the congruency of stimuli whereas the shortening of component latency was affected by the congruency of speech stimuli, which suggests that there exist two different audiovisual speech processes in the early time range. For lexical tones, a comparison of brain activity between the congruent and incongruent conditions could help further dissect the non-phonetic and phonetic levels of processing of audiovisual lexical tones.

In addition to controlling visual features, different experimental paradigms, such as detection tasks, can be employed to test non-phonetic audiovisual integration in lexical tones in future studies. In the behavioural studies of the dissertation, two experimental paradigms, identification and discrimination, were adopted and an audiovisual benefit effect was found in both paradigms. Although these two paradigms are considered to engage different levels of speech processing, where identification requires higher-level (phonetic/ phonological) processing while discrimination encourages lower-level (non-phonetic) processing (Aslin & Smith, 1988), it is still possible that discrimination can involve both phonetic-specific and non-phonetic audiovisual processing (Lalonde and Holt, 2016). Comparatively, a detection paradigm can more directly measure whether the audiovisual benefit to speech in sensitivity to the auditory signal is due to an audiovisual temporal relationship but not the phonetic content of visual speech (Eskelund et al., 2011; Kim & Davis, 2004).

Additionally, there are some other methodological limitations of the studies that can be improved in future work. First, the number of trials was insufficient for each individual tone condition, such as the trials in Experiment 1. This could influence the stability of the data when analysing the audiovisual benefit effect for individual tones. Even though the studies in the dissertation did not measure individual differences, large individual differences in the ability to integrate audiovisual speech seem to be common in audiovisual speech research (e.g. Nath & Beauchamp, 2012). To overcome these issues, the trial numbers in each condition should be increased, so that data stability can be improved, and data deviation can be alleviated. Second, the number of stimulus exemplars should be increased. The stimuli (e.g. syllable /bai/ in four tones) used in Experiments 1–3 only contained a single articulation, which could lead to the participants developing a strategy of detecting low-level features (e.g. specific speakers' idiosyncrasies) rather than phonetic categorical differences while performing the task. Employing multiple tokens for each tone syllable could reduce the effect generated by this bias. Additionally, presenting the stimuli recorded by multiple speakers would also alleviate the bias from lower feature processing. In Experiment 8 (MMN evoked by incongruent lexical tones)，the methodological issues have three aspects. The deflection evoked between standard and deviant stimuli might be affected by physical differences in the visual input. In the oddball sequence, standard and deviant audiovisual syllables were different in their visual dimensions. Although the main comparison was of physically identical auditory sounds between standard and deviant ones, the preceding visual exogenous activity might have overlapped with subsequent auditory activity. Therefore, this might have contaminated the real auditory response modified by visual speech on the endogenous level. To solve this problem, an experiment design could introduce an additivity model (AV=A+V) to calculate the auditory mismatched activity under the influence of visual information (Besle et al., 2009). By subtracting visual-only activity from audiovisual activity, the possible visual mismatched activity can be reduced, and the visual activity evoked by the physical difference between standard and deviant can be controlled (Saint-Amour et al., 2007).

The findings of the studies in the dissertation are restricted to lexical tones in isolation monosyllables. As mentioned in Section 1.2, the four lexical tones in Mandarin are systematically different in their acoustic features, such as F0 contour and duration. The acoustic characteristics of a Mandarin lexical tone can be affected by a preceding or following tone in connected speech (Xu, 1997). For example, when a T3 syllable is followed by another T3 syllable, the first T3 becomes T2 or T2-like. That is, the F0 contour of the first T3 is changes to T2 contour (Peng, 2000). The duration of the four lexical tones is less distinguishable in connected speech. Yang et al. (2017) reported that lexical tone recognition with a duration cue was only 23%, suggesting that duration is not a reliable cue for lexical tone perception in connected speech. The variation in lexical tone acoustic features consequently has an impact on the visual features of lexical tones, especially on visual duration.

5.5 Conclusions

In summary, the studies in this dissertation have investigated the audiovisual perception of Mandarin lexical tones in terms of two effects: the audiovisual benefit effect and the McGurk effect. The audiovisual benefit effect of Mandarin lexical tones was found in both behavioural and ERP studies. The behavioural results provide evidence that presenting the visual movements of lexical tones along with auditory lexical tones facilitates the perception of lexical tones. The ERP findings further suggest that visual lexical tones have an influence on the early processing of auditory lexical tones by reducing auditory N1 responses. The McGurk effect of Mandarin lexical tones was also found in the behavioural and ERP studies. Incongruent visual lexical tones changed the perception of lexical tone duration, which activated MMN with a frontal distribution, suggesting that visual lexical tones modified auditory lexical tone perception, possibly in the auditory cortex. The audiovisual benefit effect and the McGurk effect of lexical tones are due to two types of visual information from visual lexical tones: a phonetic visual cue (possibly a tone duration cue) and a non-phonetic visual cue (possibly a timing cue). This suggests that there exist two different processes in different stages during audiovisual lexical tone perception.

The research presented in this dissertation fills a gap in the studies on lexical tone perception, and more widely in the domain of audiovisual speech perception. The current studies have explored some of the important issues not previously studied, such as the time course of audiovisual processing in lexical tones and the McGurk effect of audiovisual lexical tones in behavioural and brain responses. The present research affords new insights into the audiovisual integration mechanism of lexical tones and has theoretical and practical implications. However, the studies in the dissertation are just an initial stage in research on audiovisual lexical tones, there remain unanswered questions that are worth exploring in future studies.

Bibliography

Abramson, A. S. (1979). The noncategorical perception of tone categories in Thai. In B. Lindblom & S. Öhman (Eds.), *Frontiers of speech communication research* (pp. 127-134). London: Academic Press.

Alain, C., & Tremblay, K. (2007). The role of event-related brain potentials in assessing central auditory processing. *J Am Acad Audiol, 18*(7), 573-589.

Alho, K. (1995). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear Hear, 16*(1), 38-51.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*(2), 339-355. doi: 10.1348/000712601162220

Aslin, R. N., & Smith, L. B. (1988). Perceptual development. *Annual review of psychology, 39*, 435-473. doi: 10.1146/annurev.ps.39.020188.002251

Barth, D. S., Goldberg, N., Brett, B., & Di, S. (1995). The spatiotemporal organization of auditory, visual, and auditory-visual evoked potentials in rat cortex. *Brain Res, 678*(1-2), 177-190.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1), 5-18. doi: 10.1016/j.specom.2004.10.011

Besle, J., Bertrand, O., & Giard, M. H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hear Res, 258*(1-2), 143-151. doi: 10.1016/j.heares.2009.06.016

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur J Neurosci, 20*(8), 2225-2234. doi: 10.1111/j.1460-9568.2004.03670.x

Besle, J., Fort, A., & Giard, M. H. (2005). Is the auditory sensory memory sensitive to visual information? *Exp Brain Res, 166*(3-4), 337-344. doi: 10.1007/s00221-005-2375-x

Blair, R. C., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology, 30*(5), 518-524.

Blicher, D. L., Diehl, R., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the mandarin tone2/tone3 distinction: evidence of auditory enhencement. *Journal of Phonetics, 18*, 37-49.

Boersman, P., & Weenink, D. (2013). Praat: doing phonetics by computer. *[Computer program].*(Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/).

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphéme (perte de la parole). *Bull Soc Anat Paris, 36*, 330-357.

Bryden, M. P. (1982). *Laterality: Functional Asymmetry in the Intact Brain*. New York: Academic Press.

Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S. (2000). *Perception of visual information for Cantonese tones*. Paper presented at the Eighth Australian International Conference on Speech Science and Technology, pp. 86–91, Canberra, Australia.

Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., Rattanasone, N. X., Schwarz, I.-C., & Webster, D. (2014). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics, 36*(6), 1459-1491. doi: 10.1017/S0142716414000496

Burnham, D., & Lau, S. (1998). *The Effect of Tonal Information on Auditory Reliance in the McGurk Effect*. Paper presented at the international conference on auditory0visual speech processing (AVSP) 1998, pp., Terrigal-Sydney, Australia.

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*. Paper presented at the the International Conference on Audio-Visual Speech Processing (AVSP) 2001, pp. 155-160, Aalborg, Denmark.

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H. C., Ciocca, V., Morris, R. H., 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*. Paper presented at the the 7th International Seminar on Speech Production, pp. 185-192, Ubatuba, Brazil.

Burnham, D. K., Attina, V., & Kasisopa, B. (2011). *Auditory-visual discrimination and identification of lexical tone within and across tone languages*. Paper presented at the International Conference on Audio-Visual Speech Processing (AVSP) 2011, pp. 37-42, Volterra, Italy.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci, 363*(1493), 1001-1010. doi: 10.1098/rstb.2007.2155

Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996, 3-6 Oct 1996). *About the relationship between eyebrow movements and Fo variations.* Paper presented at the Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.

Chang, C., & Yao, Y. (2007). *Tone production in whispered Mandarin*. Paper presented at the Proceedings of the 16th International Congress of Phonetic Sciences, pp. 1085-1088, Dudweiler, Germany.

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

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Percept Psychophys, 66*(5), 820-836.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *J Acoust Soc Am, 123*(4), 2356-2366. doi: 10.1121/1.2839004

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *J Acoust Soc Am, 126*(2), 858-865. doi: 10.1121/1.3158823

Cheng, Y.-Y., & Lee, C.-Y. (2018). The Development of Mismatch Responses to Mandarin Lexical Tone in 12- to 24-Month-Old Infants. *Front Psychol, 9*, 448. doi: 10.3389/fpsyg.2018.00448

Colin, C., Radeau, M., Soquet, A., Dachy, B., & Deltenre, P. (2002a). Electrophysiology of spatial scene analysis: the mismatch negativity (MMN) is sensitive to the ventriloquism illusion. *Clinical Neurophysiology, 113*(4), 507-518. doi: 10.1016/S1388-2457(02)00028-7

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clin Neurophysiol, 115*(9), 1989-2000. doi: 10.1016/j.clinph.2004.03.027

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002b). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin Neurophysiol, 113*(4), 495-506.

Crook, J. (2012). Audacity. In A. Brown & G. Wilson (Eds.), *The Architecture of Open Source Applications* (pp. 15-28). California: Creative Commons Attribution.

Curry, F. K. W. (1967). A Comparison of Left-Handed and Right-Handed Subjects on Verbal and Non-Verbal Dichotic Listening Tasks. *Cortex, 3*(3), 343-352. doi: 10.1016/S0010-9452(67)80022-4

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Percept Psychophys, 59*(2), 165-179.

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. doi: 10.1016/j.specom.2010.02.006

Czigler, I. (2014). Visual mismatch negativity and categorization. *Brain Topogr, 27*(4), 590-598. doi: 10.1007/s10548-013-0316-8

Davis, C., Kislyuk, D., Kim, J., & Sams, M. (2008). The effect of viewing speech on auditory speech processing is different in the left and right hemispheres. *Brain Res, 1242*, 151-161. doi: 10.1016/j.brainres.2008.04.077

de la Vaux, S. K., & Massaro, D. (2004). Audiovisual speech gating: Examining information and information processing. *Cognitive Processing, 5*, 106-112. doi: 10.1007/s10339-004-0014-2

Dees, M. T., Bradlow, A., Dhar, S., & Wong, C. M. P. (2007). *Effects of Noise on Lexical Tone Perception by Native and Non-native Listeners*. Paper presented at the International Congress of Phonetic Sciences (ICPhS) 2007, pp. 817-820, Saarbrücken, Germany.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods, 134*(1), 9-21. doi: 10.1016/j.jneumeth.2003.10.009

Department of Edcation. (2017, September 7). *Pupils across England start intensive lessons in Mandarin* [Press release]. Retrieved from https://www.gov.uk/government/news/pupils-across-england-start-intensive-lessons-in-mandarin

Dohen, M., & Lœvenbruck, H. (2005). *Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study.* Paper presented at the Interspeech 2005-Eurospeech. the 9th European Conference on Speech Communication and Technology, Lisbonne, Portugal.

Dohen, M., & Lœvenbruck, H. (2006). *Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variabilities*. Paper presented at the Speech Prosody 2006, pp. 221-224, Dresden, Germany.

Dohen, M., & Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Lang Speech, 52*(Pt 2-3), 177-206. doi: 10.1177/0023830909103166

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford: Oxford University Press.

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. (1961). Interval of Time Uncertainty in Auditory Detection. *J Acoust Soc Am, 33*(6), 771-778. doi: 10.1121/1.1908795

Eskelund, K., MacDonald, E. N., & Andersen, T. S. (2015). Face configuration affects speech perception: Evidence from a McGurk mismatch negativity study. *Neuropsychologia, 66*, 48-54. doi: 10.1016/j.neuropsychologia.2014.10.021

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. *Exp Brain Res, 208*(3), 447-457. doi: 10.1007/s00221-010-2495-9

Files, B. T., Auer, E. T., Jr., & Bernstein, L. E. (2013). The visual mismatch negativity elicited with visual speech stimuli. *Front Hum Neurosci, 7*, 371. doi: 10.3389/fnhum.2013.00371

Fok, C. Y.-Y. (1974). *A perceptual study of tones in Cantonese*. [Hong Kong]: Centre of Asian Studies, University of Hong Kong.

Francis, A. L., Ciocca, V., & Ng, B. K. (2003). On the (non)categorical perception of lexical tones. *Percept Psychophys, 65*(7), 1029-1044.

Fu, Q.-J., & Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing, 5*(1), 45-57. doi: 10.1179/136132800807547582

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., Li, X., & Lowe, M. (2004). Hemispheric roles in the perception of speech prosody. *Neuroimage, 23*(1), 344-357. doi: 10.1016/j.neuroimage.2004.06.004

Gandour, J. T. (1984). Tone dissimilarity judgments by Chinese listeners. *Journal of Chinese Linguisitcs, 12*, 235-261.

Gerrits, E., & Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Percept Psychophys, 66*(3), 363-376.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *J Acoust Soc Am, 109*(5), 2272-2275. doi: 10.1121/1.1362687

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am, 108*(3 Pt 1), 1197-1208.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory–visual consonant recognition. *J Acoust Soc Am, 100*(4), 2415-2424. doi: 10.1121/1.417950

Green, K. (1998). The use of auditory and visual information during phonetic processing: implication for theories of speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: advances in psychology of speechreading and audio-visual speech* (pp. 3-25). New York NY: Psychology Press.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Percept Psychophys, 38*(3), 269-276.

Greenberg, S., & Zee, E. (1979). On the perception of contour tones. *UCLA Working Papers in Phonetics, 45*, 150-165.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*(12), 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x

Hayes, B. (2009). *Introductory phonology*. Malden; Oxford: Wiley-Blackwell.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52*(11), 996-1009. doi: https://doi.org/10.1016/j.specom.2010.05.003

Hessler, D., Jonkers, R., Stowe, L., & Bastiaanse, R. (2013). The whole is more than the sum of its parts - audiovisual processing of phonemes investigated with ERPs. *Brain Lang, 124*(3), 213-224. doi: 10.1016/j.bandl.2012.12.006

Hillyard, S. A., Teder-Salejarvi, W. A., & Munte, T. F. (1998). Temporal dynamics of early perceptual processing. *Curr Opin Neurobiol, 8*(2), 202-210.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge [England]; New York: Cambridge University Press.

Huang, C.-Y., Yang, H.-M., Sher, Y.-J., Lin, Y.-H., & Wu, J.-L. (2005). Speech intelligibility of Mandarin-speaking deaf children with cochlear implants. *International Journal of Pediatric Otorhinolaryngology, 69*(4), 505-511. doi: https://doi.org/10.1016/j.ijporl.2004.10.017

Huhn, Z., Szirtes, G., Lorincz, A., & Csepe, V. (2009). Perception based method for the investigation of audiovisual integration of speech. *Neurosci Lett, 465*(3), 204-209. doi: 10.1016/j.neulet.2009.08.077

Jasper, H. H. (1958). Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalography and Clinical Neurophysiology, 10*(2), 370-375. doi: 10.1016/0013-4694(58)90053-1

Jennings, J. R., & Wood, C. C. (1976). Letter: The epsilon-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology, 13*(3), 277-278.

Jesse, A., & Janse, E. (2012). Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners. *Language and Cognitive Processes, 27*(7-8), 1167-1191. doi: 10.1080/01690965.2011.620335

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*(4), 359-384. doi: 10.1006/jpho.1999.0100

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. (2006). Perception and production of Mandarin Chinese tones. In P. Li, L. H. Tan, E. Bates & O. J. T. Tzeng (Eds.), *The Handbook of East Asian Psycholinguisitcs* (pp. 209-217). New York: Cambridge University Press.

Kent, R. D. (1997). *The speech sciences*. San Diego: Singular Publishing Group.

Kim, J., & Davis, C. (2001). *Visible speech cues and auditory detection of spoken sentences: an effect of degree of correlation between acoustic and visual properties.* Paper presented at the the International Conference on Audio-Visual Speech Processing (AVSP) 2001, Aalborg, Denmark.

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. *Speech Communication, 44*(1), 19-30. doi: 10.1016/j.specom.2004.09.008

Kim, J., & Davis, C. (2014). How visual timing and form information affect speech and non-speech processing. *Brain Lang, 137*, 86-90. doi: 10.1016/j.bandl.2014.07.012

Kim, J., Davis, C., & Groot, C. (2009). Speech identification in noise: Contribution of temporal, spectral, and visual speech cues. *J Acoust Soc Am, 126*(6), 3246-3257. doi: 10.1121/1.3250425

Kimura, D. (1973). The asymmetry of the human brain. *Sci Am, 228*(3), 70-78.

Kiriloff, C. (1969). On the Auditory Perception of Tones in Mandarin. *Phonetica, 20*, 63-67. doi: 10.1159/000259274

Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *J Cogn Neurosci, 20*(12), 2175-2184. doi: 10.1162/jocn.2008.20152

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research, 18*(1), 65-75. doi: 10.1016/j.cogbrainres.2003.09.004

Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. (2014). Audio-visual speech perception: a developmental ERP investigation. *Dev Sci, 17*(1), 110-124. doi: 10.1111/desc.12098

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication, 34*(4), 391-405. doi: https://doi.org/10.1016/S0167-6393(00)00058-3

Kujala, T., Tervaniemi, M., & Schroger, E. (2007). The mismatch negativity in cognitive and clinical neuroscience: theoretical and methodological considerations. *Biol Psychol, 74*(1), 1-19. doi: 10.1016/j.biopsycho.2006.06.001

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proc Natl Acad Sci U S A, 105*(32), 11442-11445. doi: 10.1073/pnas.0804275105

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: University Press.

Lalonde, K., & Holt, R. F. (2015). Preschoolers benefit from visually salient speech cues. *J Speech Lang Hear Res, 58*(1), 135-150. doi: 10.1044/2014_jslhr-h-13-0343

Lalonde, K., & Holt, R. F. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *J Acoust Soc Am, 139*(4), 1713-1723. doi: 10.1121/1.4945590

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *J Speech Lang Hear Res, 42*(3), 526-539.

Li, X., & Chen, Y. (2015). Representation and Processing of Lexical Tone and Tonal Variants: Evidence from the Mismatch Negativity. *PLoS One, 10*(12), e0143097. doi: 10.1371/journal.pone.0143097

Li, Y. (2016). Audiovisual Training Effects on L2 Speech Perception and Production. *International Journal of English Language Teaching, 3*(2), 14-23.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J Exp Psychol, 54*(5), 358-368.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Lang Speech, 47*(Pt 2), 109-138. doi: 10.1177/00238309040470020101

Liu, S. Y., Yu, G., Lee, L. A., Liu, T. C., Tsou, Y. T., Lai, T. J., & Wu, C. M. (2014). Audiovisual speech perception at various presentation levels in Mandarin-speaking adults with cochlear implants. *PLoS One, 9*(9), e107252. doi: 10.1371/journal.pone.0107252

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci, 8*(213). doi: 10.3389/fnhum.2014.00213

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear Hear, 19*(1), 1-36.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, Massachusetts: The MIT Press.

Luo, H., Ni, J. T., Li, Z. H., Li, X. O., Zhang, D. R., Zeng, F. G., & Chen, L. (2006). Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proc Natl Acad Sci U S A, 103*(51), 19558-19563. doi: 10.1073/pnas.0607065104

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Percept Psychophys, 24*(3), 253-257.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br J Audiol, 21*(2), 131-141.

Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Exp Brain Res, 233*(9), 2581-2586. doi: 10.1007/s00221-015-4324-7

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces : from speech perception to a behavioral principle*. Cambridge, Mass.: MIT Press.

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 19-35). Oxford: Oxford University Press.

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

Miki, K., Watanabe, S., & Kakigi, R. (2004). Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans: a magnetoencephalographic study. *Neurosci Lett, 357*(3), 199-202. doi: 10.1016/j.neulet.2003.12.082

Mixdorff, H., Charnvivit, P., & Burnham, D. (2005a). *Auditroy-visual Perception of Syllabic Tones in Thai*. Paper presented at the the International Conference on Audio-Visual Speech Processing (AVSP) 2005, pp. 3-8, Parksville, Canada.

Mixdorff, H., Hu, Y., & Burnham, D. (2005b). *Visual Cues in Mandarin Tone Perception*. Paper presented at the Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, pp. 405 - 408, Lisbon, Portugal.

Mixdorff, H., Luong, C. M., Nguyen, D. T., & Burnham, D. (2006). *Syllabic Tone Perception in Vietnamese*. Paper presented at the TAL 2006, pp. 137-142, La Rochelle, France.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *J Acoust Soc Am, 102*(3), 1864-1877. doi: 10.1121/1.420092

Moradi, S., Lidestam, B., & Ronnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. *Front Psychol, 4*, 359. doi: 10.3389/fpsyg.2013.00359

Moradi, S., Wahlin, A., Hällgren, M., Rönnberg, J., & Lidestam, B. (2017). The Efficacy of Short-term Gated Audiovisual Speech Training for Improving Auditory Sentence Identification in Noise in Elderly Hearing Aid Users. *Front Psychol, 8*, 368. doi: 10.3389/fpsyg.2017.00368

Möttönen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Res Cogn Brain Res, 13*(3), 417-425.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Percept Psychophys, 58*(3), 351-362.

Näätänen, R. (1990). The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and Brain Sciences, 13*(2), 201-233.

Näätänen, R., & Alho, K. (1997). Mismatch negativity--the measure for central sound representation accuracy. *Audiol Neurootol, 2*(5), 341-353.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin Neurophysiol, 118*(12), 2544-2590. doi: 10.1016/j.clinph.2007.04.026

Näätänen, R., & Picton, T. W. (1986). N2 and automatic versus controlled processes. *Electroencephalogr Clin Neurophysiol Suppl, 38*, 169-186.

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychol Bull, 125*(6), 826-859.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage, 59*(1), 781-787. doi: https://doi.org/10.1016/j.neuroimage.2011.07.024

Pazo-Alvarez, P., Cadaveira, F., & Amenedo, E. (2003). MMN in the visual modality: a review. *Biol Psychol, 63*(3), 199-236. doi: 10.1016/s0301-0511(03)00049-8

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169-181. doi: 10.1016/j.cortex.2015.03.006

Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S. Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics, 38*(4), 616-624. doi: https://doi.org/10.1016/j.wocn.2010.09.003

Peng, S.-H. (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In B. M. B & P. J (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 152-167). Cambridge: Cambridge University Press.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J Speech Lang Hear Res, 52*(4), 1073-1081. doi: 10.1044/1092-4388(2009/07-0276)

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253-260. doi: 10.3758/BF03214136

Ponton, C. W., Bernstein, L. E., & Auer, E. T., Jr. (2009). Mismatch negativity with visual-only and audiovisual speech. *Brain Topogr, 21*(3-4), 207-215. doi: 10.1007/s10548-009-0094-5

Reetz, H., & Jongman, A. (2009). *Phonetics:Transcription, Production, Acoustics, and Perception*. Chichester, West Sussex: Wiley-Blackwell.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: the psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *J Acoust Soc Am, 103*(6), 3677-3689.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7*, 31. doi: 10.3389/fnsys.2013.00031

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *Int J Audiol, 47 Suppl 2*, S99-105. doi: 10.1080/14992020802301167

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex, 17*(5), 1147-1153. doi: 10.1093/cercor/bhl024

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia, 45*(3), 587-597. doi: 10.1016/j.neuropsychologia.2006.03.036

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett, 127*(1), 141-145.

Sams, M., Paavilainen, P., Alho, K., & Naatanen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalogr Clin Neurophysiol, 62*(6), 437-448.

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Lang Speech, 52*(Pt 2-3), 135-175. doi: 10.1177/0023830909103165

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), B69-78. doi: 10.1016/j.cognition.2004.01.006

Schwartz, J. L., Robert-Bibes, J., & Escudier, P. (1998). Ten years after Summerfield: a taxonomy of modals for audio-visual fusion in speech perception In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing By Eye II* (pp. 85-108). East Sussex: Psychology Press Ltd.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. *Percept Psychophys, 59*(1), 73-80.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Dev Sci, 11*(2), 306-320. doi: 10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., Burnham, D., Tam, H., & Erdener, D. (2003). *Auditoryvisual speech perception development in Japanese and English speakers*. Paper presented at the International Conference on Auditory-Visual Speech Processing (AVSP) 2003, pp. 61-66, St. Jorioz, France.

Sekiyama, K., & Tohkura, Y. i. (1993). *Inter-Language Differences in the Influence of Visual Cues in Speech Perception* (Vol. 21).

Shankweiler, D., & Studdert-Kennedy, M. (1967). Identification of consonants and vowels presented to left and right ears. *Q J Exp Psychol, 19*(1), 59-63. doi: 10.1080/14640746708400069

Shi, R., Gao, J., Achim, A., & Li, A. (2017). Perception and Representation of Lexical Tones in Native Mandarin-Learning Infants and Toddlers. *Front Psychol, 8*, 1117. doi: 10.3389/fpsyg.2017.01117

Shuai, L., & Gong, T. (2014). Temporal relation between top-down and bottom-up processing in lexical tone perception. *Front Behav Neurosci, 8*, 97. doi: 10.3389/fnbeh.2014.00097

Smith, D., & Burnham, D. (2012). Faciliation of Mandarin tone perception by visual speech in clear and degraded audio: implications for cochlear implants. *J Acoust Soc Am, 131*(2), 1480-1489. doi: 10.1121/1.3672703

So, C. K. L. (2003). *Training non-native listeners to acquire Mandarin tones with visual and auditory feedback.* Paper presented at the the WorldCALL conference 2003, Banff, Alberta, Canada.

So, C. K. L. (2006). *Effects of L1 Prosodic Background and AV training on Learning Mandarin Tones by Speakers of Cantonese, Japanese, and English.* (Ph.D. dissertation), Simon Fraser University.

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport, 18*(4), 347-350. doi: 10.1097/WNR.0b013e32801776f9

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *J Exp Psychol Hum Percept Perform, 35*(2), 580-587. doi: 10.1037/a0013483

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J Cogn Neurosci, 19*(12), 1964-1973. doi: 10.1162/jocn.2007.19.12.1964

Studdert-Kennedy, M., & Shankweiler, D. (1970). Hemispheric specialization for speech perception. *J Acoust Soc Am, 48*(2), 579-594.

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: Psychology of Lipreading* (pp. 3-51). Hillsdale, NJ: Erlbaum.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philos Trans R Soc Lond B Biol Sci, 335*(1273), 71-78. doi: 10.1098/rstb.1992.0009

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology Section A, 36*(1), 51-74. doi: 10.1080/14640748408401503

Sundberg, J. (1973). *Data on maximum speed of pitch changes* (Vol. 4).

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*(2), 219-238. doi: https://doi.org/10.1016/j.wocn.2007.05.001

ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., & Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia, 63*, 43-50. doi: 10.1016/j.neuropsychologia.2014.08.008

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35*(2), 244-258. doi: https://doi.org/10.1016/j.wocn.2006.03.002

Tsao, F.-M. (2008). The Effect of Acoustical Similarity on Lexical-Tone Perception of One-Year-Old Mandarin-Learning Infants. [聲調相似度對漢語周歲嬰兒聲調知覺的影響]. *Chinese Journal of Psychology, 50*(2), 111-124. doi: 10.6129/cjp.2008.5002.01

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends Amplif, 11*(4), 233-241. doi: 10.1177/1084713807307409

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Cross-modal enhancement of speech detection in young and older adults: does signal content matter? *Ear Hear, 32*(5), 650-655. doi: 10.1097/AUD.0b013e31821a4578

Valkenier, B., Duyne, J. Y., Andringa, T. C., & Baskent, D. (2012). Audiovisual perception of congruent and incongruent Dutch front vowels. *J Speech Lang Hear Res, 55*(6), 1788-1801. doi: 10.1044/1092-4388(2012/11-0227)

van Hessen, A. J., & Schouten, M. E. (1992). Modeling phoneme perception. II: A model of stop consonant discrimination. *J Acoust Soc Am, 92*(4 Pt 1), 1856-1868.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A, 102*(4), 1181-1186. doi: 10.1073/pnas.0408949102

Wang, W. S.-Y. (1976). Language change. *Annals of the New York Academy of Sciences, 280*(1), 61-72. doi: doi:10.1111/j.1749-6632.1976.tb25472.x

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain Lang, 78*(3), 332-348. doi: 10.1006/brln.2001.2474

Watson, C. S., & Nichols, T. L. (1976). Detectability of auditory signals presented without defined observation intervals. *J Acoust Soc Am, 59*(3), 655-668. doi: 10.1121/1.380915

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *J Acoust Soc Am, 75*(6), 1866-1878. doi: 10.1121/1.390988

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Enie psychologische Studie auf anatomischer basis*. Breslau, Germany: Kohn und Weigert.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*(1), 25-47.

Wong, P., Schwartz, R. G., & Jenkins, J. J. (2005). Perception and Production of Lexical Tones by 3-Year-Old, Mandarin-Speaking Children. *Journal of Speech, Language, and Hearing Research, 48*(5), 1065-1079. doi: 10.1044/1092-4388(2005/074)

Xi, J., Zhang, L., Shu, H., Zhang, Y., & Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience, 170*(1), 223-231. doi: 10.1016/j.neuroscience.2010.06.077

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61-83. doi: 10.1006/jpho.1996.0034

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 55*(4), 179-203. doi: 28432

Yang, J., Zhang, Y., Li, A., & Xu, L. (2017). *On the Duration of Mandarin Tones*. Paper presented at the INTERSPEECH 2017, pp. 1407-1411, Stockholm, Sweden.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555-568. doi: 10.1006/jpho.2002.0165

Yip, M. (1995). Tone in Asian Languages. In J. Goldsmith (Ed.), *Handbook of Phonological* (pp. 476-494). Oxford: Basil Blackwell.

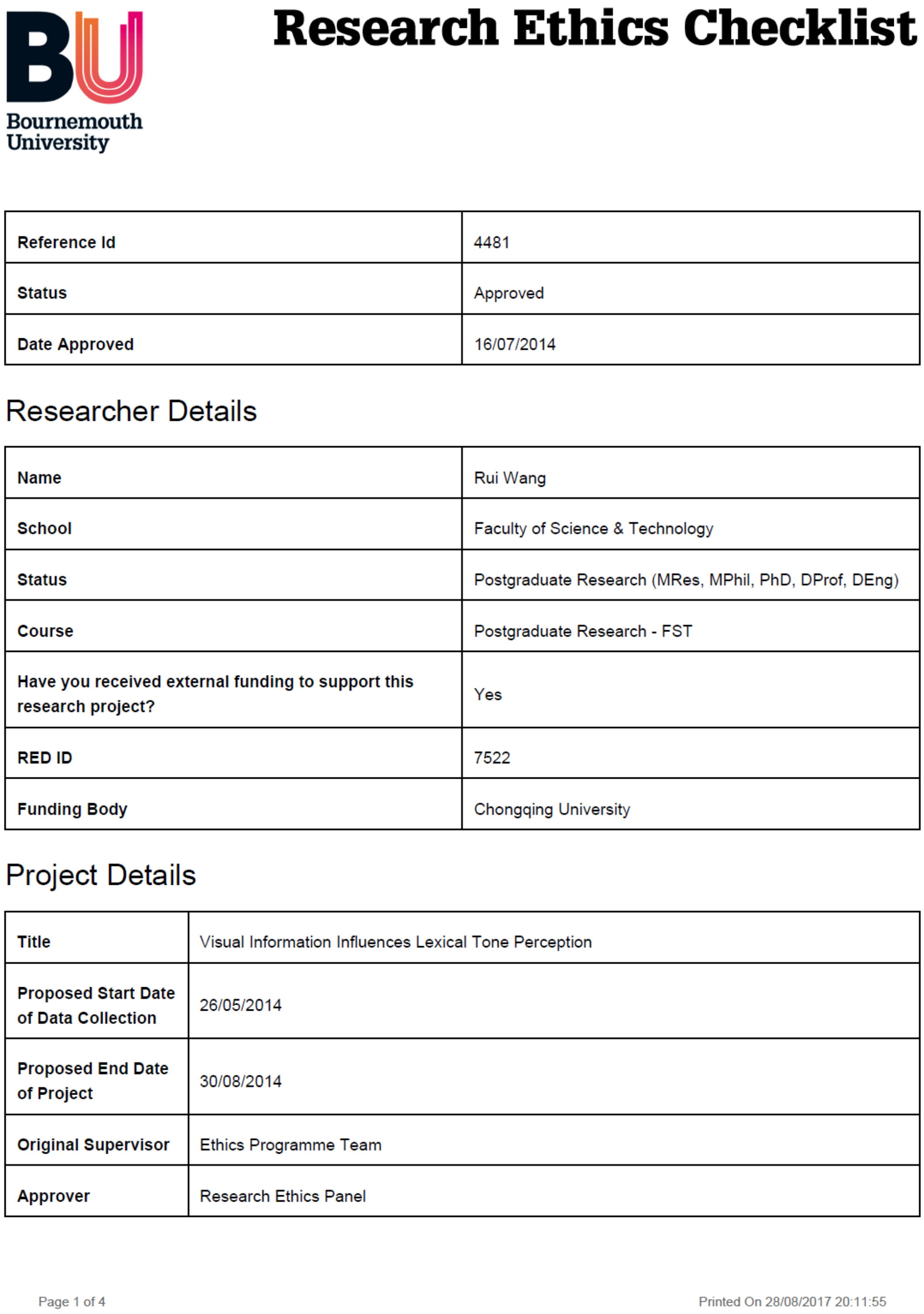
Yip, M. (2002). *Tone.* New York: Cambridge University Press.

Zhang, J. (2002). *The effets of duration and sonority on contour tone distribution-A typological survey and formal analysis.* New York: Routledge.

Zhang, J. (2014). Tones, Tonal Phonology and Tone Sandhi. In C. T. J. Huang, Y. H. A. Li & A. Simpson (Eds.), *The Handbook of Chinese Linguistics* (pp. 443-464). West Sussex: Wiley-Blackwell.

Appendices

Ethical Approval 1



Research Ethics Checklist

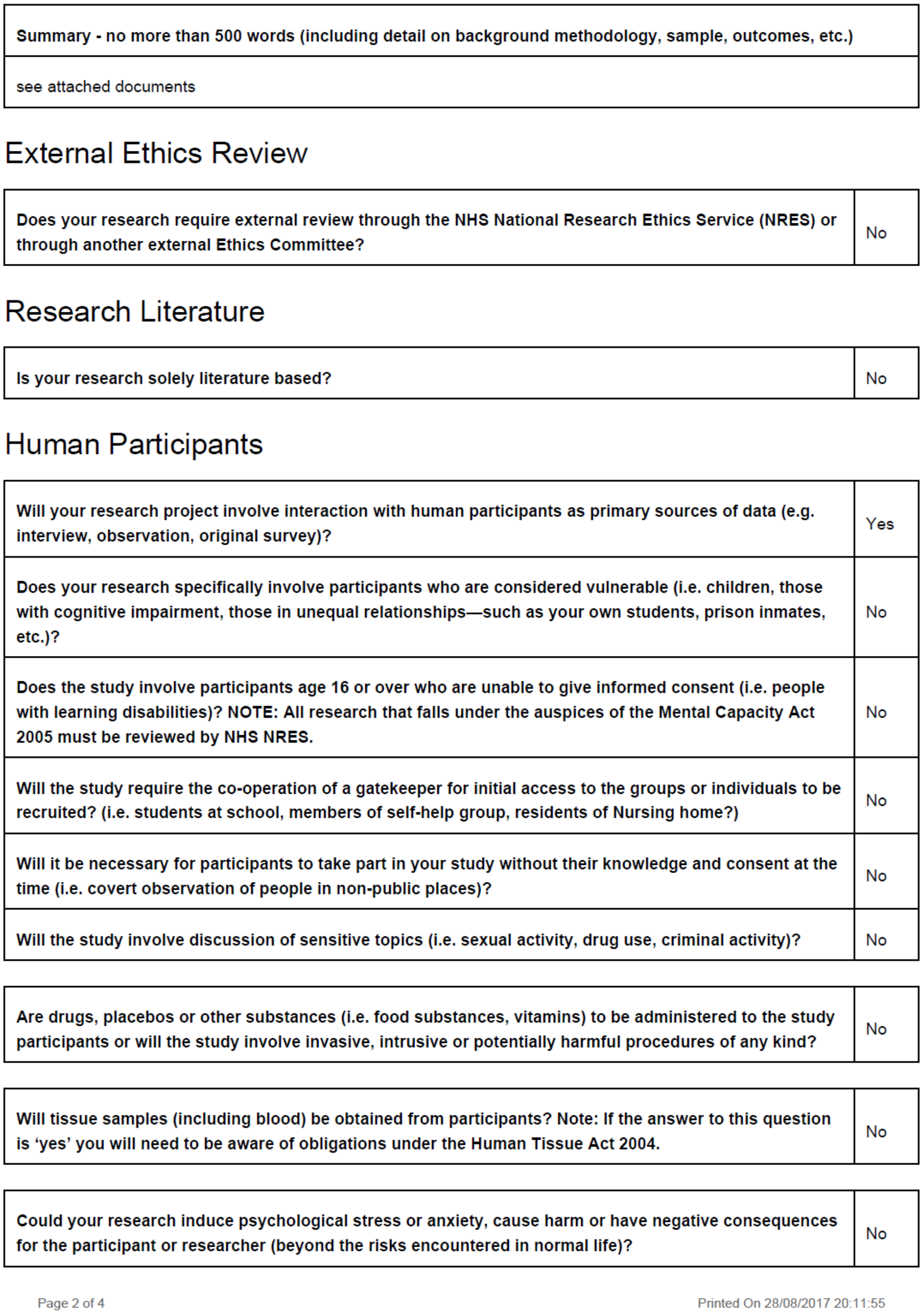
Reference Id	4481
Status	Approved
Date Approved	16/07/2014

Researcher Details

Name	Rui Wang
School	Faculty of Science & Technology
Status	Postgraduate Research (MRes, MPhil, PhD, DProf, DEng)
Course	Postgraduate Research - FST
Have you received external funding to support this research project?	Yes
RED ID	7522
Funding Body	Chongqing University

Project Details

Title	Visual Information Influences Lexical Tone Perception
Proposed Start Date of Data Collection	26/05/2014
Proposed End Date of Project	30/08/2014
Original Supervisor	Ethics Programme Team
Approver	Research Ethics Panel



Summary - no more than 500 words (including detail on background methodology, sample, outcomes, etc.)
see attached documents

External Ethics Review

Does your research require external review through the NHS National Research Ethics Service (NRES) or through another external Ethics Committee?	No

Research Literature

Is your research solely literature based?	No

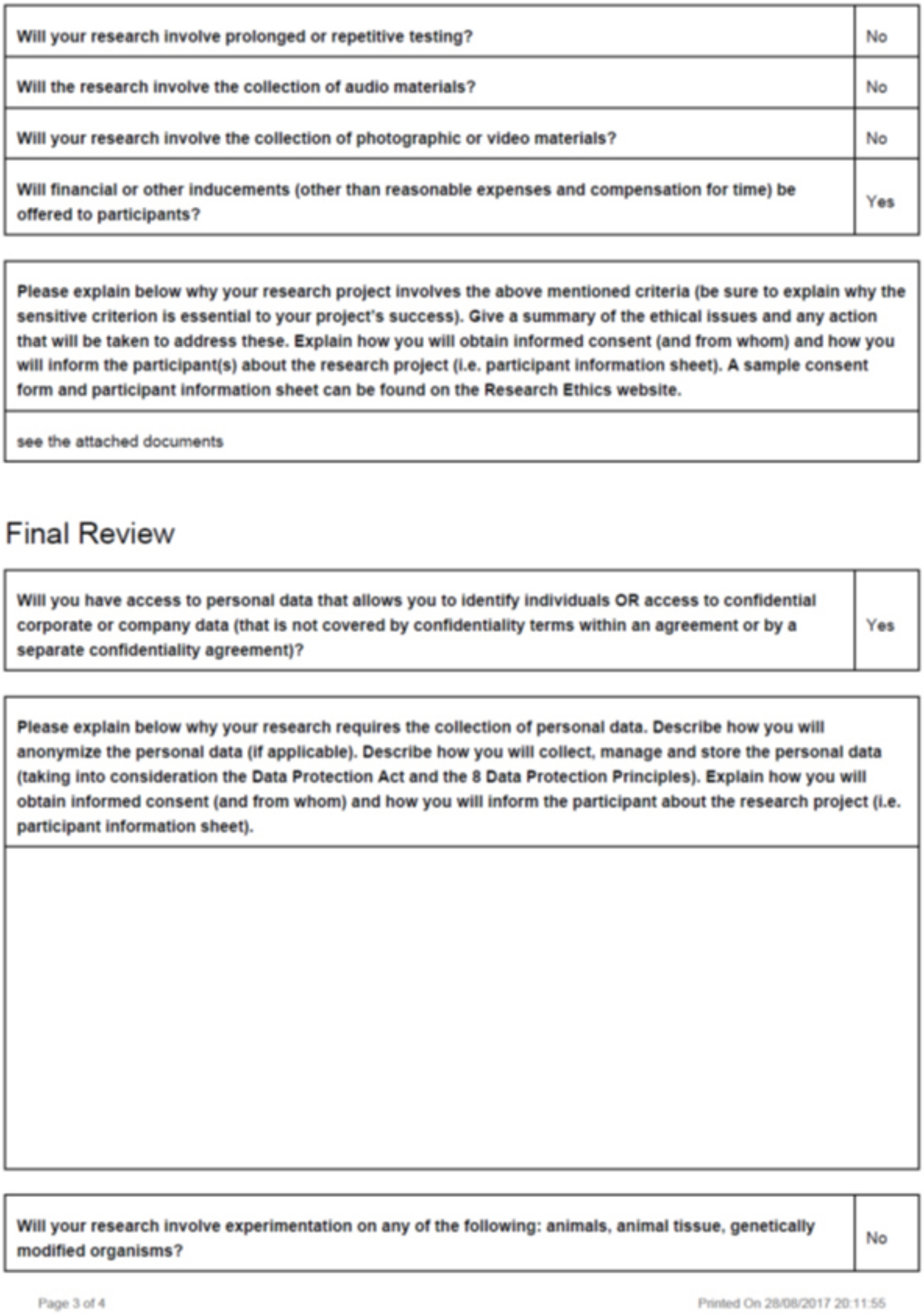
Human Participants

Will your research project involve interaction with human participants as primary sources of data (e.g. interview, observation, original survey)?	Yes
Does your research specifically involve participants who are considered vulnerable (i.e. children, those with cognitive impairment, those in unequal relationships—such as your own students, prison inmates, etc.)?	No
Does the study involve participants age 16 or over who are unable to give informed consent (i.e. people with learning disabilities)? NOTE: All research that falls under the auspices of the Mental Capacity Act 2005 must be reviewed by NHS NRES.	No
Will the study require the co-operation of a gatekeeper for initial access to the groups or individuals to be recruited? (i.e. students at school, members of self-help group, residents of Nursing home?)	No
Will it be necessary for participants to take part in your study without their knowledge and consent at the time (i.e. covert observation of people in non-public places)?	No
Will the study involve discussion of sensitive topics (i.e. sexual activity, drug use, criminal activity)?	No

Are drugs, placebos or other substances (i.e. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	No

Will tissue samples (including blood) be obtained from participants? Note: If the answer to this question is 'yes' you will need to be aware of obligations under the Human Tissue Act 2004.	No

Could your research induce psychological stress or anxiety, cause harm or have negative consequences for the participant or researcher (beyond the risks encountered in normal life)?	No



Will your research involve prolonged or repetitive testing?	No
Will the research involve the collection of audio materials?	No
Will your research involve the collection of photographic or video materials?	No
Will financial or other inducements (other than reasonable expenses and compensation for time) be offered to participants?	Yes

Please explain below why your research project involves the above mentioned criteria (be sure to explain why the sensitive criterion is essential to your project's success). Give a summary of the ethical issues and any action that will be taken to address these. Explain how you will obtain informed consent (and from whom) and how you will inform the participant(s) about the research project (i.e. participant information sheet). A sample consent form and participant information sheet can be found on the Research Ethics website.
see the attached documents

Final Review

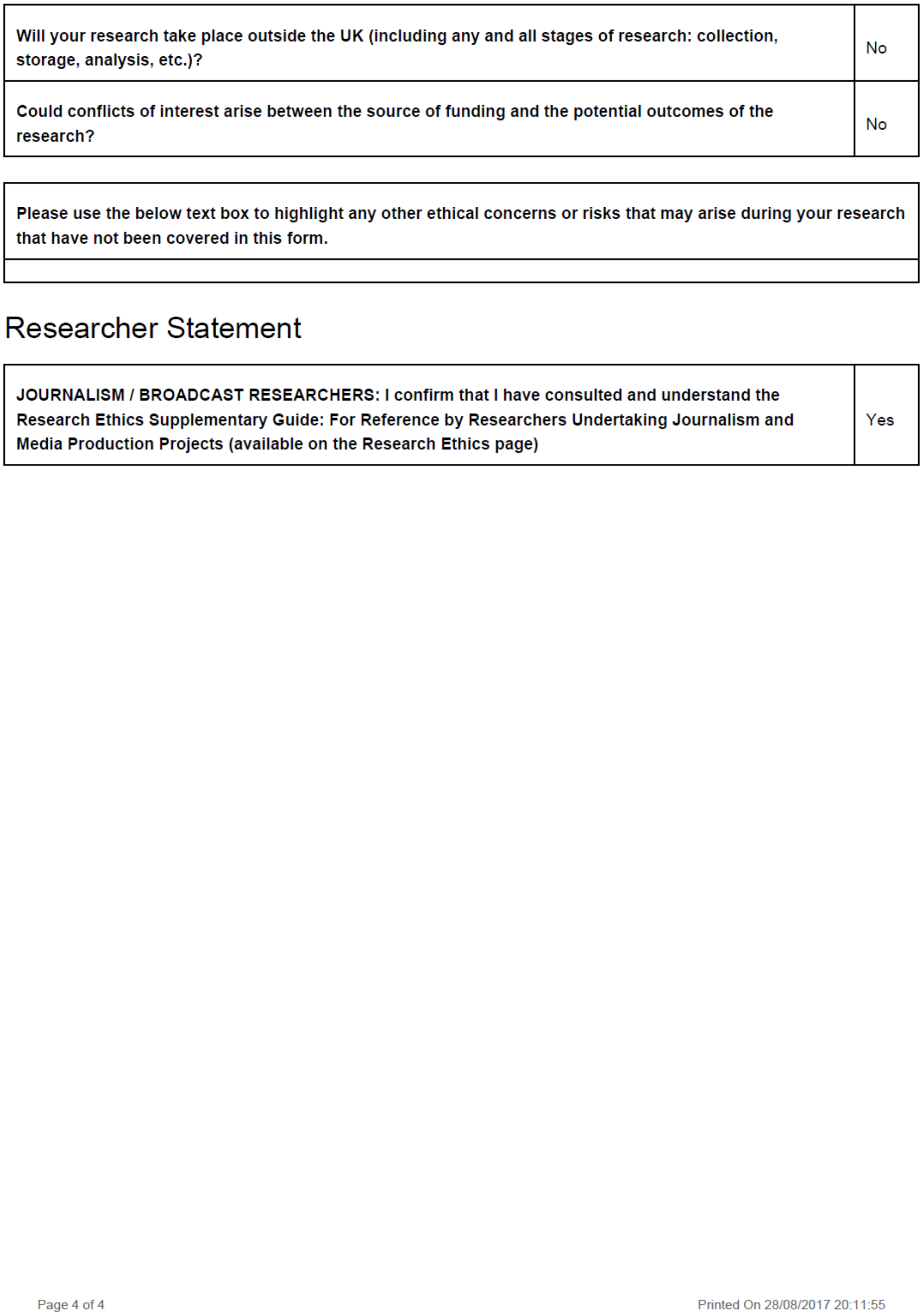
Will you have access to personal data that allows you to identify individuals OR access to confidential corporate or company data (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)?	Yes

Please explain below why your research requires the collection of personal data. Describe how you will anonymize the personal data (if applicable). Describe how you will collect, manage and store the personal data (taking into consideration the Data Protection Act and the 8 Data Protection Principles). Explain how you will obtain informed consent (and from whom) and how you will inform the participant about the research project (i.e. participant information sheet).

Will your research involve experimentation on any of the following: animals, animal tissue, genetically modified organisms?	No

Page 3 of 4

Printed On 28/08/2017 20:11:55



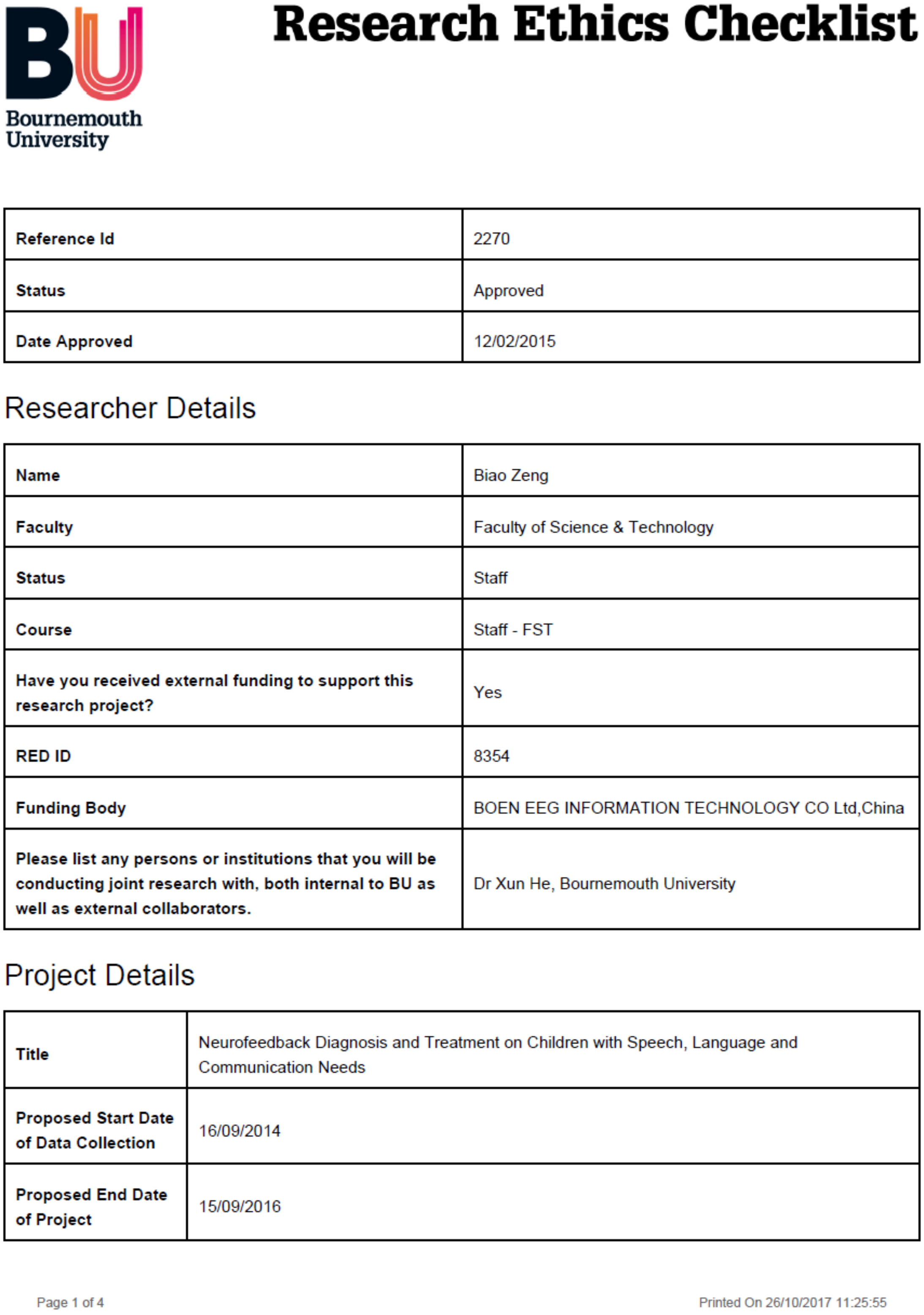
Will your research take place outside the UK (including any and all stages of research: collection, storage, analysis, etc.)?	No
Could conflicts of interest arise between the source of funding and the potential outcomes of the research?	No

Please use the below text box to highlight any other ethical concerns or risks that may arise during your research that have not been covered in this form.

Researcher Statement

JOURNALISM / BROADCAST RESEARCHERS: I confirm that I have consulted and understand the Research Ethics Supplementary Guide: For Reference by Researchers Undertaking Journalism and Media Production Projects (available on the Research Ethics page)	Yes

Ethical Approval 2



Bournemouth University

Research Ethics Checklist

Reference Id	2270
Status	Approved
Date Approved	12/02/2015

Researcher Details

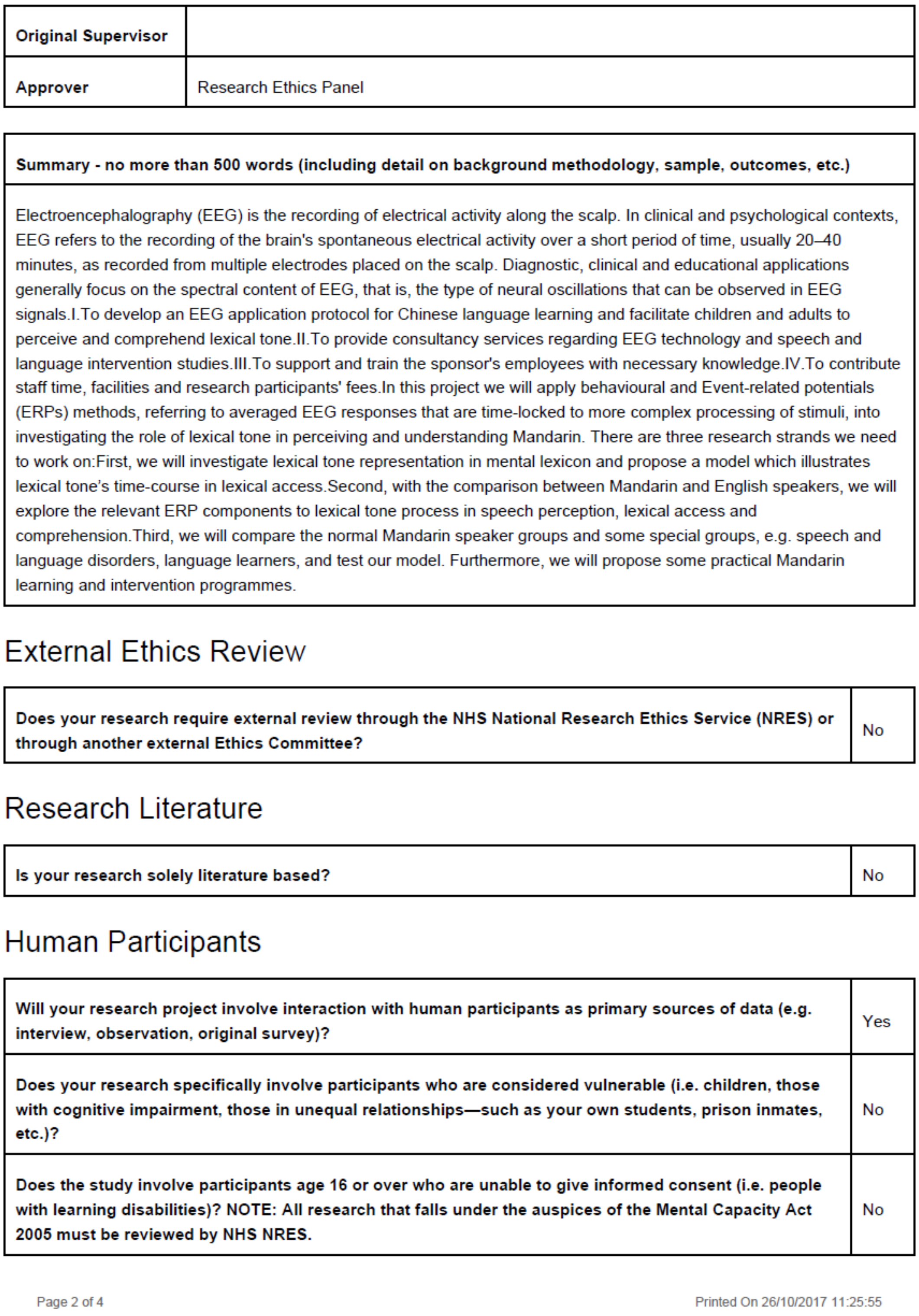
Name	Biao Zeng
Faculty	Faculty of Science & Technology
Status	Staff
Course	Staff - FST
Have you received external funding to support this research project?	Yes
RED ID	8354
Funding Body	BOEN EEG INFORMATION TECHNOLOGY CO Ltd,China
Please list any persons or institutions that you will be conducting joint research with, both internal to BU as well as external collaborators.	Dr Xun He, Bournemouth University

Project Details

Title	Neurofeedback Diagnosis and Treatment on Children with Speech, Language and Communication Needs
Proposed Start Date of Data Collection	16/09/2014
Proposed End Date of Project	15/09/2016

Page 1 of 4

Printed On 26/10/2017 11:25:55



Original Supervisor	
Approver	Research Ethics Panel

Summary - no more than 500 words (including detail on background methodology, sample, outcomes, etc.)
Electroencephalography (EEG) is the recording of electrical activity along the scalp. In clinical and psychological contexts, EEG refers to the recording of the brain's spontaneous electrical activity over a short period of time, usually 20–40 minutes, as recorded from multiple electrodes placed on the scalp. Diagnostic, clinical and educational applications generally focus on the spectral content of EEG, that is, the type of neural oscillations that can be observed in EEG signals.I.To develop an EEG application protocol for Chinese language learning and facilitate children and adults to perceive and comprehend lexical tone.II.To provide consultancy services regarding EEG technology and speech and language intervention studies.III.To support and train the sponsor's employees with necessary knowledge.IV.To contribute staff time, facilities and research participants' fees.In this project we will apply behavioural and Event-related potentials (ERPs) methods, referring to averaged EEG responses that are time-locked to more complex processing of stimuli, into investigating the role of lexical tone in perceiving and understanding Mandarin. There are three research strands we need to work on:First, we will investigate lexical tone representation in mental lexicon and propose a model which illustrates lexical tone's time-course in lexical access.Second, with the comparison between Mandarin and English speakers, we will explore the relevant ERP components to lexical tone process in speech perception, lexical access and comprehension.Third, we will compare the normal Mandarin speaker groups and some special groups, e.g. speech and language disorders, language learners, and test our model. Furthermore, we will propose some practical Mandarin learning and intervention programmes.

External Ethics Review

Does your research require external review through the NHS National Research Ethics Service (NRES) or through another external Ethics Committee?	No

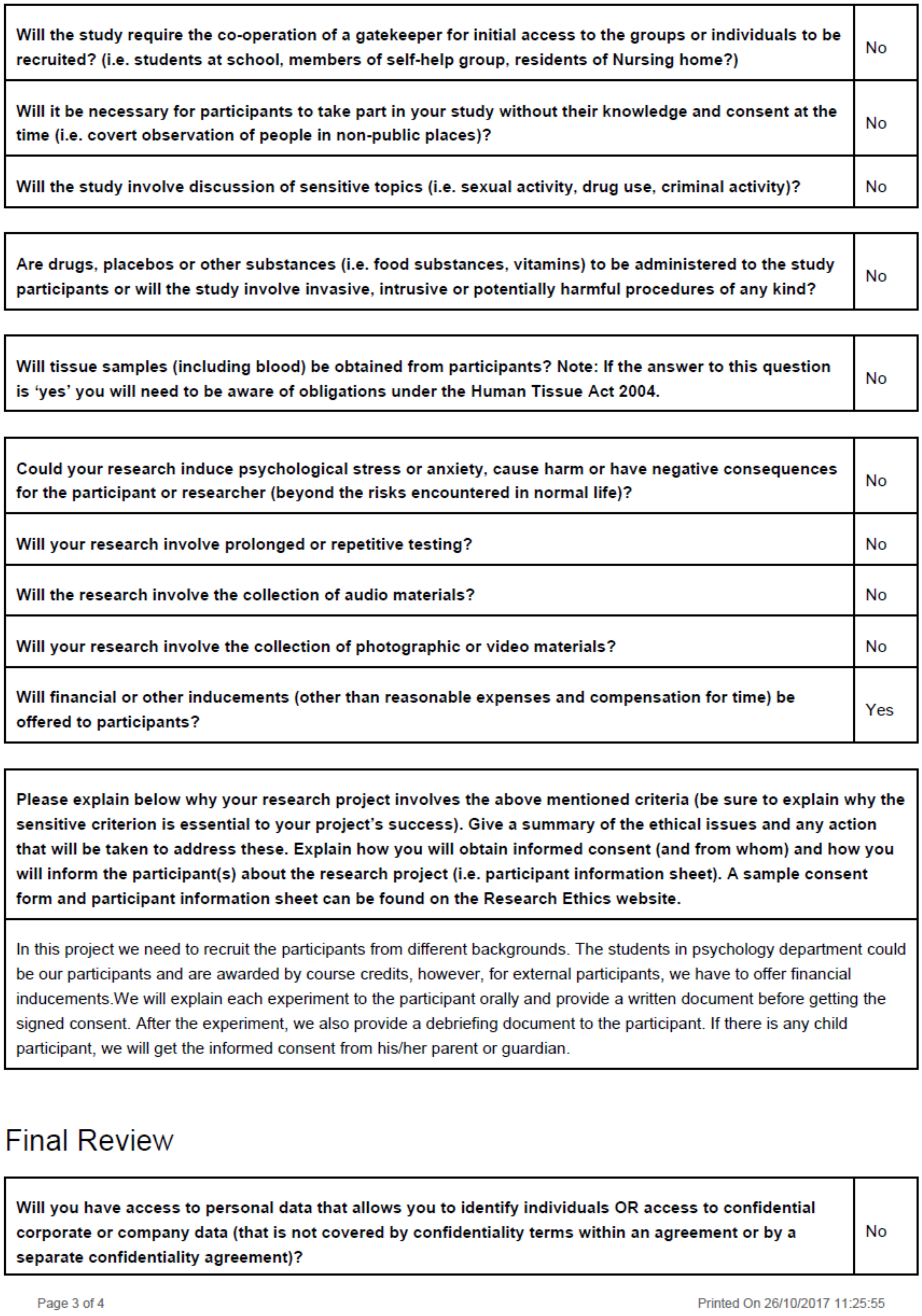
Research Literature

Is your research solely literature based?	No

Human Participants

Will your research project involve interaction with human participants as primary sources of data (e.g. interview, observation, original survey)?	Yes
Does your research specifically involve participants who are considered vulnerable (i.e. children, those with cognitive impairment, those in unequal relationships—such as your own students, prison inmates, etc.)?	No
Does the study involve participants age 16 or over who are unable to give informed consent (i.e. people with learning disabilities)? NOTE: All research that falls under the auspices of the Mental Capacity Act 2005 must be reviewed by NHS NRES.	No

Page 2 of 4 Printed On 26/10/2017 11:25:55



Will the study require the co-operation of a gatekeeper for initial access to the groups or individuals to be recruited? (i.e. students at school, members of self-help group, residents of Nursing home?)	No
Will it be necessary for participants to take part in your study without their knowledge and consent at the time (i.e. covert observation of people in non-public places)?	No
Will the study involve discussion of sensitive topics (i.e. sexual activity, drug use, criminal activity)?	No

Are drugs, placebos or other substances (i.e. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	No

Will tissue samples (including blood) be obtained from participants? Note: If the answer to this question is 'yes' you will need to be aware of obligations under the Human Tissue Act 2004.	No

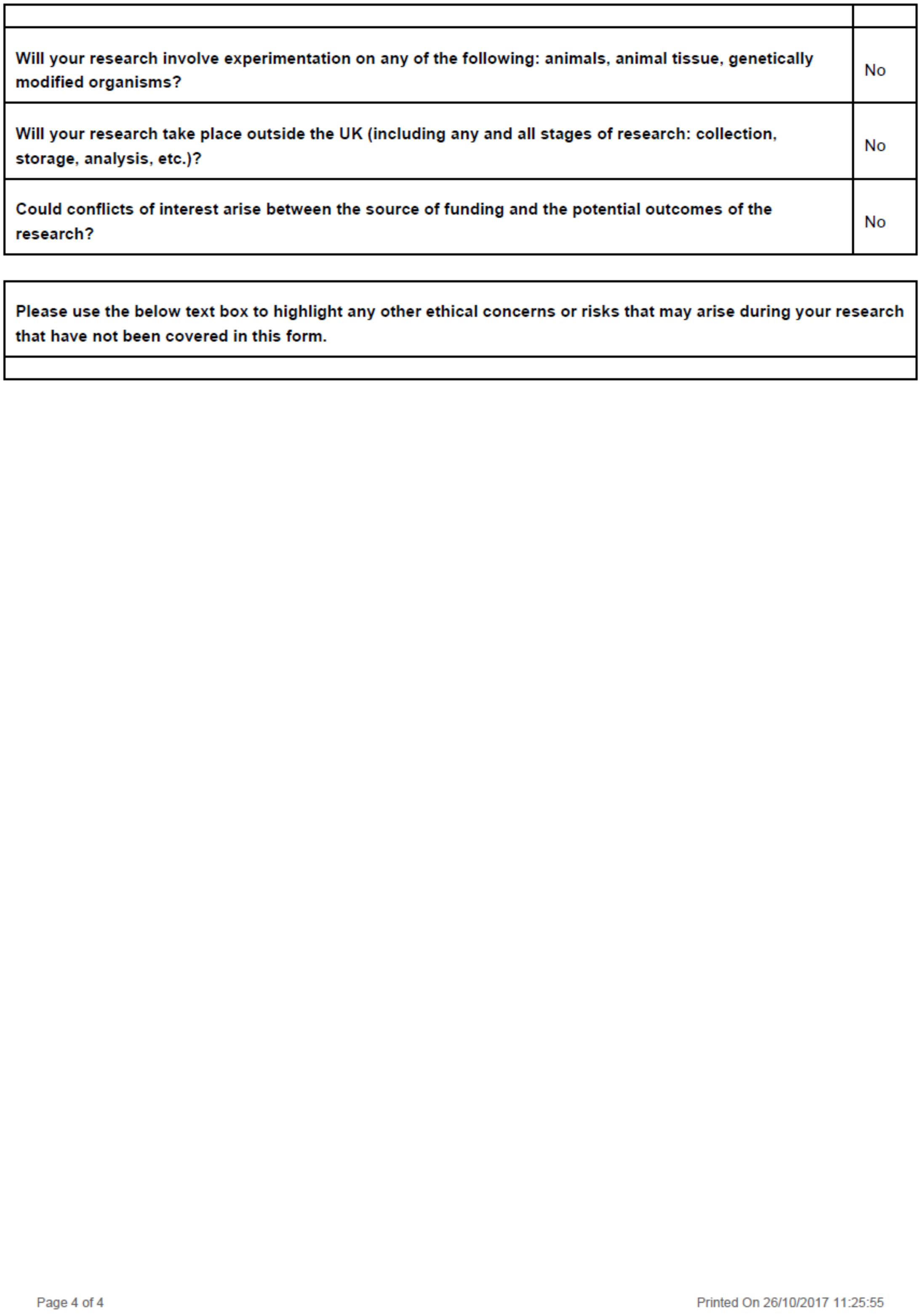
Could your research induce psychological stress or anxiety, cause harm or have negative consequences for the participant or researcher (beyond the risks encountered in normal life)?	No
Will your research involve prolonged or repetitive testing?	No
Will the research involve the collection of audio materials?	No
Will your research involve the collection of photographic or video materials?	No
Will financial or other inducements (other than reasonable expenses and compensation for time) be offered to participants?	Yes

Please explain below why your research project involves the above mentioned criteria (be sure to explain why the sensitive criterion is essential to your project's success). Give a summary of the ethical issues and any action that will be taken to address these. Explain how you will obtain informed consent (and from whom) and how you will inform the participant(s) about the research project (i.e. participant information sheet). A sample consent form and participant information sheet can be found on the Research Ethics website.

In this project we need to recruit the participants from different backgrounds. The students in psychology department could be our participants and are awarded by course credits, however, for external participants, we have to offer financial inducements.We will explain each experiment to the participant orally and provide a written document before getting the signed consent. After the experiment, we also provide a debriefing document to the participant. If there is any child participant, we will get the informed consent from his/her parent or guardian.

Final Review

Will you have access to personal data that allows you to identify individuals OR access to confidential corporate or company data (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)?	No



Will your research involve experimentation on any of the following: animals, animal tissue, genetically modified organisms?	No
Will your research take place outside the UK (including any and all stages of research: collection, storage, analysis, etc.)?	No
Could conflicts of interest arise between the source of funding and the potential outcomes of the research?	No

Please use the below text box to highlight any other ethical concerns or risks that may arise during your research that have not been covered in this form.