Visual Cues in Mandarin Tone Perception

Hansjörg Mixdorff*, Yu Hu** and Denis Burnham ***

*Faculty of Computer Science and Media, TFH Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**University of Science and Technology, Hefei, China
yuhu@iflytek.com
***MARCS Auditory Labs, University of Western Sydney, Australia
denis.burnham@uws.edu.au

Abstract

This paper presents results concerning the exploitation of visual cues in the perception of Mandarin tones. The lower part of a female speaker's face was recorded on digital video as she uttered 25 sets of syllabic tokens covering the four different tones of Mandarin. Then in a perception study the audio sound track alone, as well an audio plus video condition were presented to native Mandarin speakers who were required to decide which tone they perceived. Audio was presented in various conditions: clear, babble-noise masked at different SNR levels, as well as devoiced and amplitude-modulated noise conditions using LPC resynthesis. In the devoiced and the clear audio conditions, there is little augmentation of audio alone due to the addition of video. However, the addition of visual information did significantly improve perception in the babble-noise masked condition, and this effect increased with decreasing SNR. This outcome suggests that the improvement in noise-masked conditions is not due to additional information in the video per se, but rather to an effect of early integration of acoustic and visual cues facilitating auditory-visual speech perception.

1. Introduction

Syllabic tones in tone languages are connected with distinct F0 patterns (rising, falling etc.). Mandarin has four different lexical tones: high (1), rising (2), low (3), and falling (4) (commonly used tone indices are given in brackets).

Research has shown that these tone contour patterns can be associated with distinct F0 patterns. While this might suggest that tone is a purely acoustic phenomenon, there are now auditory-visual studies that suggest that speakers also exploit visual cues when identifying tones [1][2]. This observation provided the impetus for the current study.

There has been so far relatively limited research with respect to the integration of acoustic and visual information in the perception of tone. A recent production study [1] suggests certain restrictions with respect to the coordination of the laryngeal and articulatory systems which might be responsible for visual cues of tones. In the associated realm of prosody, it has been shown that there is a strong correlation between head movements and F0 [4]. These correlations are continuous and seem to be used by multimodal perceivers during auditory-visual perception [5], but direct studies on the perception of these movements are yet to be conducted.

2. Video material and stimulus materials

The corpus of monosyllabic tokens compiled for the experiment contains a total of 25 syllables which were uttered by a female native speaker of Mandarin with the four different tones. The 25 syllables were chosen based on the following criteria: (1) all members in each syllable set should be real words, (2) together they provide a good coverage of Mandarin vowels, as well as articulatory trajectories (for instance, tongue movements from the back to the front, from the front to the back, etc.). A list of the syllables used is shown in Table 1.

Table 1: List of syllables used in the study, Hanyu Pinyin notation.

<table>
<thead>
<tr>
<th>bo</th>
<th>ji</th>
<th>liu</th>
<th>wei</th>
<th>yi</th>
</tr>
</thead>
<tbody>
<tr>
<td>cai</td>
<td>lang</td>
<td>lun</td>
<td>xian</td>
<td>ying</td>
</tr>
<tr>
<td>chuang</td>
<td>lei</td>
<td>luo</td>
<td>xu</td>
<td>yuan</td>
</tr>
<tr>
<td>fang</td>
<td>liang</td>
<td>mao</td>
<td>yan</td>
<td>yun</td>
</tr>
<tr>
<td>hui</td>
<td>lin</td>
<td>meng</td>
<td>yao</td>
<td>zhi</td>
</tr>
</tbody>
</table>

The tokens were randomized and recorded by a professional speaker (see Figure 1) at iFlyTek Corporation Hefei with an analog video camera and digitized in MPEG2 standard (352 x 288 pixels, 25 frames per second, sampling rate 48 kHz). Each syllable was recorded four times.

Figure 1: Section of the speaker's face that was video-taped.

In order to segment the long video sequences into chunks of individual tokens, the audio tracks were down-sampled to 16 kHz and annotated using Praat TextGrid [6]. A tool was written for converting the TextGrid into a VirtualDub [7] script which in turn was used for automatically cutting the
video as well as saving the associated soundtracks to individual wave files. The videos were cut with a window starting 400 ms before the onset of the syllable and ending 400 ms after the offset.

In order to determine the potential contribution of visual cues to tone perception two different audio degradation paradigms were adopted:

1. Reduced audio, that is, stimuli which presumably do not contain acoustic tonal information.
2. Masked audio, that is, the original audio masked by varying levels of noise.

Accordingly, the original soundtracks were subjected to the following manipulations. **Devoiced stimuli** were created by LPC analysis in Praat (default settings: Prediction order 16, window-size 25 ms, step-size 5 ms) and resynthesis using pink noise as the source signal. Pink noise was chosen rather than white noise, as the resulting speech stimuli were more similar to whispered speech and more comfortable to listen to.

**Amplitude-modulated noise stimuli** were created by extracting the Intensity Tiers of the devoiced stimuli in Praat and amplitude-modulating band limited noise ($f_{center}$=2kHz). The resulting signal still contains the temporal envelope of the devoiced stimulus, but no spectral information.

**Babble noise-masked stimuli** were created by adding babble noise at SNRs of -3, -6, -9, and -12 dB, with SNR calculated for the speech portion only, not the silent intervals before and after. The appropriate values of SNR were determined by preliminary trials, identifying the region between tone identification performances on clear-audio and complete masking.

A VirtualDub script was used for replacing the original soundtracks by clear 16 kHz, babble-noise masked, devoiced and amplitude-modulated noise versions. From each syllable, two example tokens were selected, yielding 25 syllables x 4 tones x 2 tokens = 200 stimuli of each version. For easier reference we introduce abbreviations for the types of stimuli: clean 16kHz audio, **Clean-A**, clean 16 kHz audio plus video, **Clean-AV**, devoiced audio, **DeVoiced-A**, devoiced audio plus video, **DeVoiced-AV**, amplitude-modulated noise, **AmpNoise-A**, amplitude-modulated noise plus video, **AmpNoise-AV**, babble-noise masked audio, **Noise3-A**, babble-noise masked audio plus video, **Noise3-AV**, etc.

### 3. The Perception Test

Experiments were conducted using the DMDX software [7] and employed scripts provided by Caroline Jones (MARCS, UWS) that were slightly modified. Considering the large number of stimuli and the fact that the tests were to be conducted with native Mandarin speakers, an identification task rather than a discrimination task was employed. This required participants to identify the presented stimulus by choosing one of four written syllables differing only in tone.

One set of syllables was chosen for a practice session preceding the experiment proper, and the remaining 24 syllable sets were divided into four groups. During a session each subject was presented with stimuli from four different auditory, visual or auditory-visual conditions in four consecutive blocks of trials. As can be seen in Table 1, a rolling design was employed such that the four types of stimuli presented to a particular participant in one trial set were, for instance, **Clean-A**, **Clean-AV**, **DeVoiced-A**, and **DeVoiced-AV**, with each block containing a different set of syllables, and the sequence of stimulus types varying between the four trial sets.

**Table 2: Structure of experiment trials with respect to stimulus type and syllable set.**

<table>
<thead>
<tr>
<th>Block</th>
<th>Trial set 1</th>
<th>Trial set 2</th>
<th>Trial set 3</th>
<th>Trial set 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stim.type 1</td>
<td>Stim.type 2</td>
<td>Stim.type 3</td>
<td>Stim.type 4</td>
</tr>
<tr>
<td></td>
<td>syll.group 1</td>
<td>syll.group 1</td>
<td>syll.group 1</td>
<td>syll.group 1</td>
</tr>
<tr>
<td>2</td>
<td>Stim.type 2</td>
<td>Stim.type 3</td>
<td>Stim.type 4</td>
<td>Stim.type 1</td>
</tr>
<tr>
<td></td>
<td>syll.group 4</td>
<td>syll.group 1</td>
<td>syll.group 2</td>
<td>syll.group 1</td>
</tr>
<tr>
<td>3</td>
<td>Stim.type 3</td>
<td>Stim.type 4</td>
<td>Stim.type 1</td>
<td>Stim.type 2</td>
</tr>
<tr>
<td></td>
<td>syll.group 4</td>
<td>syll.group 3</td>
<td>syll.group 1</td>
<td>syll.group 1</td>
</tr>
<tr>
<td>4</td>
<td>Stim.type 4</td>
<td>Stim.type 1</td>
<td>Stim.type 3</td>
<td>Stim.type 4</td>
</tr>
<tr>
<td></td>
<td>syll.group 2</td>
<td>syll.group 4</td>
<td>syll.group 1</td>
<td>syll.group 1</td>
</tr>
</tbody>
</table>

Within each block, tokens pertaining to a syllable set were presented consecutively, but in randomized order. Each combination of syllable and tone occurred four times, that is, each of two different examples of the syllable/tone combination was presented twice. Hence each trial set consisted of 24 syllables x 4 tones x 2 versions x 2 repetitions=384 trials and took about 35 minutes to complete.

Participants listened to the stimuli over headphones connected to a PC soundcard. Each trial started with a preparation phase of one second during which the word ‘ready’ was displayed. Then the stimulus was presented, followed by the four Mandarin words represented as Mandarin characters. The order of the four words corresponded to the numbering convenctions for Mandarin tones and was left unaltered during the experiment. Following the presentation of the four words, participants made a forced choice by hitting the appropriate number key on the keyboard.

In the practice trials (using one syllable set), feedback concerning response accuracy was given, but in the main test no feedback was given.

Participants were 20 members of staff (5 male, 15 female) at iFlyTek corporation, Hefei, China, aged 23-30. They reported having normal hearing, and four had corrected vision. Most of them were familiar with the speaker who had produced the video data. Each participant performed on two of the trial sets in Table 2.

### 4. Results

The results are discussed in two parts, first with respect to conditions employing clean and devoiced audio stimuli, and second with respect to the masked audio and amplitude-modulated noise.

**Clean and Devoiced Audio Stimuli**

In the clean and devoiced audio stimulus conditions, eight participants (3 males, 5 females), all staff at CRSLP, took part. Figure 2 displays the proportion of correct responses for stimulus types **Clean-A**, **Clean-AV**, **DeVoiced-A**, and **DeVoiced-AV**.
Clean and Devoiced Stimulus Conditions

Tone | Proportion Correct | Clean-A | Clean-AV | DeVoiced-A | DeVoiced-AV
-----|-------------------|---------|----------|------------|--------------
1    | 0.23             | 0.25    | 0.23     | 0.23       |
2    | 0.23             | 0.23    | 0.23     | 0.23       |
3    | 0.23             | 0.23    | 0.23     | 0.23       |
4    | 0.23             | 0.23    | 0.23     | 0.23       |

Figure 2: Results of perception experiment employing stimuli types Clean-A, Clean-AV, DeVoiced-A, and DeVoiced-AV. The table displays from left to right: The stimulus type, the underlying tone, the mean correct identification rates, the total number of judgments N and the associated standard deviations.

Table 3: Confusion matrix for stimulus type DeVoiced-AV.

<table>
<thead>
<tr>
<th>Intended tone</th>
<th>Perceived tone in DeVoiced-AV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>27.9</td>
</tr>
<tr>
<td>2</td>
<td>13.3</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>9.4</td>
</tr>
</tbody>
</table>

Given that the number of possible tones is four, the statistical chance level for the current experiment is at 25%. As can be seen, the identification rate on clear audio is close to 100%. The average results suggest only slight differences between audio only and the corresponding audio plus video conditions. In the case of the devoiced stimuli the picture is rather mixed. Whereas the high tone (number 1) is recognized only slightly above chance level, the low tone (3) yields around 95% correct responses. The corresponding total confusion matrix for the DeVoiced-AV condition is displayed in Table 3.

As can be seen, tones 1 and 2 are highly confusable, though the confusions are not necessarily symmetrical.

Babble Noise-Masked Audio Stimuli and Amplitude-Modulated Noise

Figure 3 (top) displays the proportion correct for babble noise-masked stimuli types, as well as amplitude-modulated noise. Due to space limitations results for individual tones are only displayed for Noise12-A and Noise12-AV – see Figure 3 (bottom). As can be seen from the top figure, recognition rates are still fairly high at an SNR of -3 dB, and subjects do not significantly benefit from seeing the video. As the SNR decreases, however, the relative gain of the “plus video” conditions increases, from 1.3% at -3dB to 15.3% at -12dB.

The gains are highly significant (p<.01) for -9 and -12dB. The relative gain for amplitude-modulated noise is even higher at 36.4%, though this augmentation occurs for an audio only condition with recognition rates close to chance. In the breakdown for individual tones in the Noise12-A, and Noise12-AV conditions, it is interesting to note that the greatest augmentation occurs for tones 3 and 4, exactly those tones that are perceived well even when de-voiced (see Figure 2). It is of interest to note that this was also the case in the amplitude-modulated noise, for this condition tones 3 and 4 still reach recognition rates of 73 and 75% respectively. Together this evidence suggests that these two tones may exhibit distinct extra-fundamental frequency cues, such as duration and intensity, which are even preserved in the devoicing condition.


The recognition results for individual tones in Noise12-A and Noise12-AV shown in Figure 3 (bottom) are much more balanced than in the DeVoiced-AV condition, as is confirmed by the confusion matrix for the Noise12-A condition shown in Table 4. They suggest that tones 2 and 3, as well as tones 3 and 4 are most prone to confusion. Interestingly the latter two are much better separated in the devoiced case (Fehler! Verweisquelle konnte nicht gefunden werden.). The results of the masking experiment confirm earlier observations [2] that presenting the video along with masked audio increases
the tonal identification rate. This gain appears to increase with increased masking of the speech sound.

Table 4: Confusion matrix for stimulus type Noise12-A. Rows indicate intended tone, columns perceived tone, figures correspond to percent.

<table>
<thead>
<tr>
<th>Intended tone</th>
<th>Perceived tone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>89.6</td>
</tr>
<tr>
<td>2</td>
<td>4.2</td>
</tr>
<tr>
<td>3</td>
<td>10.4</td>
</tr>
<tr>
<td>4</td>
<td>4.2</td>
</tr>
</tbody>
</table>

At an SNR of -12dB most syllables become very difficult to identify although the tonal contour can still be detected, as informal listening demonstrated. This is, of course, the fundamental difference between the babble noise stimuli and the devoiced stimuli – the latter are devoid of tonal information and do not significantly benefit from the addition of video. It appears as if the video does not contain any additional information in and of itself that the subjects can use. In the devoiced cases video information does not and cannot augment tonal identification, because these stimuli do not have an auditory representation of the tone gesture. On the other hand, the babble noise stimuli do have such information, and the AV conditions do show augmentation above the audio alone. This auditory information for the articulatory gesture event would match, of course, visual information for the articulatory gesture and thus augment tone identification [9][10]. On the other hand, identification of tones on the devoiced stimuli works to a certain extent due to the presence of durational as well as intensity cues. Tone 3, for instance, exhibits a characteristic two-peaked intensity contour as compared to other tones.

5. Discussion and Conclusions

The perception experiment shows that under clear audio conditions correct identification rates were close to 100% and improved only slightly when the video was presented at the same time, an obvious ceiling effect.

Identification rate on devoiced audio was reduced to a mean of 60% and did not significantly improve with the addition of video. Nevertheless, tone 3 and 4 were quite reliably identified, probably due to their specific temporal characteristics - long duration and two intensity peaks in tone 3, short duration in tone 4 - which were preserved even in the amplitude-modulated noise. Furthermore, a recent vowel study [11] showed a correlation between F0 and F1 which may suggest that a certain amount of tonal information may be captured in the spectrum envelope. There were significant benefits from presenting the video with babble noise-masked stimuli, and this gain increased with decreasing SNR.

The fact that subjects were only able to augment the missing tonal information in the noise-masked conditions suggests that the benefits of the video are due to an effect of early auditory-visual integration of the time varying and modality-independent characteristics of the tone even when the underlying syllable may no longer be identified. This, however, does not imply that visual-only cues for tones do not exist; they might, for instance, simply not be captured by the video due to limited temporal and spatial resolution. Normal hearing perceivers might simply not be trained to make use of tone cues, since in a typical communicative situation, success in tone perception depends on a range of bottom-up, and top-down cues. Studies like the one presented here are nevertheless illustrative of how auditory-visual cues operate at a basic level to augment speech perception.

Future efforts will be dedicated to a more detailed analysis of perception test results with respect to the syllable types, evaluation of acoustic features of the corpus data, evaluation of OPTOTRAK data to determine the articulatory concomitants of tone, and cross-language comparisons with the other language corpora that have been collected.

ACKNOWLEDGMENTS

This work was supported by a grant from the University of Western Sydney International Research Initiatives Scheme, as well as through the “Germany/China Joint Research Programme” of the German Academic Exchange Service and the China Scholarship Council. The assistance of Dr Caroline Jones in providing DMDX scripts is gratefully appreciated.

REFERENCES