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/, /ða/, 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

 0001-4966/2008/123(4)/2356/11/$23.00

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

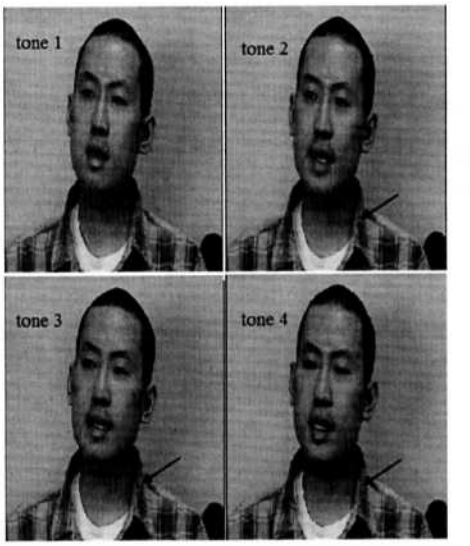

FIG. 1. Pictorial illustration of visible information for Mandarin tones at the lower parts of the neck. The bulge on the side of the neck changes across the four tones. Tone 1 has minimal movement, tones 2 and 4 have some movement, and tone 3 has the biggest bulge.

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 I. 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
	Tone 1	Tone 2	Tone 3	Tone 4
Neck	No (least) activity	Some activity	Most activity	Some (brief) activity
Chin			Females drop head/chin	
Mouth				

structed 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 I) 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 prime (d').

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

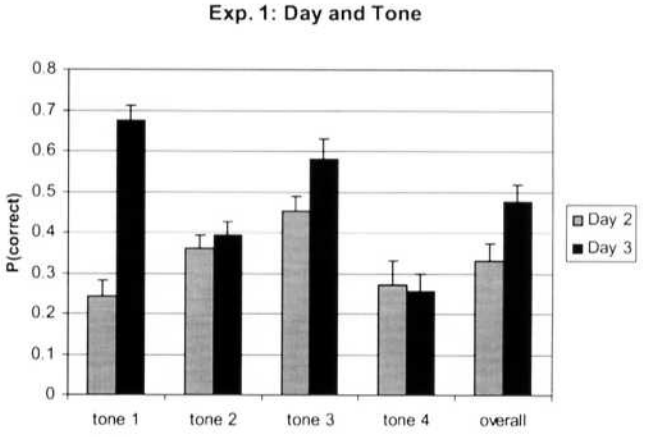

FIG. 2. Experiment 1: Tone-identification performance, plotted as the probability of correct responses (Y axis), for overall and each of the four tones on day 2 and day 3.

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 moved 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

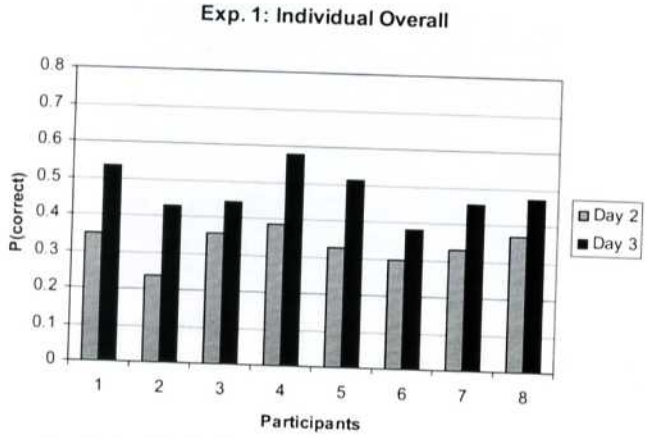

FIG. 3. Experiment 1: Individual performances, plotted as the probability of correct responses (Y axis), for the tone-identification task on day 2 and day 3.

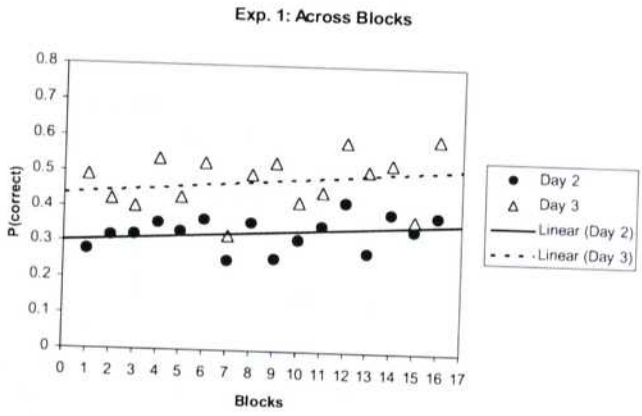

FIG. 5. Experiment 1: The overall tone-identification performance across blocks.

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 been 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 regression 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-prime (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

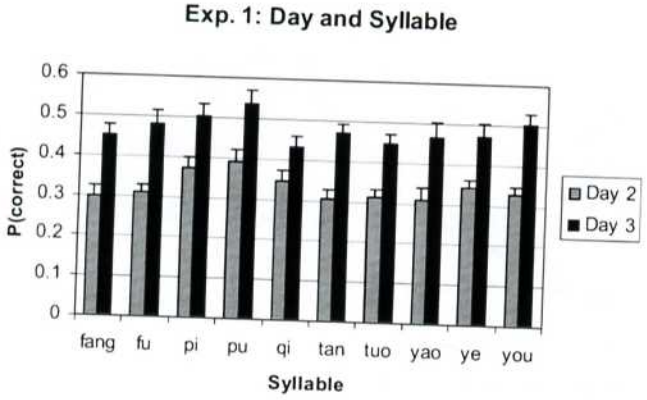

FIG. 4. Experiment 1: Tone-identification performance, plotted as the probability of correct responses (Y axis), for each of the syllables on day 2 and day 3.

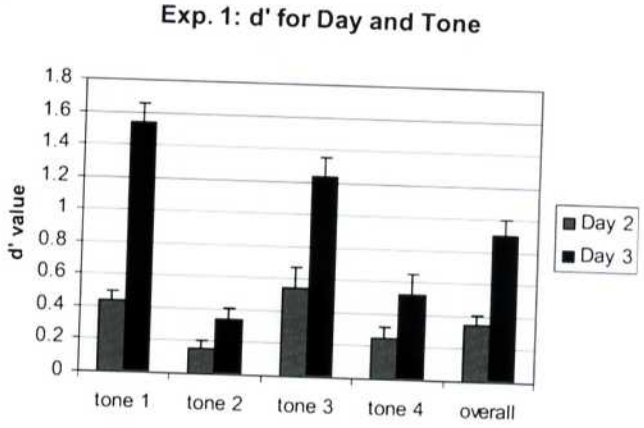

FIG. 6. Experiment 1: The d' values for overall and each of the four tones on day 2 and day 3.