AUDIO-VISUAL PERCEPTION OF CV SYLLABLE OF THE STANDARD CHINESE

Xiaosheng Pan[a,b] *& Jiangping Kong*[a,c]

[a]Department of Chinese Language and Literature, Peking University, Beijing, China;
[b]Comparative Linguistic Division, E-institute of Shanghai Universities, Shanghai, China;
[c]Center of Chinese Linguistics, Peking University, Beijing, China
itol_xs@pku.edu.cn; kongjp@gmail.com

ABSTRACT

The paper studies the audio-visual(AV) perception of consonant-vowel(CV) syllables of the Standard Chinese. The subjects were shown the synthetic AV congruent and incongruent stimuli in a random sequence, and were asked to respond by writing what they hear through IPA symbols. Results show that the vision of consonants can affect the perception of acoustic consonants. The influence of visual signal on the consonant identity is not identical across different place of articulation. Visual signal show greatest effect on labial place perception. Labio-dental place perception is moderately affected and other places are least affected. The visual signals with labial feature affect acoustic perception more than the visual signal with other features.

Keywords: audio-visual perception, McGurk effect, incongruent audio-visual stimulus

1. INTRODUCTION

Sumby and Pollack’s [12] experiment indicated that the visual cues contribute substantially to speech comprehension at low S/N ratios. McGurk and MacDonald [9] first studied the audiovisual perception of speech stimuli with conflicting cues and they found that acoustic perception could be affected by visual cues. McGurk effect will occur when the audio signal is incongruous with the visual signal. MacDonald, et al. [8] proposed that there were basically two modes of speech processing: an auditory-only speech mode and an AV one. McGurk effect appeared to be very robust since it persisted even when the visual and auditory components were from speakers different in gender [3] . In addition, audiovisual integration did not require detailed information about the speaker’s face [10].

Researches proved that McGurk effect will be varied between languages or cultures [4, 11]. Most researches concerning the AV perception in the Standard Chinese dubbed audio /pa/ onto visual /ka/ as stimulus for obtaining the McGurk effect. And previous researches focus on the AV perception for subjects with different language background, including the Standard Chinese [1, 2, 13].

Standard Chinese is a monosyllable language without complex onset, and syllable can only end with nasal coda.

The purpose of the paper is to systematically investigate the McGurk effect in AV incongruent condition in the AV perception of CV syllable of the Standard Chinese.

2. METHOD

2.1. Subject

A female speaker was recruited to produce the speech material at a normal speech rate. In the perception experiment, eleven subjects were recruited, all reporting normal hearing and normal or corrected-to-normal vision. All of them have phonetic training background but do not received any lipreading training.

2.2. Stimuli

The speaker’s face was recorded by SONY 3CCD. The camera and the face are at equal height and the distance between them was 1.2m. The acoustic signal was recorded using a lavalier wireless microphone(SONY ECM-44B) at a distance of 30cm from the speaker’s mouth. The speaker read words at a rate of 25 wordlist per minute. The speaker was instructed to pronounce each word with the same duration, to close mouth at the beginning and ending of every word, and to pause for one second between words. The video together with the audio signal for each word was cropped by Virtualdub [7] with the frame resolution being set at 420*300 pixels. Then a Matlab program was used to extract audio information of every word. The syllable boundary was marked for audio signal. The audio signal of CV syllables were cross-dubbed onto visual signal of CV syllables, aligned with the start marker position for vowels in the syllable.

The consonant can be in the initial or coda position of syllable in the Standard Chinese. The paper only focuses on the consonant as initial of CV syllable. Since consonant can't be pronounced separately, and vowel /a/ is compatible with most of high and level tone CV syllables, we chose vowel /a/ to generate CV syllable for experiment. The CV syllables are listed in Table 1.

Table 1: CV syllables with different consonants.

IPA	Pinyin	Word
a	a	阿
pa	ba	八
pʰa	pa	趴
ma	ma	妈
fa	fa	发
ta	da	搭
tʰa	ta	他
na	na	南
la	la	拉
ka	ga	旮
kʰa	ka	喀
xa	ha	哈
tsa	za	匝
tsʰa	ca	擦
sa	sa	仨
ts̯a	zha	扎
ts̯ʰa	cha	插
s̯a	sha	杀

As for /a/, in Standard Chinese, /a/ is often preceded by a glottal stop.

324 AV testing stimuli would be generated by cross-dubbing the syllables in Table 1. In this way, the number of the stimuli would be too large.

In [6] it is shown that in the silent condition, people cannot distinguish the difference of lip moving among CV syllable with the same articulatory place. The consonants with the same articulatory place produce the same visual perception. An audio is dubbed onto visual when the consonants of CV syllable with the same articulatory place affect acoustic perception in the same degree. CV syllables from Table 1 are grouped by the articulatory place of consonants. The result is listed in Table 2, which is the visual source of AV stimuli samples.

Table 2: CV syllables with different articulatory place of consonants.

Articulatory Place	IPA	Word
zero(Glottal)	-a	阿
Bilabial	pa	八
Labio-dental	fa	发
Dental Sibilant	tsa	匝
Alveolar	ta	搭
Retroflex	ts̯a	扎
Velar	ka	旮

126 AV stimuli were synthesized with the visual source from 7 CV syllables in Table 2 and audio source from 18 CV syllables in Table 1.

2.3. Experimental procedure

Ten samples were randomly selected from the corpus to make the subjects familiarize the task. 126 samples was adopted to test whether McGurk effect occurs when the AV was incongruent in the consonant of CV syllable. The subjects wore headphones AKG-K240 and were seated with their faces at 60 cm in front of a computer screen. Each stimulus was present once in random sequence. To every subject, the sequence was the same. The subjects informed the experiment were concerned with AV perception, but they were not told that the stimuli were synthetic. They were asked to watch and listen to the video sample in the AV condition, and to write down the IPA symbols of what they have heard. The subjects were supervised during the whole session to make sure they were focused on the screen all the time. The total experiment lasted about 1 hour, rest time was given after every 15-mintue session.

3. RESULT

To test the impact of audio signal of /Ca/ to be affected by the visual signal, each audio signal of CV syllables was cross-dubbed onto seven visual signals with different places. 77 results were obtained from 11 subjects.

The first column of the table is the consonant of the audio part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the audio part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in all 77 results.

Table 3 shows the AV perception result of CV syllables with vowel /a/ in audio signal. The result is grouped by the consonant of the audio part of stimuli.

The first column of the table is the consonant of the audio part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the audio part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in all 77 results.

Table 3: AV perception result of CV syllable with vowel /a/ in audio signal.

Consonant	Number	Percentage
pʰ	57	74%
p	44	57.1%
m	32	41.6%
f	30	39%
t	23	29.9%
n	18	23.4%
l	10	13%
tʰ	10	13%
s	9	11.7%
tʂ	8	10.4%
tʂʰ	7	9.1%
k	7	9.1%
x	7	9.1%
ts	7	9.1%
-	5	6.5%
kʰ	5	6.5%
tsʰ	3	3.9%
ʂ	3	3.9%

To test the impact of audio signal to be affected by visual signal with certain articulatory place, 18 audio signals with different consonants are dubbed onto a visual signal, which has a certain articulatory place, into 18 synthetic AV stimuli. The total number of AV perception about the specific articulatory place.

Table 4: AV perception result of audio signal with vowel /a/.

Articulatory Place	Number	Percentage
Dental Sibilant	60	30.3%
Bilabial	52	26.4%
Alveolar	41	20.7%
Retroflex	39	19.7%
Velar	31	15.7%
Labio-dental	30	15.2%
Zero(Glottal)	22	11.1%

Table 4 shows the AV perception result of CV syllables with vowel /a/ in audio signal. The result is grouped by the different articulatory places of the visual part of stimuli. The first column of the table is the articulatory places of the visual part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the articulatory place of the visual part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in total samples.

The result from The first column of the table is the consonant of the audio part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the audio part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in all 77 results.

Table 3 shows that the acoustic consonants affected by visual signal are /pʰ/, /p/, /m/, /f/, /t/, /n/, /l/, /tʰ/, /s/, /tʂ/, /tʂʰ/, /k/, /x/, /ts/, /-/, /kʰ/, /tsʰ/, /ʂ/ in descending order of percentage column respectively. If it is grouped by articulatory place of acoustic consonant, bilabials are proved to be greatly affected, which is followed by dental sibilants, and the other places are least affected.

It can be concluded from Table 4 that when labial in the visual part of stimulus, it is the easiest to affect the acoustic perception than other articulatory place feature.

4. DISCUSSION AND CONCLUSION

It can be concluded that there are two rules of consonant perception. First one is that the acoustic perception is affected by visual signal depending on the consonantal place of audio signal. The second one is the articulatory place of visual signal is also an important factor, which can affect acoustic perception.

It is found from The first column of the table is the consonant of the audio part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the audio part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in all 77 results.

Table 3 that audio signal for some consonants, such as /tsʰ/, /ʂ/, do not affected by visual signal of stimuli. It means that the articulatory place of the consonants is easily perceived from audio signal. The articulatory place information from audio signal is strong enough to make subjects ignore the articulatory place they perceive from visual signal of stimuli.

McGurk, et al. [9] also observed "combinations" such that when an auditory /kaka/ presented together with a visual /papa/ evoked the percept of /kapka/ or /papka/. The reason is visual signal /papa/ has feature of labial. In our experiment, in all cases which are perceived in accordance with the visual signal of stimuli, there are a few cases perceived as "combinations" as McGurk has found. We can get the evidence from the second and third rows of Table 4. The percentage of the cases, which are affected by visual signal with feature of labial, is only about 30%. In post-test questionnaire, all subjects admit that they have seen the action of close labial, but they do not record it because they know there isn't complex onset in Standard Chinese, although they are told that complex onset maybe exist in testing samples. Language background tends to constrain the way the subjects perceive the stimuli.

When audio part of stimulus is /fa/ in The first column of the table is the consonant of the audio part of stimulus. The second column is the number of cases which are perceived in accordance with the visual part of stimulus when the audio part is corresponding to the first column of the table. The third column is the percentage of the number of cases perceived in accordance with visual part of stimulus in all 77 results.

Table 3, 14 out of 30 cases are perceived as /sa/. About 50% of total cases are affected by visual signal. Zhang's study [5] provides the similarity of acoustic of initials in Standard Chinese. /x/ is the closest one to /f/, and the next are /s/ and /ʂ/. We know articulatory place of /s/ is the closest one to /f/ among /x/, /s/ and /ʂ/. That is why a lot of /fa/ is perceived as /sa/.

The difficulty of the experiment is that we cannot find enough subjects who have been well phonetically trained. If we choose subjects without phonetically training, and ask them to pronounce what they hear and record the IPA symbols by a phonetist, there are two problems. The first one is the subjects often cannot pronounce what they hear, the second one is that the result may be affected by the phonetist's subjective experience.

5. ACKNOWLEGMENTS

This research was funded by China National Natural Science Funds (Grant 61073085) and Funded by Ministry of Science & Technology of China (Grant 2009BAH41B00). We thank two anonymous reviewers for the comments on the earlier version of the paper.

6. REFERENCES

[1] Chen, Y., Hazan, V. 2007. Developmental factor in auditory-visual speech perception-The McGurk effect in Mandarin-Chinese and English speakers. *Proceedings of AVSP2007* Netherlands, 1-3.

[2] Chen, Y., Hazan, V. 2007. Language effects on the degree of visual influence in audiovisual speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences* Saarbrücken, 6-10.

[3] Green, K.P., et al. 1991. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception, & Psychophysics* 50, 524-536.

[4] Hayashi, Y., Sekiyama, K. 1998. Native-foreign langage effect in the McGurk effect: A test with Chinese and Japanese. *Proceedings of AVSP98* Sydney, Australia, 61-66.

[5] Jialu, Z. 2005. The distinctive features for standard Chinese(Putonghua). *Acta Acustica* 30, 506-514.

[6] Junru, W. 2008 A cognitive experiment on the lip-reading strategies of consonant phoneme in Mandarin. *The 8th Phonetic Conference of China* Beijing.

[7] Lee, A. 2001. VirtualDub home page, URL: *http://www. virtualdub. org/index*.

[8] MacDonald, J., McGurk, H. 1978. Visual influences on speech perception processes. *Perception and Psychophysics* 24, 253-257.

[9] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices, *Nature* 264, 746-748.

[10] Rosenblum, L.D., Salda a, H.M. 1996. An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology Human Perception and Performance* 22, 318-331.

[11] Sekiyama, K., Tohkura, Y. 1993. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics* 21, 427-444.

[12] Sumby, W., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America* 26, 212-215.

[13] Wang, Y., et al. 2008. Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America* 124, 1716.