Visual benefit in lexical tone perception in Mandarin: An event-related potential study

Rui Wang[1], *Xun He*[2*], *Biao Zeng*[3*]

[1]School of Foreign Languages, Guangdong Pharmaceutical University, China
[2]Department of Psychology, Faculty of Science and Technology, Bournemouth University, UK
[3] School of Psychology and Therapeutic Studies, Faculty of Life Science and Education, University of South Wales, UK

rui.wang@gdpu.edu.cn, xhe@bournemouth.ac.uk, biao.zeng@southwales.ac.uk
* Corresponding authors

List of Abbreviations

ANOVA: analysis of variance

AO: auditory-only

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

AV: audiovisual

d’: sensitivity index

EEG: electroencephalogram

ERP: event-related potential

SOA: stimulus onset asynchrony

VO: visual-only

Abstract

Congruent visual information enhances auditory speech perception. This visual benefit has been widely observed in perception of segments and linked to reduced amplitudes and latencies of auditory N1 and P2 event-related potential (ERP) components when visual information was present. However, it remains unclear whether lexical tone perception in Mandarin also shows this visual benefit. This question is theoretically important given the low visual saliency of lexical tones. The current study compared the N1/P2 reduction in Mandarin lexical tones and consonants perception with a discrimination task. Result showed amplitude reductions in N1/P2 and a latency reduction in N1 for audiovisual lexical tone perception. These findings suggest that lexical tone perception was also helped by visual information as found in consonants. Furthermore, this visual benefit in N1 for lexical tone perception was delayed relative to consonants.

Keywords: Mandarin lexical tones; Audiovisual speech perception; Event-related potential (ERP); Integration

INTRODUCTION

In natural speech, speakers convey and exchange audiovisual information rather than auditory-only information. Visualising an interlocutor's face, particularly the mouth area, typically improves the perception of speech (Summerfield, 1987). Audiovisual speech perception has been extensively studied in segmental consonants and vowels, but is lacking in prosodic information. The visual cues of consonants and vowels represent the articulatory gestures or visemic features (e.g. bilabial /b/, fricative /v/, mouth roundness /o/ or flatness /i/ in speech), whereas the visual cues of prosodic information (intonation and tone) are much more implicit. Intonation is a form of prosodic information which refers to the rise and fall of pitch over entire phrases and sentences. It conveys emotional, pragmatic, and social information, e.g. questioning, doubting and satire. Several studies have reported the role of upper facial cues in this form of prosodic information. The upper facial cues can facilitate the listener's ability to identify intonation through head movements (Cvejic et al., 2010) and eyebrow movements (Kim & Davis, 2014). Tone information, on the other hand, expresses lexical meaning of a syllable or a word. The tonal perception can also be improved with watching a speaker's face (Smith & Burnham, 2012).

Lexical tone, as another form of prosodic information, differs to consonants and vowels. It widely exists in many East Asian languages, such as Mandarin Chinese, Thai, and Vietnamese. A lexical tone is the fundamental frequency or pitch variation over a syllable or mora able to distinguish the lexical or grammatical meaning of a word. For example, in Mandarin, the syllable /ma/, when produced with a high-level tone, means *mother*, in a rising tone means *hemp*, in a dipping tone means *horse*, and means *to scold* in a falling tone. In a tonal language, a typical anatomy of monosyllables includes segmental consonant and vowel as well as lexical tone. From a phonological perspective, a segment, usually a syllable nucleus, is defined as tone-bearing unit (Yip, 2002), which can be associated with a tone. In Mandarin, a lexical tone is born by syllable nucleus and preceded by a consonant. Autosegmental theory (Goldsmith, 1979) has received widespread attention for its claim that tones are represented in a separate tier from segments, even though both are co-registered at the phonetic level.

Because lexical tones are produced by the vibration of vocal cords (Yip, 2002), there are few explicit visual articulatory cues from preceding mouth movements. However, visual features being less visible do not necessarily suggest that visual cues of lexical tones do not exist. Some studies have discovered that adding visual information improved the identification of lexical tones in adverse auditory conditions (e.g. Burnham et al., 2001). Additionally, lexical tone perception could benefit from various visual cues: rigid motion (Burnham et al., 2006), head movement (Chen & Massaro, 2008), and visual timing (Xie et al., 2018). The current study would look into the neural underpinnings of the visual benefit in lexical tone perception.

At the neural level, the visual benefit effect of segments has been observed in auditory event-related potentials (ERPs). Auditory ERPs are electrical brain activities evoked by auditory stimuli. The N1 and P2 components of auditory ERPs are biomarkers of early auditory perception thus often used to illustrate the visual benefit effect in audiovisual speech perception. N1 is a negative ERP component peaking at about 100 ms in the fronto-central area on the scalp, followed by P2, which is a positive ERP component peaking at about 180 ms. In the presence of visual benefit during audiovisual speech perception, auditory N1 and P2 were found to be smaller in amplitude (Besle et al., 2004) and shorter in latency (van Wassenhove et al., 2005) than in auditory-only speech perception.

Besle et al. (2004) found that the amplitude of the auditory N1 component evoked by speech syllables was reduced with the presence of visual articulatory gestures. They proposed that this N1 reduction was associated with phonetic pre-activation from preceding lip movement in the auditory cortex through the poly-modal area superior temporal sulcus. Another study (van Wassenhove et al., 2005) reported shortened N1 and P2 latencies as a result of visual benefit, and noted that the latency reduction was greatest with the most visually salient /p/ (bilabial) and weakest with the least visually salient /k/ (velar), suggesting the importance of visemic features salience. Their findings agree with the notion that phonetic predictiveness of visual speech leads to an ERP latency reduction. Another hypothesis suggests that the reduction effect in the auditory N1 component was not due to visual signals predicting the content of the auditory signal, but instead due to an alerting effect of visual signal on the forthcoming auditory signal, because the N1/P2 reduction effect in audiovisual stimuli was found in both speech and non-speech stimuli (e.g. sawing wood) but only when visual movements preceded auditory stimuli (Stekelenburg & Vroomen, 2007). Despite that whether the visual speech predicts the content (phonetic identity) or alerts the upcoming auditory signal remains in dispute, it can be argued that the neural reduction effect was related to the degree of predictiveness of visual speech.

It has been suggested that preceding lip movements (articulation) provide phonetic information and thus alleviate the processing load in the auditory modality, thereby accelerating and/or reducing auditory responses (e.g. Besle et al., 2009). This may be true for consonants and vowels, whose lip movements usually start a few hundred milliseconds ahead of auditory signals; therefore, visual speech might inform on following auditory speech processing (Besle et al., 2004). However, the audiovisual integration in lexical tone perception is little understood. For lexical tones, the preceding lip movements are unlikely, if possible at all, to predict the upcoming auditory tone signal. Therefore, it is theoretically important to investigate whether audiovisual processing of lexical tones would also cause similar reduction effects in the N1/P2 ERP components (relative to auditory processing only).

The reduction effects in amplitudes and latencies may reflect different mechanisms. For instance, in a study by Knowland et al. (2014), the N1/P2 reduction effect was studied with monosyllabic words in congruent (the visual and auditory information matched) and incongruent conditions (e.g. an auditory syllable *lay* paired with a visual syllable *row*). The reduction in amplitude was found to be independent of congruency, whereas the reduction in latency was sensitive to the congruency of audiovisual syllables. Despite the possibility of reflecting different mechanisms, both the amplitude and latency reductions in the auditory N1/P2 components have been taken as the neural indices of visual benefit in audiovisual speech perception (Alsius et al., 2014; van Wassenhove et al., 2005). Pilling (2009) also confirmed that this N1/P2 reduction effect represents the audiovisual integration process rather than other cognitive factors such as attention shifting to visual modality or top-down inhibition in the audiovisual modality.

Lateralisation is closely related to speech perception thus an important aspect of the current research. Speech processing takes place dominantly in the left hemisphere (Hickok & Poeppel, 2007). 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. There is evidence that the processing of lexical tones is lateralised to the left hemisphere 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, 2006; Shuai & Gong, 2014). For audiovisual integration, mainstream theories suggest dominance of the left hemisphere (Calvert, 2001; Campbell, 2008). This was supported by ERP evidence (Reale et al., 2007). However, some neuroimaging studies did not agree on this question (Okada et al., 2013). It is not clear whether the dominance of audiovisual integration shows the same lateralisation found in general speech perception. The perception lateralisation of audiovisual lexical tones was also studied in the current study.

The current study investigated whether lexical tone perception shows visual benefit indexed by the amplitude/latency reduction of auditory N1 and P2 ERP components. Because consonants have more salient visual features than those of lexical tones, it was hypothesised that the N1/P2 reduction effect would be weaker and later in lexical tones than in consonants. To quantify the visual benefit, an additivity model (Barth et al., 1995) was employed for a comparison between the Auditory-Only (AO) and Audiovisual (AV) modalities. This model asserts that the neural activity of AV processing equals the sum of AO and Visual-only (VO) processing activities (i.e. AV = AO + VO) if auditory and visual information are processed separately. The violation of the equation indicates audiovisual integration. The critical comparisons were between the AO condition and the auditory-only responses produced by the AV stimuli (AO'). The AO' activities were derived by subtracting the VO activities from the AV activities (i.e. AO' = AV – VO) (the details are shown in Fig. S1 in Supplementary Materials). A reduction effect was defined as weaker activities in the AO' than AO conditions.

METHOD

Participant recruitment

Native Mandarin speakers aged between 18 and 45 years were recruited from the Bournemouth University student community as participants in the current study. They all reported normal or corrected-to-normal visual acuity and no hearing impairment. To clearly show hemispheric differences, only right-handed participants were allowed. The experimental protocol was approved by Research Ethics Panel of Bournemouth University in accordance with the Declaration of Helsinki. Informed consent was obtained from each participant before the experiment took place.

Materials and procedure

The stimuli were six Mandarin monosyllables *bai, dai, tai, bao, dao, and tao* (International Phonetic Alphabet [pai], [tai], [t′ai], [pɑu], [tɑu] and [t′ɑu]) with four lexical tones (6 × 4 = 24 syllables), presented in AO, AV and VO conditions. These syllables were produced by two native male Mandarin speakers. 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 soundtracks of the videos were edited in Audacity (Audacity Team, 2020) and Adobe Audition CC (Adobe Systems, California). All auditory tracks were digitised at 48,000 Hz, with a 32-bit amplitude resolution, and were root mean square normalised to -12 dB. The durations and auditory onsets of AV stimuli were kept constant. AO and VO stimuli were derived from the AV clips and kept identical in duration as well. The physical properties of all the syllable stimuli used in the current study are summarised in Table S1 in Supplementary Materials.

Visual and auditory onsets were measured for consonants and lexical tones. The difference between the visual and auditory onsets indicated how much visual cues preceded acoustic information. Fig. 1 illustrates that the stimulus onset always preceded the audio onset by 234 ms. The visual onset of consonant took place shortly before the audio onset. However, as there was no technique to recognise or extract visual cues of lexical tones, the visual onset of a lexical tone was assumed to be identical as that of a consonant. Because Mandarin lexical tones were born by rimes or vowels, a lexical tone’s audio onset was defined as vowel onset. Overall, the average visual-audio gap of consonant onsets was 149 ms, and that of lexical tones was 197 ms.

The experiment was conducted in a sound-attenuated and dimly lit room. The participants sat in front of a 17-inch CRT monitor at a viewing distance of 70 cm. Sound was played via two Genelec 8030A loudspeakers (Genelec Oy, Iisalmi) placed by the sides of the monitor. The loudness of the presented sound was approximately 65 dB sound pressure level. The experimental stimuli were presented using E-Prime 2.0 (Psychology Software Tools, Sharpsburg).

Fig. 1 illustrates the trial sequences. A same-different discrimination paradigm (go/no-go task) was employed in both the lexical tone and consonant experiments using the same set of syllables. A fixation cross (0.8×0.8˚) remained on the centre of the screen throughout each block except when video clips were played. Each trial had two consecutive syllables. The first syllable (Stim1) was always presented in AO and had an average duration of 726 ms (see Table S1 in Supplementary Materials for detailed audio duration information). The second syllable (Stim2) was randomly presented in one of the modalities (AO, VO, or AV; visual image was displayed 13.6×19.3˚ at the screen centre) and had a duration of 1,368 ms (measured from stimulus onset to stimulus offset in Fig. 1). In each trial, the stimulus onset asynchrony (SOA; the time difference between the onsets of Stim1 and Stim2) was randomised between 1,250 and 1,650 ms. The two syllables differed in lexical tone in 20% of trials in the lexical tone experiment, and differed in consonant (20% of trials) in the consonant experiment. The participants were 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, and to press the spacebar as quickly and accurately as possible (within 3,000 ms from Stim2 onset) when different tones or consonants were detected within a trial. The intervals between trials were randomised 2,700-3,300 ms. In total, there were 540 trials in six blocks.

[Insert Fig. 1 about here]

EEG recording and pre-processing

Electroencephalogram (EEG) was recorded at a sampling rate of 500 Hz and with a physical bandpass filter (0.1-250 Hz) using a Brain-Amp DC system (Brain Products GmbH, Gilching) Thirty-two 10-20 system (Jasper, 1958) recording sites (Fp1/2, AF3/4, Fz/3/4/7/8, FC1/2/5/6, Cz/3/4, T7/8, CP1/2, TP7/8, Pz/3/4/7/8, PO7/8, O1/2) and the right mastoid were used, all physically referenced to the left mastoid and re-referenced off-line to an averaged-mastoids reference. The impedance level was kept below 20 kΩ.

Raw data were processed offline with EEGLAB 14.0.0 (Delorma & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014). Data were bandpass filtered (0.1–30 Hz, 48 dB/oct roll-off) and segmented into 1,200 ms-epochs (200 ms before to 1,000 ms after stimulus onsets), and baseline-corrected (200 ms period pre-stimulus). Segments with activities stronger than ±100 μV were rejected as artefacts before the averaging process.

Statistical analysis

Statistical analysis was performed using JASP 0.12 (JASP Team, 2020) with a preference for parametric analysis whenever possible. Repeated-measures analyses of variance (ANOVAs) with fixed-effect models and Bayesian ANOVAs were employed with all statistical assumptions tested, including data normality (Shapirao-Wilk test) and homogeneity of variance (Levene's test). In case of assumption violation, box plot was used for outlier detection. If the assumptions were met after outlier removal, repeated-measures ANOVA and Bayesian ANOVA were applied, with Greenhouse-Geisser correction employed for sphericity assumption violation. All post hoc comparisons were corrected with the Bonferroni procedure following the presence of significant main effects in the ANOVAs. If no outlier could be identified or outlier removal could not help following violated assumptions, non-parametric tests (e.g., Mann-Whitney) were applied as well.

RESULTS

Participant characteristics

Twenty volunteers (13 females, 7 males) participated in the lexical tone experiment, with an average age of 25.8 ± 4.4 years. Another group of nineteen (11 females, 8 males; aged 26.6 ± 5.7 years) volunteers participated in the consonant experiment. No participant in one experiment took part in the other experiment. All participants were Chinese, and were native Mandarin speakers and right-handed per the recruitment requirement. They were studying at the undergraduate or graduate level at Bournemouth University at the time of the study.

Sensitivity

Perception performance was quantified as sensitivity (*d'*) values, which were 4.34 ± 0.40 (AV), 4.25 ± 0.35 (AO), 1.93 ± 0.39 (VO) for consonants, and 4.19 ± 0.48 (AV), 3.94 ± 0.64 (AO), 0.28 ± 0.38 (VO) for lexical tones. Data from both experiments were analysed with an ANOVA and a Bayesian ANOVA, using a between-subject factor Language Unit (lexical tone, consonant) and a within-subject factor Modality (AO, AV, VO). Greenhouse-Geisser correction was applied when the sphericity assumption was violated. The post hoc comparisons were corrected with the Bonferroni procedure.

There were significant main effects of Modality [the consonant condition outperformed the lexical tone condition, $F_{2,\,74} = 976.83$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.964$, $BF_{10} = 5.7 \times 10^{13}$] and Language Unit ($F_{1,\,37} = 39.97$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.519$, $BF_{10} = 5.7 \times 10^{13}$), and a significant interaction ($F_{1.69,\,62.58} = 52.49$, $p < .001$, $\eta^2_{\mathrm{p}} = 0.587$, $BF_{10} = 7.4 \times 10^{12}$). Further analyses showed the lowest *d*' values in VO (*p*s < .001) and a visual benefit in lexical tones (AV > AO, $p = .026$). This visual benefit, however, was absent in consonants ($p = 1.0$), probably due to ceiling performance. It is also worth noting that, consistent with consonants being more visually salient than lexical tones, consonants in VO were reliably perceived, whereas lexical tones in VO showed much worse ($p < .001$) and near-zero sensitivity.

Given that normality (Shapiro-Wilk *p*s < .034 in the two AV conditions) and homogeneity ($p = .002$ for the AO conditions) assumptions were violated, and no outliers were detected, non-parametric tests were also employed. Wilcoxon tests confirmed the above findings, namely sensitivity differences across all three modalities in lexical tones (AV vs. VO $p = .028$, other *p*s < .001) but no visual benefit in consonants (AV vs. VO $p = .22$, other *p*s < .001). A Mann-Whitney test also confirmed the tone-consonant difference in VO ($p < .001$).

ERPs

AO' ERP waveforms were calculated as the differences between AV and AO waveforms (AV – VO) and were contrasted with AO waveforms. After the subtraction, the zero time points were shifted 234 ms to the right of the time axis and re-aligned with the auditory onsets (see Fig. 1).

The 200-ms pre-auditory-onset period was used as the new baseline. Amplitudes and latencies of the auditory N1 and P2 components were examined with ANOVAs and Bayesian ANOVAs. There was a between-subject factor Language Unit (lexical tone vs. consonant) and a within-subject factor Modality (AO vs. AO'), with an additional factor Lateralisation (left vs. right hemisphere) for N1.

Fig. 2 illustrates the results of N1. The mean amplitudes of N1 were measured 168–188 ms after auditory onset at electrodes FC1/2, FC5/6, and C3/4. These were determined with collapsed localisers (Luck & Gaspelin, 2017) of the visual benefit effects across the lexical tone and consonant conditions. Box plots detected two outlier participants, one in each language-unit group. After removal of these outliers, no normality violation was detected by Shapiro-Wilk tests (*p*s > .053). The homogeneity assumption was also met (Levene's tests *p*s > .59). In the ANOVAs, main effects were observed for Modality [AO (-3.03 ± 0.26 μV) higher than AO' (-1.26 ± 0.25 μV), $F_{1, 35} = 46.94$, $p < .001$, $\eta^2_p = 0.573$, $BF_{10} = 1.0 \times 10^{15}$] and Lateralisation [left hemisphere (-2.36 ± 0.24 μV) stronger than right hemisphere (-1.92 ± 0.22 μV), $F_{1, 35} = 19.76$, $p < .001$, $\eta^2_p = 0.361$, $BF_{10}= 3.628$]. The main effect of Language Unit was also significant [tones (-2.65 ± 0.31 μV) stronger than consonants (-1.63 ± 0.32 μV), $F_{1, 35} = 5.24$, $p = 0.028$, $\eta^2_p = 0.130$] but only showing anecdotal evidence ($BF_{10} = 2.356$). No interaction approached significance (*p*s > .30, BF_{10} between 0.265 and 0.381).

The N1 latencies were determined with peak detection within a time window of 132-204 ms and analysed with three-way ANOVAs. The normality assumption was met (*p*s > .19). Only a strong Modality main effect was found [AO' (152 ± 2 ms) faster than AO (160 ± 2 ms), ($F_{1, 37} = 21.50$, $p < .001$, $\eta^2_p = 0.368$, $BF_{10} = 2.0 \times 10^8$]. All other effects were not significant (*p*s > .34, BF_{10} between 0.174 and 0.453). The homogeneity assumption was not met (*p*s < .050 except for one cell), however further Mann-Whitney tests confirmed the absence of any between-subject difference (*p*s > .46).

[Insert Fig. 2 about here]

Fig. 3 demonstrates the P2 results. The mean amplitudes of P2 were quantified between 240 and 260 ms over electrodes FC1/2 and Cz, and were assessed with two-way ANOVAs. The results showed a significant Modality effect [AO (3.05 ± 0.30 μV) stronger than AO' (1.74 ± 0.39 μV), $F_{1, 37} = 15.33$, $p < .001$, $\eta^2_p = 0.293$, $BF_{10} = 84.489$]. No other effects reached significance [ps > .26, $BF_{10} = 0.541$ (Language Unit) or 0.305 (interaction)]. P2 latencies were detected within a window of 210-290 ms. Only a main effect of Modality was found [AO' (240 ± 4 ms) earlier than AO (247 ± 3 ms), $F_{1, 37} = 4.29$, $p = 0.045$, $\eta^2_p = 0.104$] but only anecdotal (BF_{10}= 1.460) (other ps > .61, $BF_{10} = 0.407$ or 0.318).

[Insert Fig. 3 about here]

Time-course and lateralisation of N1 reduction effects

In the analyses reported above, the non-significant interactions involving Lateralisation seem to suggest the absence of lateralisation in the N1 reduction effect (an index of visual benefit), not to mention any differential effects between lexical tones and consonants. However, in Fig. 2, the AO – AO' difference waveforms seemed to be stronger in the right hemisphere for lexical tones and stronger in the left hemisphere for consonants. To further assess lateralisation and the time course of audiovisual integration, further exploratory analyses for N1 were carried out.

The N1 amplitude reduction effects (representing the visual benefits) were measured in the AO – AO' difference waveforms over representative electrodes FC1/2/5/6, C3/4, and T7/8. The peak latencies of the reduction effects were subjected to two-way ANOVAs with factors Language Unit (lexical tone, consonant) and Lateralisation (left vs. right). Normality was largely followed (Shapiro-Wilk tests ps > .78 except for one cell, $p = .049$), with no outlier detected. Homogeneity was also confirmed across groups (Levene's tests ps > .67). The ANOVAs only found a strong significant Language Unit effect [consonants (174 ± 2 ms) 14-ms earlier than lexical tones (188 ± 2 ms), $F_{1, 37} = 20.91$, $p < .001$, $\eta^2_p = 0.361$, $BF_{10} = 398.68$]. Other effects were not significant (ps > .29, $BF_{10} = 0.349$ or 0.465).

The above results were then consulted to determine the mean-amplitude measurement windows for the reduction effect amplitudes (162-182 ms and 180-200 ms for consonants and lexical tones respectively). The normality and homogeneity assumptions were met (all ps > .36). Two-way ANOVAs did not find any significance (Fs < 2.76, ps > .10, $BF_{10} < 1.012$). A t-max permutation test (Groppe et al., 2011) was also applied to examine the hemispheric differences at a family-wise alpha level of 0.05. For each language unit, all time points between 160 and 200 ms at eighteen electrodes (nine per hemisphere) were included. No significant effect was found. While revealing an earlier visual benefit in consonants than lexical tones, these results confidently and consistently pointed to a lack of visual benefit lateralisation in the ERPs.

DISCUSSION

In the current study, despite having a low level of saliency, the visual information still helped the lexical tone perception in behaviour. This visual benefit was also observed as N1 and P2 amplitude reductions when visual inputs were available in lexical tone perception, as well as in consonants. A latency reduction effect was also revealed in N1 for both language units. Additionally, the visual benefit effects in N1 occurred later in lexical tones than in consonants. The results suggest that visual information can significantly improve the processing of auditory lexical tones in the same way the visual information benefits auditory perception of consonants. Auditory N1 and P2 are known to reflect activities involving sensory processing that are sensitive to physical variations in stimuli (Näätänen & Winkler, 1999). The reduction in the lexical tone task within this time range (particularly in the N1 time interval) supports the audiovisual integration of lexical tones that began in sensory processing (pre-linguistic processing), similar to the audiovisual integration of consonants in this time range. The finding that lexical tone perception, which was more difficult than consonant perception in VO as shown in sensitivity results, evoked stronger N1 is also consistent with the notion that N1 amplitudes reflect the processing load of auditory stimuli (Besle et al., 2004).

Compared with 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 the VO condition, lip-reading lexical tones was much more difficult than lip-reading consonants, as mouth movement-related visual cues provided very 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 did not only depend on how much phonetic information visual input could convey. Instead, it could be due to visual timing information (Kim & Davis, 2014) from visual inputs. Duration is an important cue (both visual and auditory) of lexical tones and can be used to discriminate tones (e.g. the dipping tone has the longest duration). The contribution of duration, however, is yet to be confirmed given that the N1 and P2 latencies (around 160 ms and 240 ms respectively in the current study) were much shorter than durations of lexical tones (in average 677 ms in the current study) thus it is less likely that the duration information dominated the visual benefit in ERPs. On the other hand, the presence of N1/P2 reduction in the current findings is consistent with the alerting effect of visual information on auditory processing in natural speech perception (Pilling, 2009; Stekelenburg & Vroomen, 2007). This alerting effect was shared by consonants and lexical tones in the current study thus could be the reason why similar reduction effects were found for these language units.

It is worth noting that there was no difference in the magnitudes of N1/P2 reductions between consonants and lexical tones. This is striking given that consonants also benefitted from apparent visual information, for instance, articulatory gesture. This benefit was unlikely for lexical tones. In the current study, the only difference found between language units was that consonants received visual benefit about 14 ms earlier than lexical tones. This time-course difference in audiovisual integration could be explained in two ways. First, the audiovisual integration is determined by the phonetic saliency or predictiveness of visual information for the phonetic content of auditory speech (van Wassenhove et al., 2005). With the visual lexical tone and consonant stimuli being physically identical in the current experiments, the differential reduction effect could only be a result of differential engagement of audiovisual integration due to visual saliency. One can anticipate a consonant from a distinctive place of articulation (e.g. bilabial) in preceding mouth movements hundreds of milliseconds prior to the auditory signal. However, leading lip movements are less likely to predict lexical tones. The visual cues of lexical tones may provide other cues such as duration (Smith & Burnham, 2012), head movement (Burnham et al., 2006) and laryngeal movement (Chen & Massaro, 2008), but not place of articulation. In contrast, visual consonants have stronger predictability for ensuing auditory consonants. The time difference in the reduction effects between the two speech units was very likely due to the difference of visual saliency between lexical tones and consonants.

Alternatively, lexical tones were born by vowels in Mandarin monosyllables and preceded by consonants in time. The visual onset of a lexical tone was assumed to be identical to that of a consonant. However, the audio onset of a lexical tone, which was defined as the onset of the tone-bearing vowel, differed from that of the consonant. As such, there was a large time difference between the visual-audio gaps of consonants (149 ms) and those of lexical tones (197 ms). The relatively long visual-audio gap in lexical tones could have delayed the integration process. This suggests that the reduction effect found in lexical tones might not solely depend on the amount of visual information, but also the timing of visual inputs (Kim & Davis, 2014). Overall, the time-course difference in audiovisual integration between consonants and lexical tones may be a result of different causes of the reduction effects, namely a timing-dominant cause for lexical tones (Xie et al., 2018) and a visemic-timing combined cause for consonants.

The current data also found a left-hemisphere dominance for auditory perception of both consonants and lexical tones. This is consistent with the suggestion that left hemisphere would dominate if lexical tones were perceived as linguistic categories (Jongman et al., 2006). However, the visual benefit in the current data, as shown by the Modality (AO vs. AO') × Lateralisation interaction in N1 data and the lateralisation analysis, did not show any lateralisation. This differs from past ERP studies (e.g. Reale et al., 2007) but agrees with neuroimaging research (e.g. Okada et al., 2013). More light should be shed by further research on this.

In sum, the current findings for the first time demonstrated visual benefit of lexical tone perception in brain activities (reduced and accelerated N1 and P2 activities) and highlighted the importance of visual inputs for auditory speech perception even with a low visual saliency.

ACKNOWLEDGEMENTS

The study was funded by Shaoxing 330 Scheme for Overseas Chinese Research Elites (China) and Bournemouth University PhD studentship. We would also like to thank Ms. Ziyi Wang for assistance in some data analysis.

GLOSSARY

Lexical tone

A lexical tone is the fundamental frequency or the pitch variation over a syllable that can distinguishes lexical or grammatical meaning of a word.

Mandarin

Mandarin refers to the official language of China. Mandarin contains four tone categories based on different types of pitch variation: tone 1 is high-level tone; tone 2 is high-rising tone; tone 3 is a low-falling-rising tone; and tone 4 is a high-falling tone.

Prosodic information

Prosodic information includes fundamental frequency, duration and intensity of intonations, tones and stresses.

Visemic feature

Visemic feature means the visual features such us place of articulation that can distinguish speech sounds.

REFERENCES

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Frontiers in Psychology, 5*, 727. https://doi.org/10.3389/fpsyg.2014.00727

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 Research, 678*(1-2), 177–190. https://doi.org/10.1016/0006-8993(95)00182-p

Besle, J., Bertrand, O., & Giard, M. H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research*, *258*(1-2), 143–151. https://doi.org/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. *The European Journal of Neuroscience*, *20*(8), 2225–2234. https://doi.org/10.1111/j.1460-9568.2004.03670.x

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. *Proceedings of the International Conference on Audio-Visual Speech Processing (AVSP) 2001, Denmark.* 155-160.

Burnham, D.K., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., 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. *Proceedings of the 7th International Seminar on Speech Production. Ubatuba, Brazil.*185–192.

Calvert G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex, 11*(12), 1110–1123. https://doi.org/10.1093/cercor/11.12.1110

Campbell R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America, 123*(4), 2356–2366. https://doi.org/10.1121/1.2839004

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. https://doi.org/10.1016/j.specom.2010.02.006.

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

Goldsmith, J. (1979). *Autosegmental Phonology*. New York: Garland.

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. https://doi: 10.1111/j.1469-8986.2011.01273.x

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

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

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 psycholinguistics* (pp. 209-217). Cambridge: Cambridge University Press.

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication, 57*, 317-330. https://doi.org/10.1016/j.specom.2013.06.003

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

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

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Näätänen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, *125*(6), 826–859. https://doi.org/10.1037/0033-2909.125.6.826

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PloS ONE*, *8*(6), e68959. https://doi.org/10.1371/journal.pone.0068959

Pilling M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1073–1081. https://doi.org/10.1044/1092-4388(2009/07-0276)

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., Howard, M. A., & Brugge, J. F. (2007). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience*, *145*(1), 162–184. https://doi.org/10.1016/j.neuroscience.2006.11.036

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

Smith, D., & Burnham, D. (2012). Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: implications for cochlear implants. *The Journal of the Acoustical Society of America*, *131*(2), 1480–1489. https://doi.org/10.1121/1.3672703

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964–1973. https://doi.org/10.1162/jocn.2007.19.12.1964

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 lip-reading* (pp. 3-51). London: Lawrence Erlbaum Associates.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Xie, H., Zeng, B., Wang, R. (2018). Visual timing information in audiovisual speech perception: evidence from lexical tone contour. *Proceedings of Interspeech 2018, India*, 3781-3785. https://doi.org/10.21437/Interspeech.2018-1285.

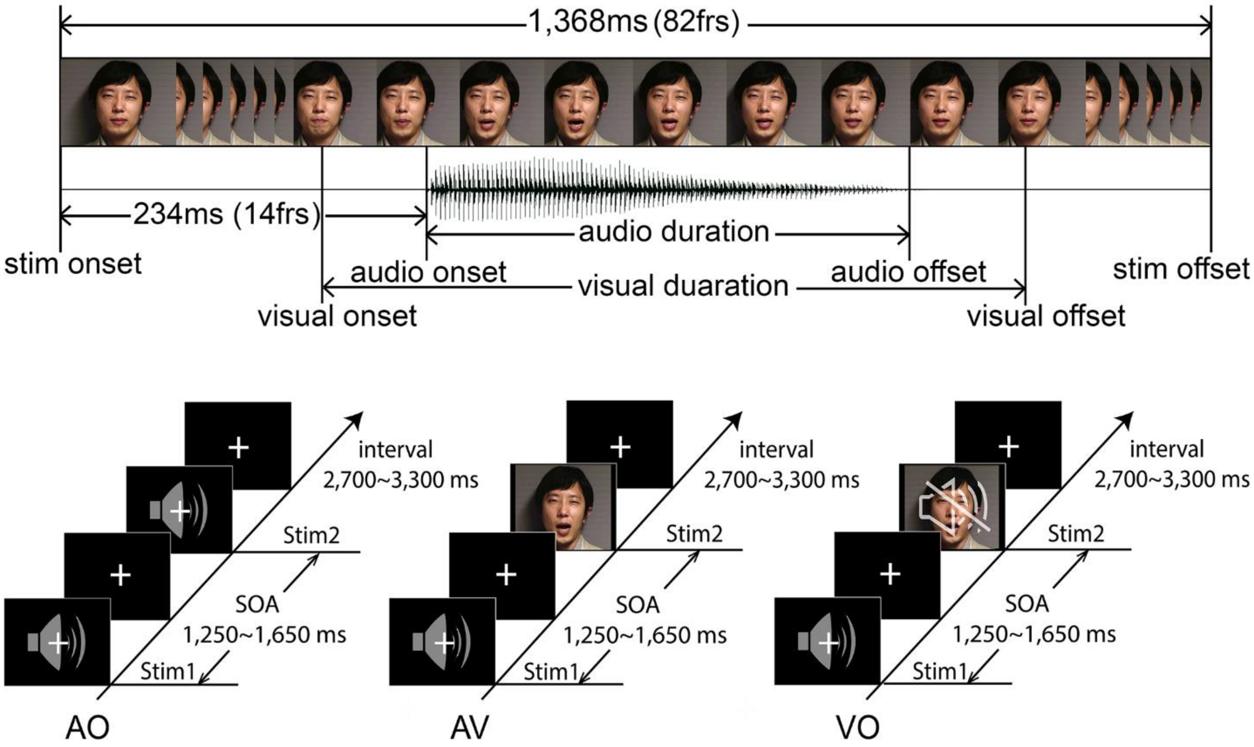
Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

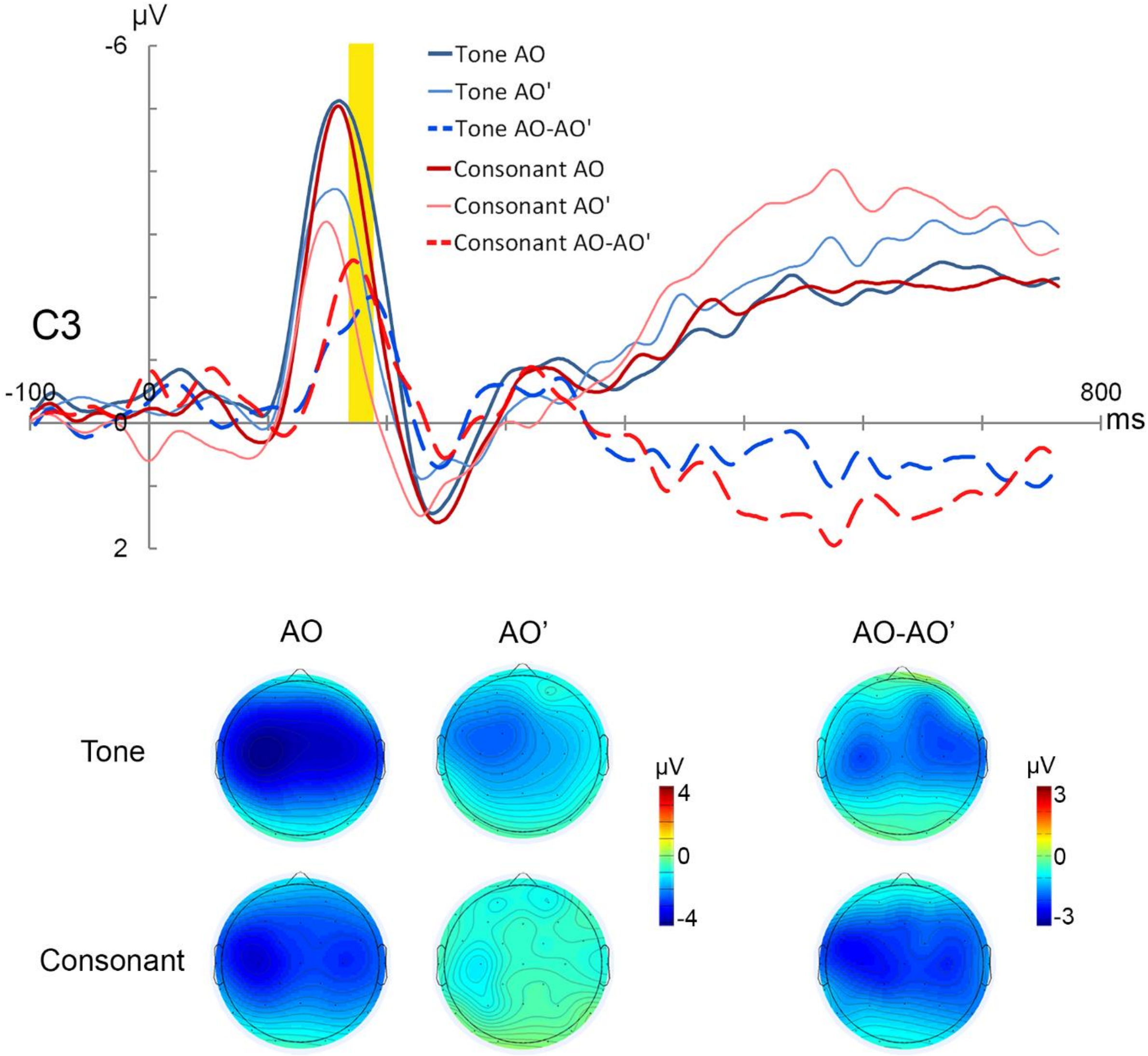
Figure Captions

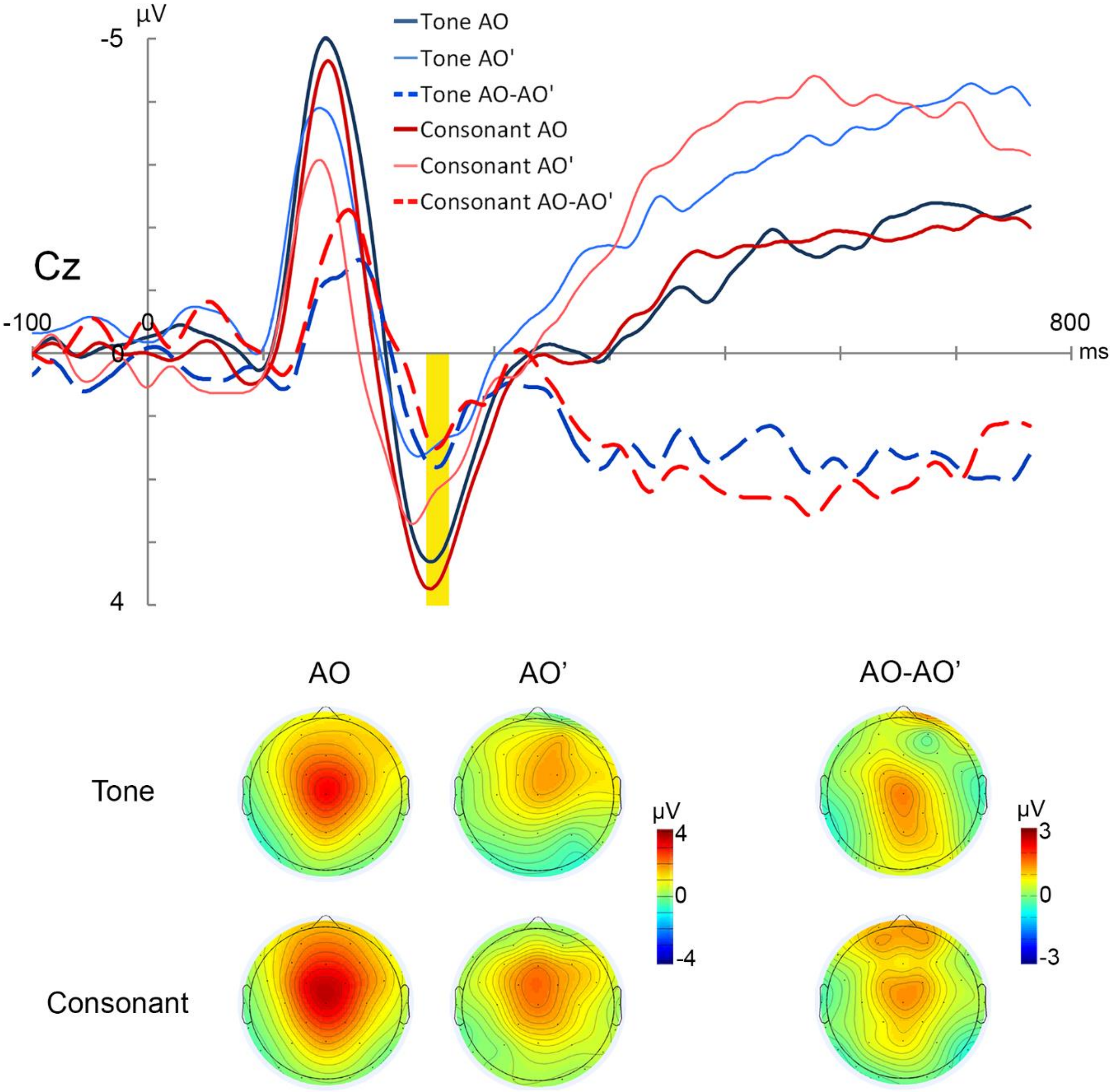
Fig. 1. The upper panel shows the structure of the audiovisual stimuli (frs = frames; SOA = stimulus onset asynchrony). Visual duration refers to the duration of visual mouth movement from opening to closure. The lower panel represents the trial sequences of audio-only (AO), audiovisual (AV) and visual-only (VO) conditions. Each trial presented two syllables (Stim1 and Stim2). Stim1 was always AO, whereas Stim2 was randomly presented in the three modalities within each block. Participants responded to lexical tone or consonant differences between Stim1 and Stim2 within a trial. The face images are used in the figure with the permission of the actor.

Fig. 2. ERP waveforms showing the auditory N1 component (maximal over C3). Topography maps illustrate the activity distributions over the analysis window (168-188 ms), which is highlighted in yellow.

Fig. 3. ERP waveforms showing the auditory P2 component (maximal over Cz). Topography maps illustrate the activity distributions over the analysis window (240-260 ms), which is highlighted in yellow.

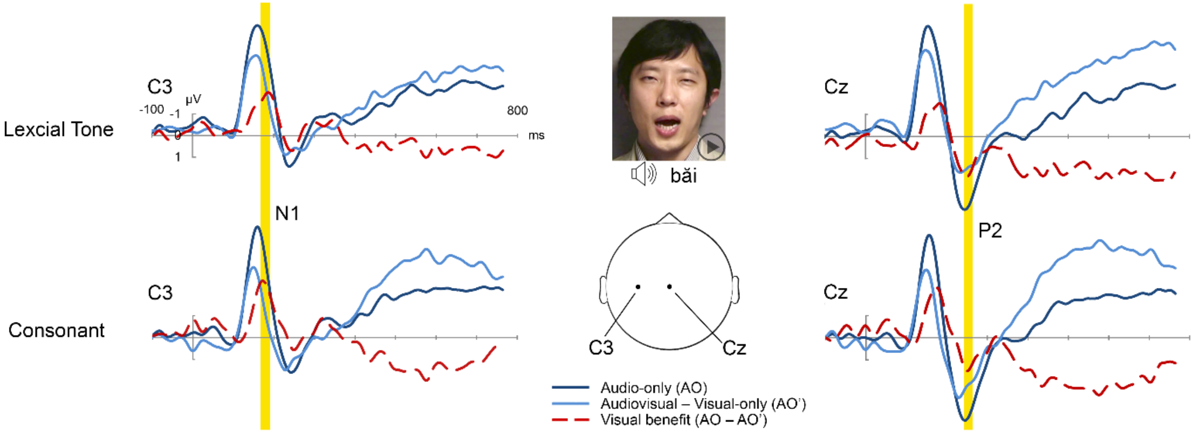

1,368ms (82frs)
234ms (14frs)
stim onset
audio onset
audio duration
audio offset
stim offset
visual onset
visual duaration
visual offset
interval
2,700~3,300 ms
Stim2
SOA
1,250~1,650 ms
Stim1
AO
interval
2,700~3,300 ms
Stim2
SOA
1,250~1,650 ms
Stim1
AV
interval
2,700~3,300 ms
Stim2
SOA
1,250~1,650 ms
Stim1
VO


µV
-6
2
C3
-100
0
0
800
ms
Tone AO
Tone AO'
Tone AO-AO'
Consonant AO
Consonant AO'
Consonant AO-AO'
AO
AO’
AO-AO’
Tone
Consonant
µV
4
0
-4
µV
3
0
-3


μV
-5
4
Cz
-100
0
0
800
ms
Tone AO
Tone AO'
Tone AO-AO'
Consonant AO
Consonant AO'
Consonant AO-AO'
AO
AO'
AO-AO'
Tone
Consonant
μV
4
0
-4
μV
3
0
-3

Graphical Abstract

Similar visual benefit: lexical tone vs. consonant



Supplementary Materials

Fig. S1. Illustration of the subtraction method generating the AO' waveforms, which are compared against the AO waveforms (AO = auditory perception; AO' = AV – VO, corresponding to auditory processing in the audiovisual condition with the visual processing removed with the subtraction). After the subtraction, the time zero is changed from the video onset to the audio onset (234 ms later). The arrow shows this change of time zero (shifting the vertical axis to the right by 234 ms), with time values on the horizontal axis measured relative to the new time zero. This new time zero was used in statistical analysis of the AO and AO' waveforms in the current study.

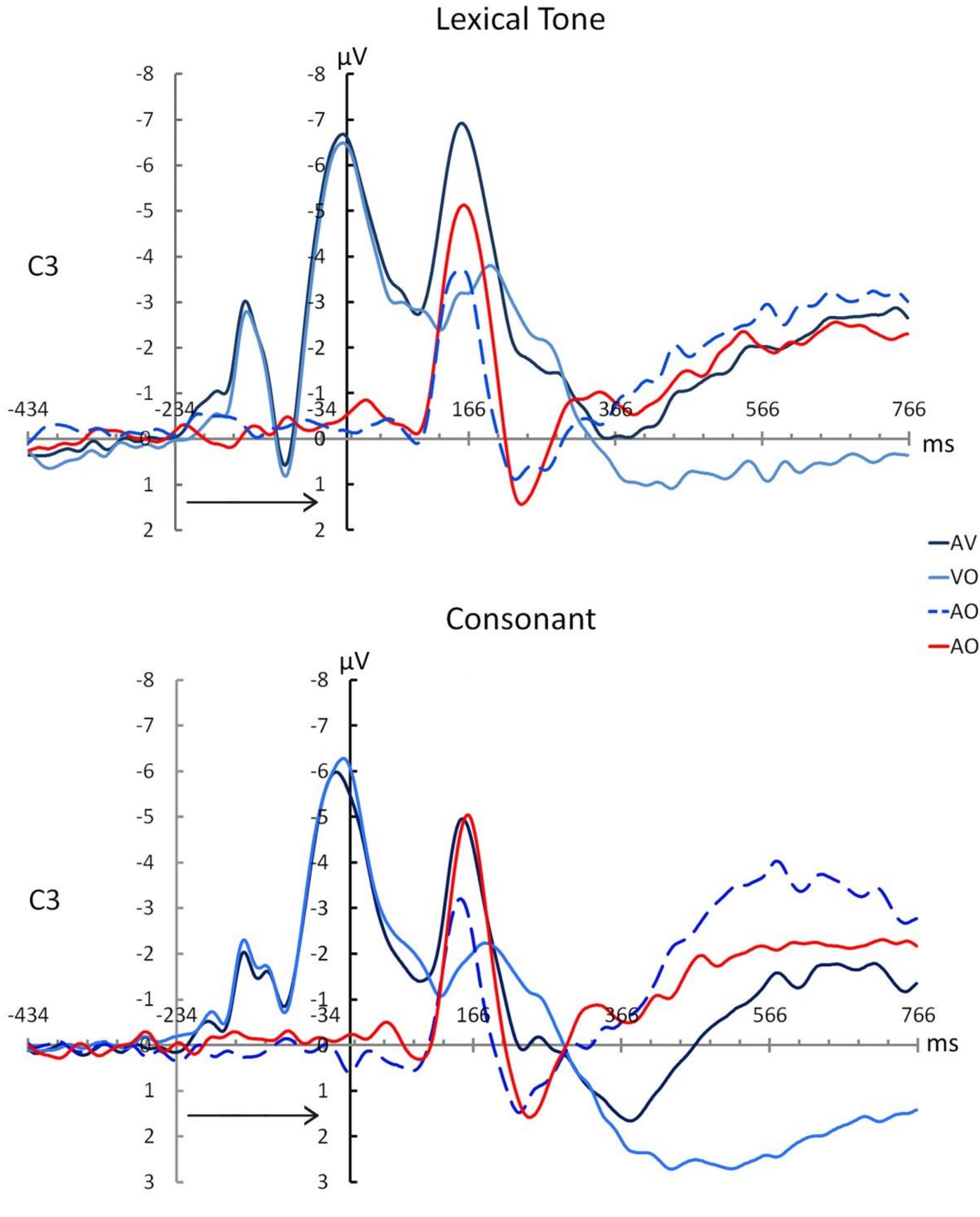


Table S1. Corpus of Mandarin syllables used in the experiments, including six syllables *bai, dai, tai, bao, dao, and tao* with four tones. T1 to T4 represents the high-level tone, the rising tone, the dipping tone and the falling tone respectively.

Syllable	Tone	Speaker	F0 (Hz)	Intensity (dB)	Voice onset time	Tone/Vowel Audio duration (ms)	Audio duration (ms)	Visual duration (ms)
bai	T1	1	124	66	16	848	864	1285
		2	136	65	10	1091	1101	1133
	T2	1	116	69	18	772	790	1134
		2	129	64	9	976	985	1200
	T3	1	95	62	17	940	957	1301
		2	112	62	10	1130	1139	1317
	T4	1	118	65	14	424	438	1084
		2	140	66	11	294	305	700
dai	T1	1	128	68	14	675	689	951
		2	139	64	13	874	887	1134
	T2	1	118	68	15	740	755	1034
		2	116	63	14	814	828	1134
	T3	1	96	65	41	819	860	1084
		2	105	61	8	957	965	1168
	T4	1	125	68	13	419	432	901
		2	138	65	11	306	317	901
tai	T1	1	130	67	129	586	715	901
		2	137	65	87	871	958	1051
	T2	1	113	66	168	747	915	1168
		2	117	66	88	878	966	1168
	T3	1	97	65	171	811	982	1185
		2	113	63	104	1018	1122	1251
	T4	1	121	67	127	384	511	934
		2	99	65	89	404	493	1051
bao	T1	1	122	71	17	665	682	1018
		2	128	69	84	740	824	918
	T2	1	112	70	16	653	669	1084
		2	112	69	18	652	670	918
	T3	1	97	68	16	744	760	1101
		2	115	64	11	816	827	968
	T4	1	116	69	15	426	441	968
		2	109	68	8	241	249	667
dao	T1	1	126	70	15	651	666	951
		2	131	68	11	757	768	951
	T2	1	114	70	24	646	670	968
		2	115	67	12	580	592	1134
	T3	1	99	68	24	757	781	1034
		2	119	62	18	849	867	1034
	T4	1	117	68	13	407	420	851
		2	134	69	15	252	267	584
tao	T1	1	130	70	112	564	676	934
		2	132	67	83	746	829	1084
	T2	1	116	70	125	605	730	1068
		2	118	69	79	754	833	1218
	T3	1	98	68	129	654	783	1151
		2	105	64	113	874	987	1285
	T4	1	110	69	114	426	540	901
		2	100	67	118	244	362	851