

Seeing pitch: Visual information for lexical tones of Mandarin-Chinese

Trevor H. Chen and Dominic W. Massaro^{a)}

University of California, Santa Cruz, Santa Cruz, California 95064

(Received 26 September 2006; revised 14 November 2007; accepted 7 January 2008)

Mandarin perceivers were tested in visual lexical-tone identification before and after learning. Baseline performance was only slightly above chance, although there appeared to be some visual information in the speakers' neck and head movements. When participants were taught to use this visible information in two experiments, visual tone identification improved significantly. There appears to be a relationship between the production of lexical tones and the visible movements of the neck, head, and mouth, and this information can be effectively used after a short training session. © 2008 Acoustical Society of America. [DOI: 10.1121/1.2839004]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Gv, 43.71.Bp [DOS]

Pages: 2356–2366

I. INTRODUCTION

A developing principle is that humans perceive by using multiple sources of information (Massaro, 1998). In the case of face-to-face speech, we use (at least) audition and vision to perceive what is spoken. Dozens of empirical studies and theoretical analyses indicate that perceivers combine or integrate audible and visible speech (e.g., Massaro, 1998; Massaro *et al.*, 2001). For example, when hearing the sound of an auditory /ba/ and seeing the mouth movement of a visual /ga/, perceivers usually perceive /da/, /ɔa/, or /va/ (Massaro, 1998; McGurk and MacDonald, 1976). Although there might be interlanguage differences in the degree of visual influence for different segments (e.g., Hayashi and Sekiyama, 1998; Sekiyama, 1997; Sekiyama and Tohkura, 1991, 1993), there is strong evidence for the principle that speakers of different languages integrate auditory and visual speech in a similar manner (e.g., Chen and Massaro, 2004; Massaro *et al.*, 1995, 1993).

Research on audiovisual speech perception has paid more attention to segmental information and less attention to lexical-tone information. In what seems to be the first study on the audiovisual perception of lexical tones, native identification of the six Cantonese tones was tested with auditory (sound), visual (face), and bimodal (both) stimuli (Burnham *et al.*, 2000). Performance averaged about 20% correct across certain visual-only conditions, which were statistically significant above the chance level of (1/6 Cantonese tones) 16.67% (Burnham *et al.*, 2000). In a same-different discrimination study on Cantonese tones, native Thai and Australian-English speakers also performed significantly better than chance under visual-only conditions (Burnham *et al.*, 2001).

In a study on the identification of Mandarin tones (Mixdorff *et al.*, 2005b), native Mandarin speakers identified tones with sound alone as well as sound plus the video (watching the lower part of a speaker's face). Under clear auditory conditions, the addition of the video did not significantly improve performance. However, under some noise-

masked conditions, the addition of video did significantly improve tone-identification performance relative to that of sound alone (but the effect was fairly small). The same patterns of results were also found in the native tone-identification of Vietnamese (Mixdorff *et al.*, 2006) and Thai (Mixdorff *et al.*, 2005a) as well as in the non-native discrimination of Cantonese tones by Thai speakers (Burnham *et al.*, 2001).

In another study (Burnham *et al.*, 2006), head motion was found to be informative for the perception of Cantonese tones. This interesting result fits well with the finding that head motion is also related to intonation (Yehia *et al.*, 2002). Others studies explored visual information for various supra-segmental prosodic (intonation, stress, etc.) features and found that head and eye-brow movements can be perceptually informative (e.g., Munhall *et al.*, 2004; Srinivasan and Massaro, 2003; Yehia *et al.*, 2002).

On the other hand, there may be additional visual information for lexical tones, but perhaps the perceivers did not use it because they were not fully aware of where to look and what to look for. Mandarin-Chinese presents an interesting case because the physical-acoustical properties of its lexical tones are well studied and well known (e.g., Chen, 1999; Connell *et al.*, 1983; Garding *et al.*, 1986; Jongman *et al.*, 2005). This tone-rich language may provide additional insights about the production of lexical tones and possible relationships with visible speech movements. There are four lexical tones in Mandarin, commonly referred to as tones 1, 2, 3, and 4. Based on the fundamental frequency (F0) patterns, tone 1 has been described as high level (5-5), tone 2 midrising (or mid-high-rising; 3-5), tone 3 mid-falling-rising (or low-dipping or low-falling-rising; 2-1-4), and tone 4 high-falling (5-1) (Chao, 1968; Lee-Schoenfeld, 2002). Mandarin tones also tend to differ on other dimensions such as vowel duration and amplitude. For example, vowel duration tends to be longest for tone 3 and shortest for tone 4; amplitude tends to be lowest for tone 3 and highest for tone 4 (Tseng, 1981).

^{a)}Electronic mail: massaro@fuzzy.ucsc.edu

A. Perception of Mandarin tones

Mandarin tone judgments are influenced by multiple sources of information: F0 pattern (both height and contour), vowel duration, and amplitude (Tseng *et al.*, 1986). F0 pattern appears to be the most influential information, with F0 height and F0 contour about equally effective. These cues are independently evaluated and optimally integrated for tone perception (Massaro *et al.*, 1985).

Although there is agreement that F0 pattern (perceived as voice pitch) is the most important/dominant phonetic cue for Mandarin tones, there are other acoustic dimensions that can be perceptually informative. Duration is systematically different depending on the tone. In isolation, phrase-final position, or citation form, tone 3 tends to be longer than tone 2, which tends to be longer than tones 1 and 4 (Blicher *et al.*, 1990; Chao, 1968). This duration difference is acoustically salient: Longer durations of auditorily presented /bi/, /ba/, and /bu/ elicited more tone 3 (and fewer tone 2) identifications for both native Mandarin and English speakers (Blicher *et al.*, 1990).

Another informative acoustic dimension is amplitude. Whalen and Xu (1992) manipulated natural speech (/ba/ and /yi/) by eliminating F0 but retaining amplitude contours. Tones 2, 3, and 4 were acoustically distinguishable on the basis of amplitude (and not duration) alone. They also found a positive correlation between F0 and amplitude (Whalen and Xu, 1992). Moreover, even in the presence of the F0 pattern, duration and amplitude can each be used as functional cues for tone judgments (Tseng *et al.*, 1986). Finally, vowel quality (i.e., the type of the vowel) did not systematically influence tone identification (Tseng *et al.*, 1986; Massaro *et al.*, 1983).

B. The current study

Given the previous findings on the auditory perception of Mandarin lexical tones, one can speculate on the possible visible dynamics in producing these different tones. Duration differences might be seen from the speaker even if the sound is not available. For example, visible speech rate has been shown to influence the perception of voice onset time of initial consonants (Green and Miller, 1985). Also, it may be possible that loudness or intensity can be reflected by the degree of hyperarticulation or exaggeration of mouth movements (Kim and Davis, 2001) or simply perceived effort (Rosenblum and Fowler, 1991). It is also conceivable that speakers may somehow express lexical tone information in terms of some visible paralinguistic cues, whether consciously or unconsciously.

The goals of the current study are to determine: (1) a baseline accuracy of Mandarin lexical-tone identification from visual-only information; (2) if there are systematic visible changes from lexical tone production and the nature of this information; and (3) whether Mandarin speakers can be taught to use this information to significantly improve visual identification performance.

In addition, it is also interesting to examine performance for each of the four tones. For example, tones 2 and 3 tend to be the most acoustically confusable pair (Blicher *et al.*,

1990), and it will be interesting to see if this is also the case visually. Finally, we ask the question whether it is easier to recognize one's own visible speech more accurately than the speech of others. Previously, one study found no overall significant differences in speech-reading performance (of numbers in French) comparing watching one's own face with watching the faces of others (Schwartz and Savariaux, 2001). The present study will assess whether this is also the case for Mandarin lexical tones.

II. EXPERIMENT 1: TRAINING VISUAL SPEECH PERCEPTION

There may be additional visual information for lexical tones, but perceivers may not have taken full advantage of this information because they were not fully aware of where to look and what to look for. If we learn about the nature of this information, it may be possible to teach the participants to use it. Experiment 1 was a within-subjects design involving three different sessions. The strategy was to measure visual tone identification before and after training on potential visual information.

A. Method

1. Participants

Eight Chinese participants were recruited from the University of California, Santa Cruz (UCSC). They are all native speakers of Mandarin. Four of them are from Mainland China: Two females (one was 26 years old, and the other chose not to reveal her age but appeared to be in her 30's) who had been in the United States for about 3.5 and 1.75 years, and two males (ages 31 and 29) who had been in the United States for about 1.5 and 3 years. The other four participants are from Taiwan: Two females (ages 19 for both) who had been in the United States for about 4 and 9 years, and two males (ages 25 and 21; one of them was the senior author) who had been in the United States for about 12 and 8 years. The ages when exposure to English began were 10 and 12 for the females from Mainland China (FC), 13 and 12 for the males from Mainland China (MC), 13 and 7 for the females from Taiwan (FT), and 12 and 14 for the males from Taiwan (MT). They were paid at the rate of \$10/h for their participation.

2. Stimuli

We made sets of audio/video recordings from four speakers (one female from Mainland China, FC; one male from Mainland China, MC; one female from Taiwan, FT; and one male from Taiwan, MT) pronouncing 40 (10 syllables \times 4 tones) Mandarin-Chinese characters or words (chosen from Liu and Samuel, 2004), which are shown in Appendix A. These four speakers also served as participants for this experiment. The words were all familiar to the participants. Speakers from Mainland China read abbreviated (simplified) words, and speakers from Taiwan read traditional words. The words to be read were displayed as slides (Microsoft POWERPOINT™) on a standard computer screen. The words were read both in isolation and following a neutral context phrase in separate blocks of 40 trials. The trials were randomized in

four different versions—two versions for characters in isolation (i.e., citation form) and two versions for characters in a neutral sentential context (i.e., “the next word is”). The speakers were recorded pronouncing all words in the four versions of randomization.

3. Design

The experiment took place on three days: 1, 2, and 3. Day 1 was the recording session: The four speakers were told that the experimental purpose was to understand “how people say words,” and their task was to do their best to pronounce the characters or words “clearly and distinctively.” After day 1 was completed, eight participants (including the original four speakers) were later contacted to schedule for day 2.

On day 2, they participated in a tone-identification task: They were told to watch the video (with no sound available) and choose (circle) what character or word was said for each trial. There were four versions of randomization [2 conditions (context or citation) \times 2 versions for each], and 16 blocks (4 speakers \times 4 versions) were arranged from pseudo-randomization (randomization but making sure that no speakers of the same gender appeared in consecutive blocks). Appendix B shows an example of one of the response sheets used for the tone-identification task. The total number of trials was 640 (16 blocks \times 40 words each) for each of the participants in each session. There were approximately 5 s between the word presentations on isolation trials and 7 s on context trials. The video was displayed on a JVC color video monitor (TM-131 SU) screen, which was approximately 11.25 in. in width and 8.5 in. in height. The faces averaged approximately 4.75 in. in width and 5.7 in. in height.

A few days after day 2, all of the participants were contacted to schedule for another day (day 3). On day 3, this time they were taught to use a specific strategy for tone identification, and they completed the tone-identification task again. Afterwards, they completed an optional questionnaire.

4. Procedure

The participants were not told about their subsequent participation. On day 1, they were not told that they would participate for day 2; on day 2, they were not told that they would participate for day 3. There were at least 14 days between day 1 and day 2, and there were at least 6 days between day 2 and day 3. All instructions and interactions with the experimenter (MT) were spoken in Mandarin. One MT (senior author) served as his own experimenter. This person was one of the participants, and he had not anticipated his subsequent participation for day 2 (although he expected a subsequent participation for day 3).

5. Day 1: Recording

After day 1 was completed for one speaker (MT), the recording was visually examined for possible sources of visual information. The first author observed that some visible information for Mandarin lexical tones seemed to be available from the activity around the lower parts of the neck and from head movements. Its visual-perceptual clarity surprised

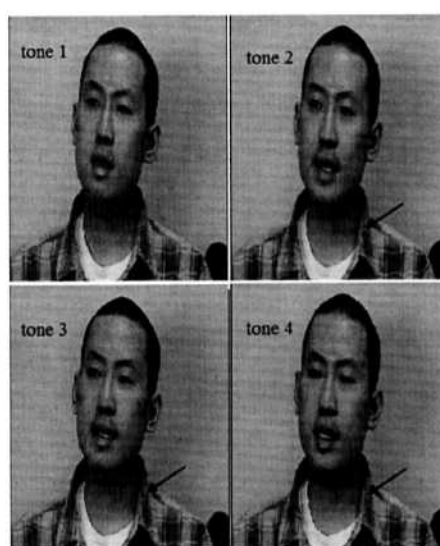


FIG. 1. Pictorial illustration of visible information for Mandarin tones at the lower parts of the neck. The bulge on the side of the neck changes across the four tones. Tone 1 has minimal movement, tones 2 and 4 have some movement, and tone 3 has the biggest bulge.

both a native and a non-native Mandarin speaker. The judgments about the sources of information were made and agreed on by the experimenter and another researcher. It appeared that there were the least (or minimal) neck movement for tone-1 words, some movements for tone-2 words, most movements for tone-3 words, and some (brief) movements for tone-4 words. Figure 1 illustrates the visible information for each of the four Mandarin tones. The bulge on the side of the neck changes across the four tones. Tone 1 has minimal movement, tones 2 and 4 have some movement, and tone 3 has the biggest bulge. Although the speakers were not told about this, an effort was made to indirectly persuade two participants during the recordings to replace or put back hair and/or clothing because they covered parts of the neck.

One speaker's (FC) hair was covering the left and right sides of her lower neck, while another speaker's (MC) shirt was buttoned-up and the collar covered parts of his lower neck. Indirect efforts trying to uncover all areas of their neck were unsuccessful. However, examination of the video, after all recording sessions were complete, revealed that all speakers showed some visible information. In particular, FC appeared to drop/dip her head/chin on tone-3 words (despite her hair covering her neck); MC's glottis and a part of his uncovered neck appeared to display activity patterns consistent with those hypothesized from MT; and for FT, her neck also appeared to display activity patterns consistent with MC and MT, at the same time her head/chin movements seemed consistent with speaker FC. These observed visible patterns are summarized in Table I, and this table was used in an information sheet to inform participants about the strategy for day 3. (It is possible that the lexical tones differed in duration, which could be seen visually. These possible durational cues were also mentioned on day 3.)

6. Day 2: Identification

For day 2, participants were not given the information sheet and not told about any strategy; they were only in-

TABLE 1. Summary sheet used to inform participants about the visible-information strategy (Experiments 1 and 2).

	Tone 1	Tone 2	Tone 3	Tone 4
Pitch (frequency)	High-level	Mid-rising	Mid-falling-rising	High-falling
Loudness (amplitude/intensity)	In-between	In-between	Quiet	Loudest
Duration (Time)	Short	Long	Longest	Shortest
Neck	Tone 1 No (least) activity	Tone 2 Some activity	Tone 3 Most activity	Tone 4 Some (brief) activity
Chin			Females drop head/chin	
Mouth				

structured to watch the speaker and choose (circle) what character or word was spoken in each trial. Participants used response sheets in the forms similar to that shown in Appendix B, except those from Mainland China saw abbreviated (simplified) characters and those from Taiwan saw traditional characters (Appendix B shows traditional characters). The experimental sessions were approximately 2 h long, with 10 min breaks after approximately every 15 min.

7. Day 3: Training and identification

For day 3, participants were informed about the acoustic-physical characteristics of Mandarin tones and how these dimensions may relate to visible activities of the neck, head, and mouth movements. Participants were allowed to take the summary information sheet (Table 1) into the subject-testing room during the experiment. They were instructed to pay attention to mouth, head/chin movements, and especially activities of the neck. Although there were no special written descriptions for the mouth on the summary information sheet, it was specifically pointed out that duration (time) differences may be reflected from the mouth. During the strategy-training time, they were shown a short VHS tape (for no more than about 15 min) that included samples of representative trials (in order to illustrate the strategy), and they were also given roughly 10–20 practice trials with feedback (to help learn this strategy). The whole training time lasted for no more than about 45 min for each participant. After this training time, participants completed the tone-identification task again.

B. Results

The results from the identification task were pooled over versions of randomization because an initial analysis showed no significant differences between them. The independent variables were day (two levels: day 2 and day 3), v-gender (two levels: gender of the speakers in the video, not gender of the perceiver), context (two levels: with or without), and lexical tone (four levels). The dependent variables were accuracy of tone-identification performance and d' .

Figure 2 shows the accuracy of tone-identification performance plotted as a function of each of the four tones on

Exp. 1: Day and Tone

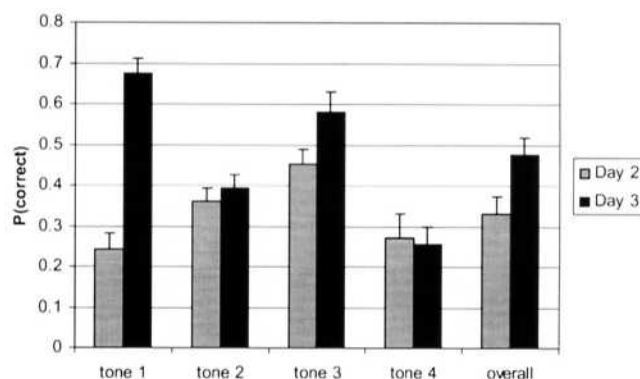


FIG. 2. Experiment 1: Tone-identification performance, plotted as the probability of correct responses (Y axis), for overall and each of the four tones on day 2 and day 3.

day 2 and day 3, as well as overall across the four tones. Analysis of variance revealed that performance on day 3 (mean=0.48) was significantly higher than day 2 (mean=0.33), $F(1,7)=68.79$, $p<0.001$. In follow-up t-tests, both day-2 and day-3 mean performances were significantly greater than the chance level of 0.25 [day-2: $t(127)=5.16$, $p<0.001$; day-3: $t(127)=11.49$, $p<0.001$]. Also, performance was generally better when the speaker of the video was a female than when the speaker was a male, $F(1,7)=20.06$, $p<0.01$. It is conceivable that, because the two females in the videotape tended to move their head/chin in a dipping fashion (consistent with the tone-3 frequency pattern) while pronouncing tone-3 words, this extra information would have helped increase performance over the two male speakers whose head movements, if any, were not as salient. The significant interaction between tone and v-gender [$F(3,21)=22.79$, $p<0.001$] indicated that the female-speaker advantage was most pronounced for tone-3 words.

There was a significant main effect for tone [$F(3,21)=11.77$, $p<0.001$] and a significant interaction between day and tone, $F(3,21)=12.54$, $p<0.001$. Consistent with our hypotheses, tone-1 and tone-3 words improved more than tone-2 and tone-4 words. We computed the 95% confidence intervals (CI) around the means of each of the tones for both days; this allows comparisons to the chance value of 0.25. Within a given experiment, if a CI is completely above 0.25, its mean is significantly higher than this chance value. On day 2, tone 1 was not significantly different from chance, but on day 3 it became significantly above the chance level. The tone-1 performance was the lowest among the four tones on day 2 but highest among the tones on day 3. For tones 2 and 3, their CI's were above the chance level on both days, although tone 3 performance improved much more relative to that of tone 2. Improvement on these tones did not appear to come at the detriment of other tones. On both days, tone-2 performance was slightly higher than chance, and tone-4 performance was not significantly different from chance.

The main effect of context was not significant, $F(1,7)=0.94$, $p=0.37$. The only other significant interaction was between v-gender and context [$F(1,7)=22.10$, $p<0.01$], with the female-speaker advantage more pronounced under

Exp. 1: Individual Overall

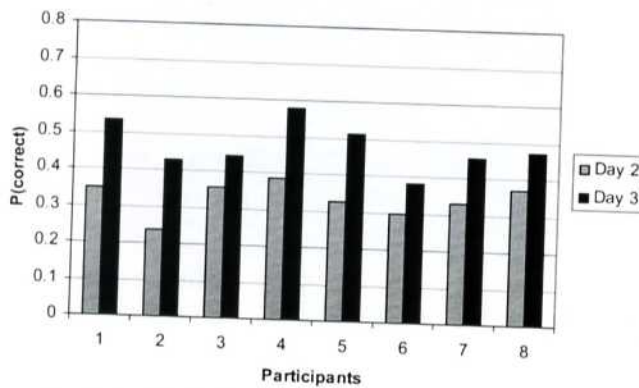


FIG. 3. Experiment 1: Individual performances, plotted as the probability of correct responses (Y axis), for the tone-identification task on day 2 and day 3.

the no-context condition. Figure 3 plots the accuracy performance of all the individual participants for the tone-identification task on day 2 and day 3. As can be seen in Fig. 3, every individual improved from day 2 to day 3.

In another analysis of variance, day (two levels) and syllable (ten levels) were included as independent variables. This analysis revealed significant effects for day [$F(1,7) = 68.79, p < 0.001$] and syllable [$F(9,63) = 3.71, p < 0.01$]. There was no significant interaction between day and syllable [$F(9,63) = 1.13, p = 0.35$]. Figure 4 plots the tone-identification accuracy performance for the syllables on day 2 and day 3. As can be seen in Fig. 4, the day-3 training advantage was reflected for all of the syllables.

Given the seemingly dramatic improvement on visual tone identification after training, an important question is whether the day-3 advantage was mainly due to simply experience in the identification task. Obviously, participants had practice on day 2, and one might argue that improvement could be due to previous exposure, learning, and/or memory. To test whether this practice was a major factor, we analyzed the results across trial blocks on days 2 and 3. Figure 5 plots the overall performance across blocks on day 2 and day 3, which shows no learning within a given day. Linear regres-

Exp. 1: Day and Syllable

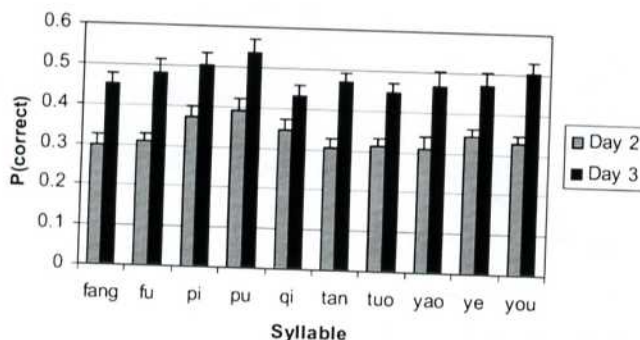


FIG. 4. Experiment 1: Tone-identification performance, plotted as the probability of correct responses (Y axis), for each of the syllables on day 2 and day 3.

Exp. 1: Across Blocks

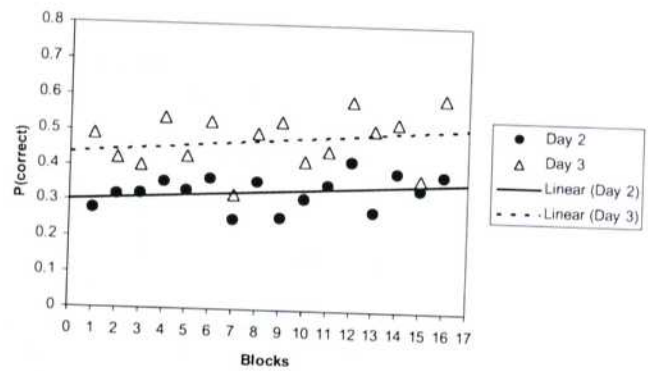


FIG. 5. Experiment 1: The overall tone-identification performance across blocks.

sion analyses showed that the slopes of the best fit lines were not significantly above zero for either day 2 (slope=0.004, $p=0.17$) or day 3 (slope=0.005, $p=0.27$), suggesting that experience in the task was not a significant factor.

We calculated d' values as a measure of identification performance independent of any decision bias. We computed hit and false alarm (FA) rates for each participant, for each tone, and for each day. A given participant's tone-1 hit rate, for example, is the probability of correctly identifying that tone (i.e., the number of correct tone-1 responses divided by the number of tone-1 trials). A given participant's FA rate for tone 1, for example, is the probability of mistakenly responding tone 1 to stimulus tones 2, 3, and 4 (i.e., the number of incorrect tone-1 responses divided by the total number of trials of tones 2, 3, and 4). The d' is an index of how well the participant distinguishes one lexical tone from the others. The bigger the d' value, the better the participant is at recognizing the tone.

Figure 6 plots the d' values for overall and the four tones on days 2 and 3. We carried out an analysis of variance on d' values. The independent variables were tone (four levels) and day (two levels). This analysis revealed significant effects for day [$F(1,7) = 60.55, p < 0.001$], tone [$F(3,21) = 105.29, p < 0.001$] and their interaction [$F(3,21) = 17.50, p < 0.001$]. The d' values for day 3 were higher than those

Exp. 1: d' for Day and Tone

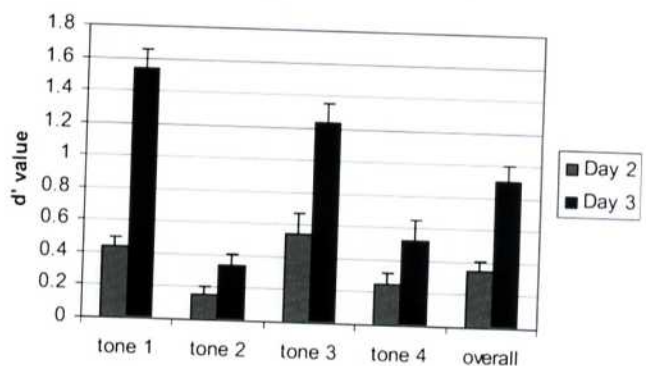


FIG. 6. Experiment 1: The d' values for overall and each of the four tones on day 2 and day 3.

for day 2 (both were significantly greater than zero), and this was the case for all four tones, with the biggest increase at tones 1 and 3. This suggests that the improvement was unlikely due to guessing or item biases—participants were better at identifying all four of the target tones on day 3 than on day 2.

Finally, an analysis of variance compared the accuracy performance from watching one's own face and that from watching another's face. Four of the eight participants watched both their own faces and other's faces, so only data from these participants were used in this analysis. There was no significant difference in performance between watching one's own face (mean=0.467) and watching another's face (mean=0.396), $F(1, 3)=1.62$, $p=0.29$. There was also no interaction between face (own or another) and day, $F(1, 3)=0.12$, $p=0.75$.

C. Discussion

Our results suggest that (1) untrained native-Mandarin visual tone identification is close to but statistically significantly above chance, and (2) when participants are taught a specific identification strategy (i.e., using mouth, head/chin movements, and especially activities of the neck), their tone-identification performance improves significantly to a level well above chance. Specifically, tone-1 and tone-3 words improved more than tone-2 and tone-4 words. Identification was better when the speakers were female (particularly for tone 3), probably because they tended to dip their head more prominently when speaking tone-3 words.

Our baseline results without training are consistent with the findings of Burnham *et al.* (2000) in that their Cantonese observers also performed just slightly above chance. Also consistent with the study by Schwartz and Savariaux (2001), we found no significant difference in performance between watching one's own face and watching others' faces.

Overall, the accuracy of tone-identification performance improved from about 33% before training to about 48% after training. Half of the participants in this experiment recorded the test stimuli, and this could have given them some advantage. The recording experience might have rendered them somewhat "special." However, separate analyses indicated that these participants' performance and response patterns were very similar to the rest of the participants, as the significant effects of day and v-gender and their significant interaction were observed even when the participants were divided into these two groups. In both groups, the improvement was greatest for tones 1 and 3. Nevertheless, a next step was taken to determine if the findings could be replicated.

Furthermore, it is still logically possible that the advantage after training was due to experience in the task. Experiment 2 compares the positive effect of instructions without prior experience in the task. In addition, we eliminated neck information for another group of participants in order to determine to what extent it contributed to performance of untrained participants.

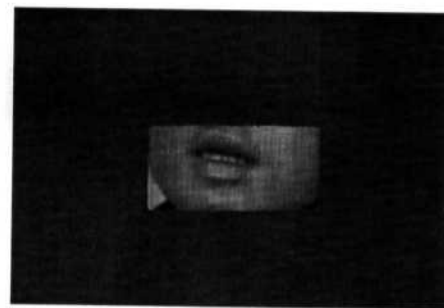


FIG. 7. Experiment 2: A picture showing the TV screen in the box condition.

III. EXPERIMENT 2: REPLICATION AND EXTENSION

Experiment 2 was a 1 day, mixed-subjects design. Participants completed the tone-identification task under one of three conditions: regular (normal), strategy (special instruction), or partially obstructed (only seeing the mouth). If the instruction group performed higher than the regular group, this would provide strong evidence in favor of the potential effectiveness of the visual information.

A. Method

A total of 24 Chinese participants were recruited from UCSC. They are all native speakers of Mandarin. Their ages ranged from 19 to 33 years old. Of the 21 participants who revealed their age (2 chose not to reveal their age but appeared to be in their 20's, and 1 hinted that she was in her early 40'), the average age was 24.3. Of those who chose to answer the question, participants on average had been in the United States for 5.2 years, and English exposure began around the average age of 10.7. They were either paid a \$15 gift certificate (for the UCSC Bay-Tree Bookstore) or given course credits for their 2 h participation.

There were eight participants under each of the three conditions: normal, box, and strategy. The same set of stimuli from the previous experiment was used, and participants completed the same tone-identification task: They were told to watch the speaker (with no sound available) and choose (circle) what character or word was said for each trial. Under the normal condition, participants simply watched the video screen and performed the task. Under the box condition, participants could only see the mouth region of the speakers—a piece of cardboard covered the screen except for a cut-out rectangle hole (8 cm width \times 4 cm height) in its approximate center, showing the middle of the screen. Figure 7 shows a picture of the screen in the box condition. Under the strategy condition, participants were taught and instructed to use our specific strategy in the same manner as the last day of Experiment 1. The rest of the procedure is the same as the previous experiment. All instructions and interactions were spoken in Mandarin.

B. Results and discussion

We carried out analysis of variance and pairwise comparisons. The independent variables were group (between-subjects, three levels: box, normal, strategy), v-gender (two

Exp. 2: Group and Tone

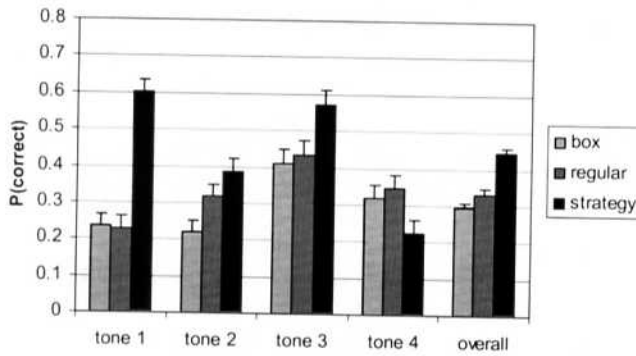


FIG. 8. Experiment 2: Tone-identification performance, plotted as the probability of correct responses (Y axis), for overall and each of the four tones for the three groups.

levels: gender of the speakers in the video, not gender of the perceiver), context (two levels: with or without), and lexical tone (four levels).

Figure 8 shows each of the tone's overall identification performance for all three groups. There was a significant overall difference among the three groups, $F(2,21)=37.60$, $p<0.001$. Specifically, group 3 (strategy, mean=0.45) scored significantly higher than both group 2 (normal, mean=0.33) and group 1 (box, mean=0.30). The difference between groups 1 and 2 was not quite significant [$F(1,14)=3.97$, $p=0.07$], so it remains an open question whether neck information is used by naïve untrained observers. There was a significant main effect for tone [$F(3,63)=13.08$, $p<0.001$], and there was a significant interaction between group and tone [$F(6,63)=8.91$, $p<0.001$]. We obtained CI around the means for each tone and for each group. For group 1 (box), only tone 3 was significantly higher than chance. For group 2 (normal), tones 3 and 4 were significantly higher than chance. For group 3 (strategy), tones 1, 2, and 3 were significantly higher than chance. Improvement on these tones did not appear to come at the detriment of other tones. Performance was the highest for group 3 on all of the tones (with the biggest differences at tones 1 and 3) except tone 4. There was no significant difference between the three conditions (box, normal, strategy) in terms of tone-4 performance.

Performance was again generally better when the speaker in the video was a female than when the speaker was a male, $F(1,21)=23.97$, $p<0.001$. The interaction between v-gender and tone was again significant [$F(3,63)=31.42$, $p<0.001$], with the video-female advantage most pronounced for tone 3. There was no significant main effect for sentential context [$F(1,21)=0.49$, $p=0.49$], and no significant interaction between tone and context, [$F(3,63)=1.06$, $p=0.38$]. The only other significant interaction was between v-gender and context [$F(1,21)=8.64$, $p<0.01$], with the female-speaker advantage again being more pronounced under the no-context condition.

In another analysis of variance on tone-identification performance, group (three levels) and syllable (ten levels) were included as independent variables. This analysis re-

Exp. 2: Group and Syllable

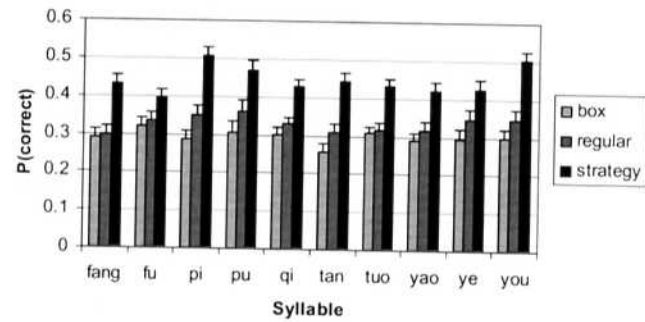


FIG. 9. Experiment 2: Tone-identification performance, plotted as the probability of correct responses (Y axis), for each of the syllables of the three groups.

vealed significant effects for group [$F(2,21)=37.60$, $p<0.001$] and syllable [$F(9,189)=2.56$, $p<0.01$]. There was only a marginally significant interaction between group and syllable, $F(18,189)=1.62$, $p=0.06$. Figure 9 shows the tone-identification accuracy for the syllables of the three groups. As can be seen in Fig. 9, group 3 (strategy) had the highest tone-identification accuracy for all of the syllables.

We computed the d' values for all participants, for each tone and each group. An analysis of variance on the d' values with the independent variables tone (four levels) and group (between-subjects, three levels) revealed significant effects for group [$F(2,21)=34.15$, $p<0.001$], tone [$F(3,63)=84.65$, $p<0.001$], and their interaction [$F(6,63)=24.73$, $p<0.001$]. Pairwise comparisons showed that the d' values for group 3 (strategy) were higher than those for group 2 [$F(1,14)=36.74$, $p<0.001$] and group 1 [$F(1,14)=60.20$, $p<0.001$], while the difference between groups 2 and 1 was not significant [$F(1,14)=3.77$, $p<0.07$].

Figure 10 plots the d' values for each of the four tones of the three groups and also averaged across the four tones. As can be seen from Fig. 10, the d' values from group 3 were the highest for all four tones, and the differences were the most pronounced at tones 1 and 3. Specifically, as neck movement for tone 1 was minimal or nonexistent compared to the other three tones, it seemed to be the most visually distinctive tone. Tones 2, 3, and 4 were more visually con-

Exp. 2: d' for Group and Tone

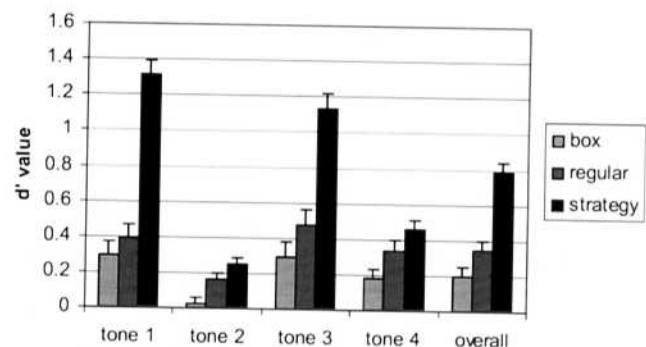


FIG. 10. Experiment 2: The d' values for overall and each of the four tones of the three groups.

fusable because they all had some neck movements. Neck movements for tone 3 tended to be more pronounced than those for tones 2 and 4, and neck movements for tone 4 tended to be slightly shorter than (and differ in timing from) tone 2. Some speakers (especially females in our video) spoke tone-3 words with a dipping head/chin movement, thus providing an extra source of visual information, which presumably helped improve tone-3 performance. Tones 2 and 4 were relatively harder to visually distinguish from each other, with the difference in onset and timing being the only reliable information observed.

Although the current study found that training improves visual tone identification, an argument might be that this training was only effective for some of the current particular speakers in our video. To explore this important question, we examined whether performance improvement occurred for each of the individual speakers. The tone-identification improvement was significant for all four speakers in both experiments 1 and 2. For both experiments, we conducted separate analyses of variance for each of the four speakers. In experiment 1, comparing performance on day 2 and day 3, identification accuracy significantly improved for each of the four speakers: Mainland-China female [$F(1,7)=16.78, p < 0.01$], Mainland-China male [$F(1,7)=7.63, p < 0.05$], Taiwan female [$F(1,7)=50.40, p < 0.001$], and Taiwan male [$F(1,7)=65.81, p < 0.01$]. For all four speakers, day 3 performance was significantly greater than that of day 2 as well as greater than the chance level. Moreover, in experiment 2, comparing performance among groups 1 (box), 2 (regular), and 3 (instruction), identification accuracy also significantly improved for each of the four speakers: Mainland-China female [$F(2,21)=6.04, p < 0.01$], Mainland-China male [$F(2,21)=9.45, p < 0.01$], Taiwan female [$F(2,21)=30.77, p < 0.001$], and Taiwan male [$F(2,21)=37.99, p < 0.001$]. For all four speakers, performance in group 3 was greater than that of groups 1 and 2 as well as greater than the chance level.

Another argument might be that the training may only be effective for the particular stimuli used under the particular condition. This may be true, because the current speakers were asked to speak "clearly and distinctively." It is not clear whether these potential sources of visual information can be effective in the everyday real world outside of the laboratory. Nevertheless, the current findings are unexpected and may serve as a first step raising the possibility of ecological and functional visual tonal information in the real world. Future research can test more speakers under various conditions or environments.

In both experiments 1 and 2, performance on the tone-identification task was better when the participants followed a specific strategy of using neck and head/chin movements. In addition, performance was generally higher when the speaker was female, especially for tone 3. However, perhaps it is important to emphasize that the interactions between tone and v-speaker likely arose from the specific behavioral patterns of the individuals in our video, rather than from their gender per se. For example, the two female speakers in our video tended to dip their head/chin while pronouncing tone-3 words, whereas the males did not.

Mandarin tones also differ in duration, and it is possible that duration differences might be visible. We measured the duration (in seconds) of the spoken words in citation (using both the spectrogram and the wave forms), and indeed the general pattern of duration differences between tones was replicated. Overall, tone-3 words (mean=0.647 s) were longer than tone-2 words (mean=0.623 s), which were longer than tone-1 (mean=0.593 s) and tone-4 (mean=0.483 s) words. An analysis of variance was carried out on stimulus duration, with the independent variables of syllable, tone, and version. The only significant results were the main effects for tone [$F(3,9)=6.39, p < 0.05$] and syllable [$F(9,27)=9.59, p < 0.01$]. None of the interactions was significant. Pairwise comparisons showed that tones 1, 2, and 3 were all significantly longer in duration than tone 4, but there were no significant differences among tones 1, 2, and 3. Although the tones did not differ much in duration, it is still possible that duration information contributed to performance.

IV. GENERAL DISCUSSION

We found that naïve participants only recognize Mandarin visual tones slightly above chance, and this finding is consistent with previous research on Cantonese tones (e.g., Burnham *et al.*, 2000). In addition, the present study found that there is more visual information for Mandarin tones than anticipated. With instruction, participants appeared to be able to use the speakers' neck activities and head/chin movements to identify tone. Previous research found that head movements are related to the perception of intonation (e.g., Yehia *et al.*, 2002; Srinivasan and Massaro, 2003) and Cantonese tones (Burnham *et al.*, 2006), and it is likely that this can also be the case for other lexical tones (e.g., Mandarin tone 3), at least when speaking clearly and distinctively.

It is possible that this finding represents only one case of a general phenomenon. Singers' neck and head movements may also reflect musical changes in pitch, intensity, and/or duration. Other examples may be observed from talking (stress) and yelling. For the neck, it is possible that its visible activity may be most pronounced with changes in frequency (though different parts of the neck may be visible for high or low frequencies).

Initially, the neck seemed to be an unexpected source of visual information. Although the speech production of pitch changes is typically described in terms the vocal fold's vibration caused by subglottal pressure and laryngeal musculature/tension (Ladefoged, 2001; Lass, 1976; Lehiste, 1970; Lieberman, 1967; Mullin *et al.*, 2003), whether there is any activity visible on the neck surface is virtually never discussed or even mentioned. However, the current study's finding is not that surprising in hindsight. At high frequency, the glottis tends to rise (this can be felt by putting fingers on the glottis); at low frequency, the muscles at the lower part of the neck tend to bulge outward (this can be felt by putting fingers on the lower neck). Our tentative speculation is that some muscles around the neck move (or move along with) the glottis, and the movement is visible, but which specific muscles are involved in lexical-tone production remain to be

identified. The current study at least revealed that some of their activities could be seen on the neck surface.

We believe that the improvement in visual tone-identification performance in the current study was mainly due to training the participants to use movements of the head and neck. Although extensive training with prolonged feedback usually improved performance (e.g., Gesi *et al.*, 1992; Massaro *et al.*, 1993; Massaro and Light, 2004; Wang *et al.*, 1999), these long-term training sessions spanned across days to weeks. In our current study, the one-time training was mostly verbal instruction and show-and-tell, and there were only a few practice trials in which feedback was given. The instruction, show-and-tell, plus practice trials lasted no longer than about 45 min. In experiment 1, there was a sudden change in the patterns of responses after training relative to before training. While we do not deny that feedback helps, it is likely that telling participants where to look and what to look for was sufficient to enable the participants to identify the tones.

In our study, we showed that performance significantly improved when participants were taught a specific visual-information strategy. This one-time training was mostly instruction and show-and-tell with feedback; participants were only given about 10–20 practice trials before the test session but no feedback during the test session. Teaching this strategy improved performance both within subjects (experiment 1) and between subjects (experiment 2). It is true that we only compared training versus no training, and we did not carry out an “incorrect training” group. However, Burnham *et al.* (2000) found that feedback (feedback versus no feedback as a between-subjects factor, total $N=48$) did not improve native-Cantonese visual-tone identification performance. Their interpretation was that, “the hypothesis that feedback would enhance performance was not upheld” (Burnham *et al.*, 2000, p. 90). Similarly, our subjects did not improve across blocks of testing (see Fig. 5). The failure to find that feedback improves visible tone identification contrasts sharply with the learning that occurs when speechreading segmental differences such as place and manner of articulation (Gesi *et al.*, 1992; Massaro *et al.*, 1993). These studies together show that, without the appropriate strategy of where to look and what information to use, subjects do not learn tone identification with feedback, although they can learn to speechread visible properties of speech articulation such as manner and place of articulation.

To provide an idea about the effectiveness of the visual information, consider a study that provided auditory training to American listeners in the identification of Mandarin tones (Wang *et al.*, 1999). The training was eight sessions over 2 weeks, and there was an overall improvement around 21% (there was probably no significant ceiling effect, as only one out of the eight participants scored 100% on the posttest). Our visual training was one session of no more than 45 min, and there was an overall improvement in the following identification task of around 13%. The relative improvements for the most visually distinctive tones (tone 1 increased about

40% and tone 3 about 13%) in native visual identification were actually comparable to improvements in trained non-native auditory identification (tone 1 increased about 15% and tone 3 about 18%).

Future research can evaluate the extent to which visual training generalizes to new speakers and words, as Wang *et al.* (1999) had done for auditory training. Other future research can examine the possible presence, nature, and usefulness of neck and/or head movements for other languages. These patterns can be compared to the Mandarin visual tonal information to determine if there is any consistent pattern across languages among high/mid/low level tones, tones with a rising/falling/changing contour, tone at specific pitch ranges, etc.

Potential applications/implications. Learning a tonal language like Mandarin by hard-of-hearing individuals may be facilitated by having the speaker pronouncing distinctively as well as teaching the visual information. Massaro and Light (2004) trained seven students with hearing loss to discriminate minimal pairs of words bimodally (auditorily and visually). The students were also trained to produce various speech segments by using visual information about how the inside oral articulators work during speech production. The articulators were displayed from different vantage points so that the subtleties of articulation could be optimally visualized. The speech was also slowed down to emphasize and elongate the target phonemes, allowing for clearer understanding of how the target segment is produced in isolation or with other segments. Intelligibility ratings of the posttest productions were significantly higher than pretest productions, indicating significant learning. An analogous training regimen could be instantiated for the perception and production of lexical tone.

Visual tonal information may also help individuals learn tones in a new language. Massaro and Light (2003) investigated the effectiveness of Baldi (a synthetic talking head) for teaching non-native phonetic contrasts, by comparing instruction illustrating the internal articulatory processes of the oral cavity versus instruction providing just the normal view of the tutor’s face. Eleven Japanese speakers of English as a second language were bimodally trained under both instruction methods to identify and produce the American English /t/ and /l/ in a within-subject design. Both the perception and production of words by the Japanese trainees generally improved from one day of training to the next. The first language speakers of a non-tonal language like English have difficulty perceiving lexical tones at the initial stages of learning, and extra visual information may improve not just tonal perception but also tonal production, as relevant head and neck movements may help the learner more easily increase awareness and monitor their own speech production processes.

Visual information adds to the richness of our multisensory experience, and we envision this line of inquiry to have the potential of training tone perception and production, pro-

moting prosodic/pitch awareness in speech/music, and adding realism and usefulness to synthetic talking heads.

V. CONCLUSION

When unexpected findings occur in science, they may help refine theories or even create paradigms. For example, the discovery of the existence and influence of visible segmental speech helped to inspire a field of research in audiovisual speech and related technology. In an analogous way, it is possible that the realization in the existence and nature of visible tonal information may eventually help further refine theories of speech perception and production. Starting with the original studies by researchers such as Burnham (Burnham *et al.*, 2000; 2001) and Mixdorff (Mixdorff *et al.*, 2005a; b, 2006), there is gradually accumulating evidence suggesting that visual information can be useful in the identification (and discrimination) of lexical tones. We encourage further research on the relationship between changes in acoustic-physical dimensions (frequency, amplitude, and duration) and visible movements of the neck, head, and mouth (for audiovisual integration of prosody, see Srinivasan and Massaro, 2003). It would be interesting to determine if extended training can improve the performance even more. Other possible future research may examine tone production and visual perception of other tonal languages, and how visual and auditory information for tones are integrated.

Tone languages are estimated to be spoken by over half of the world's population (Connell, 2000; Ladefoged, 2001), and more than 60 million Chinese are deaf (Jiang *et al.*, 1999). Mandarin (and comparative) psycholinguistics can be a useful tool to study perception, cognition, and memory. Psycholinguistic theories and findings based on just one language may be limited in the same manner as anthropological theories and findings based on just one culture. We hope this direction of inquiry will contribute to speech science and technology, while simultaneously expanding our understanding of the mind, as well as increasing our appreciation of a language rooted in a culture still thriving after more than 5000 years.

ACKNOWLEDGMENTS

The research and writing of this paper was supported in part by grants from the National Science Foundation (NSF CHALLENGE Grant No. CDA-9726363 and NSF Grant No. BCS-9905176), a grant from the Public Health Service (Grant No. PHS R01 DC00236), cooperative grants from the Intel Corporation and the University of California Digital Media Program (D97-04), the Eugene Cota-Robles Fellowship from the University of California, Santa Cruz, and the Federico and Rena Perlino Scholarship Award. The authors thank editors and reviewers for their valuable time and comments and Michael M. Cohen for providing expert technical assistance.

APPENDIX A: THE 40 MANDARIN WORDS (SELECTED FROM LIU AND SAMUEL, 2004) USED AS STIMULI.

Pinyin/Bpmf	IPA Transcription	Tone 1	Tone 2	Tone 3	Tone 4
Fang / 方	fɑŋ	方 / 方	房 / 房	访 / 访	放 / 放
Fu / 夫	fʊ	夫 / 夫	服 / 服	府 / 府	富 / 富
Pi / 匹	pʰi	批 / 批	皮 / 皮	匹 / 匹	僻 / 僻
Pu / 扑	pʰu	扑 / 扑	菩 / 菩	谱 / 谱	瀑 / 瀑
Qi / 起	tɕʰi	妻 / 妻	骑 / 骑	起 / 起	气 / 气
Tan / 谈	tʰan	贪 / 贪	谈 / 谈	坦 / 坦	探 / 探
Tuo / 脱	tʰuɔ	脱 / 脱	驼 / 驼	妥 / 妥	唾 / 唾
Yao / 妖	jəu	妖 / 妖	摇 / 摇	咬 / 咬	耀 / 耀
Ye / 夜	jɛ	噎 / 噎	爷 / 爷	也 / 也	夜 / 夜
You / 优	jəu	优 / 优	游 / 游	有 / 有	又 / 又

The 40 Mandarin characters or words used as stimuli in the study (selected from Liu & Samuel, manuscript). Participants from Mainland China read abbreviated (simplified) forms (left), and participants from Taiwan read traditional forms (right). The words were spoken under two conditions: Neutral (a character or word by itself) or the context of "下一个字是 ____" (i.e., "the next word is ____").

APPENDIX B: AN EXAMPLE OF A RESPONSE SHEET (TRADITIONAL CHARACTERS). PARTICIPANTS FROM MAINLAND CHINA SAW SIMPLIFIED WORDS, AND PARTICIPANTS FROM TAIWAN SAW TRADITIONAL WORDS.

1	噎 爺 也 夜	21	脫 駝 妥 唾
2	方 房 訪 放	22	扑 菩 譜 瀑
3	妻 騎 起 氣	23	优 游 有 又
4	批 皮 匹 僻	24	貪 談 坦 探
5	妖 搖 咬 耀	25	方 房 訪 放
6	批 皮 匹 僻	26	貪 談 坦 探
7	脫 駝 妥 唾	27	脫 駝 妥 唾
8	夫 服 府 富	28	优 游 有 又
9	噎 爺 也 夜	29	批 皮 匹 僻
10	妻 騎 起 氣	30	貪 談 坦 探
11	噎 爺 也 夜	31	扑 菩 譜 瀑
12	夫 服 府 富	32	妖 搖 咬 耀
13	妻 騎 起 氣	33	脫 駝 妥 唾
14	扑 菩 譜 瀑	34	方 房 訪 放
15	夫 服 府 富	35	批 皮 匹 僻
16	貪 談 坦 探	36	噎 爺 也 夜
17	妻 騎 起 氣	37	优 游 有 又
18	优 游 有 又	38	妖 搖 咬 耀
19	夫 服 府 富	39	扑 菩 譜 瀑
20	妖 搖 咬 耀	40	方 房 訪 放

