Clinical Linguistics & Phonetics

The development of visual speech perception in Mandarin Chinese-speaking children

Liang Chen & Jianghua Lei

To cite this article: Liang Chen & Jianghua Lei (2017): The development of visual speech perception in Mandarin Chinese-speaking children, Clinical Linguistics & Phonetics, DOI: 10.1080/02699206.2016.1277391

To link to this article: http://dx.doi.org/10.1080/02699206.2016.1277391

Published online: 14 Feb 2017.
The development of visual speech perception in Mandarin Chinese-speaking children

Liang Chen and Jianghua Lei

Communication Sciences and Special Education, University of Georgia, Athens, GA, USA; Department of Special Education, Central China Normal University, Wuhan, China

ABSTRACT

The present study aimed to investigate the development of visual speech perception in Chinese-speaking children. Children aged 7, 13 and 16 were asked to visually identify both consonant and vowel sounds in Chinese as quickly and accurately as possible. Results revealed (1) an increase in accuracy of visual speech perception between ages 7 and 13 after which the accuracy rate either stagnates or drops; and (2) a U-shaped development pattern in speed of perception with peak performance in 13-year olds. Results also showed that across all age groups, the overall levels of accuracy rose, whereas the response times fell for simplex finals, complex finals and initials. These findings suggest that (1) visual speech perception in Chinese is a developmental process that is acquired over time and is still fine-tuned well into late adolescence; (2) factors other than cross-linguistic differences in phonological complexity and degrees of reliance on visual information are involved in development of visual speech perception.

ARTICLE HISTORY

Received 26 October 2016
Revised 16 December 2016
Accepted 24 December 2016

KEYWORDS

Chinese; cross-linguistic differences; development; visual speech perception

Introduction

During face-to-face communication, a speaker’s articulatory movements are visible. The articulatory movements of a talker’s lips, jaws and adjacent facial musculature (visual speech) are a powerful and integral component of speech perception (Calvert et al., 2004; Dodd et al., 2008; Jordan et al., 2014; McGurk & MacDonald, 1976). The visual sensory information that can be extracted via visual speech perception plays a very important role in the discrimination of sounds as well as in overall comprehension (Davies et al., 2009; Gagné, 1994; Picou et al., 2011; Reisberg et al., 1987; Tye-Murray et al., 2008). Seeing the articulating face of a talker, for example, can improve auditory speech intelligibility substantially in quiet and noisy environments (Sumby & Pollack, 1954). Auditory–visual (AV) speech recognition has consistently been shown to be more accurate than auditory-only (A) or visual-only (V) speech recognition.

In normal speech and language development, children attend to both visual speech and auditory speech and integrate the two sources of information even in infancy (Burnham & Dodd, 2004; Rosenblum et al., 1997). Visual speech perception has also been found to contribute to language acquisition from infancy (Lewkowicz & Hansen-Tift, 2012;
Teinonen et al., 2008). It has been found to be closely associated with language processing and production skills (Campbell, 1989; Massaro, 1987; Woodhouse, 2007), while deficits in visual speech perception may contribute to ongoing difficulties in speech comprehension and delays in early language development in atypical populations such as autistic children (Smith & Bennetto, 2007). Studies have shown that AV speech continues to assist older child and adult listeners in comprehension of speech in daily social situations (Smith & Bennetto, 2007). Deprivations of access to either visual speech as in the case of blind children (Mills 1987), or auditory speech as in the case of deaf and hard of hearing children, or deficit in AV speech integration as in the case of children with autism (Smith & Bennetto, 2007) can be expected to contribute to atypical speech development.

Although the importance of visual speech perception is well established, we are only just beginning to unravel its complexity, despite the impressive growth of interest in AV speech perception in children. In contrast to performance in adults, AV speech perception in children and the time course of developmental change is not well understood (Jerger, et al. 2009). In particular, the development of the ability to perceive visual speech has yet to be fully understood. The limited existing studies have revealed cross-linguistic differences in the development of visual speech perception ability. On the one hand, a developmental increase across ages has been found for English-speaking children (Hockley & Polka, 1994; Massaro, 1984; McGurk & MacDonald, 1976) and Japanese-speaking children (Sekiyama & Burnham, 2008). Hockley (1994), for example, examined the developmental process of AV speech perception using the McGurk paradigm (McGurk & MacDonald, 1976), in which a visual recording of a person saying a particular syllable is synchronized with the auditory presentation of a different syllable. Fifteen adults and 46 school aged children (range = 4;7–12;4), all native English speakers in Canada, were tested with identification of CV syllables /ba/, /va/, /θa/, /da/ and /ga/ under four conditions: Visual only, auditory-only, AV (congruent) and AV (incongruent). The children were divided into four age groups: 5, 7, 9 and 11 years. Results revealed a clear developmental pattern in speech perception in the visual only condition. It was found that as children grow older, their visual speech perception skills improve as well. These results seem to suggest a gradual developmental improvement in visual speech perception with maturation from 5 to 11 years and progressing on to adulthood.

Hnath-Chisolm et al. (1998) obtained normative data as a function of age on a test designed to assess sensory-level speech perception capacity. Forty-four English speaking children between the ages of 5 and 11 years completed the test under auditory only, visual only and AV combined conditions. Results revealed significant influences of age on performance within each condition for children below age 7 years. They suggested that visual speech perception (speech reading) abilities became stable near 7 years of age.

In a more recent study, Sekiyama and Burnham (2008), the same paradigm was used to test English and Japanese children aged 6-, 8-, 11-year olds and adults. Their results showed developmental changes in accuracy of visual speech perception for both Japanese and English up to 11 years, and this held for both response accuracy and reaction time. No difference was found between the 11-year olds and the adults in either English or Japanese. In addition, researchers observed that the performance was better in English speakers than in Japanese speakers on average (87.2% vs. 83.0%) with statistically significant difference only at 11 years.
On the other hand, such a developmental improvement in visual speech perception has not been found in French-speaking children (Tremblay et al., 2007). Tremblay et al. (2007) used the McGurk paradigm to examine the development of AV speech perception in 38 French speakers who were divided into three age groups: 5–9, 10–14 and 15–19. Results showed that the performance was similar across age groups for visual-only trials that were used to assess the development of visual speech perception, suggesting no developmental increase in visual speech perception in French.

The mixed results of prior studies suggest the possibility of language specificity with respect to the development of visual speech perception over time. The purpose of this study, therefore, was to explore the possibility by examining the development of visual speech perception in Mandarin Chinese-speaking children.

Method

Participants

Participants were 74 school-aged children. There were 28 primary school students (mean age = 7;7, SD = 0.52), 25 middle school students (mean age = 13;9, SD = 0.47) and 21 high school students (mean age = 16;7, SD = 0.47). All of them were enrolled in schools around Wuhan city of Hubei Province in central China at the time of data collection. All of the participants were native Mandarin Chinese speakers and had normal or corrected-to-normal vision. No neurological, psychological, cognitive or language impairment have been reported. They had neither received speech nor language services in the past nor were they receiving such services at the time of the data collection. None had prior experience as participants of experiments involving audio, visual or AV speech recognition. The study protocol and written informed consent form was approved by the Institutional Review Board of the University of Georgia. All children involved in the present study were given parental permission to participate, and all written informed consent forms signed by the participants (and their parents) were obtained before the test procedures took place.

Materials

Three sets of test stimuli were created with words selected from textbooks of language arts for first graders in China. There was one set each for the identification of initials (consonants), simplex finals and compound finals. Each set consisted of six target sounds, and each target sound was assessed with eight words, four that contained the target sound and four that contained a sound with an identical place of articulation (the stimuli are shown in the Appendix). The target consonants were /b/, /z/, /t/, /tʰ/, /qʰ/, /k/. The simplex finals (monophthongs) were /a/, /ɔ/, /ɤ/, /i/, /ʊ/ and /y/. Compound finals included /aɪ/, /ʊɔ/, /jɛ/, /an/, /ʊŋ/ and /aʊ/. Two important considerations have guided the section of the target sounds in each category. First, we wanted to make sure that the target sounds are found in words that the youngest Chinese children are able to recognize. Second, we wanted the sounds to be as diverse and representative of each category as possible. The task was relatively easy for the simplex finals, as they basically included the six phonemic vowels in Mandarin Chinese (Lin, 2001). We also followed Lin’s (2001) categorization of finals into
open-mouth (/aɪ/, /an/, /ɑʊ/) and close-teeth/mouth (/jɛ/, /əŋ/, /ʊɔ/) to allow for diversity in compound finals. For consonants, we ended up selecting six sounds based on the phonemic distance in a hierarchical cluster analysis (Zhao & Li, 2009). Stimulus creation started with a video and audio recording of a female speaker of standard Chinese (the common language) in a sound-attenuated room. The video recordings were made with a Sony TRV38E DV camcorder. Video clips were edited so that the start and end frames of each token showed the female speaker from the shoulders up producing one of the test words. All materials were recorded with sound, and three types of stimuli were prepared with Adobe Authorware: Visual-only (V), auditory-only (A) and AV stimuli. However, the test stimuli were administered without sound so that the participants had to rely entirely on speech reading or guessing. In other words, only the V-type stimuli were used in the present study to examine the participants’ ability to extract linguistic information from the visual speech signal.

Procedure

Participants were tested individually and completed a speech reading task that assessed both accuracy rates and response times. Participants sat in front of a computer and responded to test stimuli by pressing buttons. Instructions were given in sign language and in writing, and participants completed practice trials to demonstrate that they understood the instructions.

After seeing each word produced by the female speaker, participants viewed a response screen displaying the Yes and No buttons together with the question asking whether the word contains one of the target sounds. They made judgments as to whether the word produced contained a target sound by pressing either the Yes button or the No button. Participants were encouraged to guess if they were uncertain about a response. After each response, no feedback regarding the correctness of responses was provided, and participants were asked to click a ‘continue’ button to watch the next video stimulus. The presentation software recorded the accuracy rate and response time automatically. During the presentation, the experimenter monitored participants in order to make sure that they were watching the screen. Each participant required approximately 20 min to complete the test.

Results

The dependent variables of interest were mean accuracy rates and response times. These data are presented in Table 1 as a function of age and stimulus type. An analysis of variance (ANOVA) was performed on each dependent variable separately, with age group

<table>
<thead>
<tr>
<th>Table 1. Accuracy rate (% correct) and mean response time (RT in seconds) according to group and test stimuli types.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Primary</td>
</tr>
<tr>
<td>Middle</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>
(7-, 13- and 16-year olds) and test stimuli type (initials, simplex finals and compound finals) as between-subjects factors.

**Accuracy rate**

The mean accuracy rates are plotted in Figure 1. Here, it is clear that the visual speech perception of simplex finals was more accurate than that of initial consonants or compound finals, which in turn was more accurate than that of initial consonants. A 3 (age group) × 3 (test stimuli type) ANOVA was used to investigate differences in the development of visual speech perception accuracy. This yielded statistically significant results for the interaction of age group with test-stimulus type, \( F(4, 142) = 13.94, p < 0.001 \), as well as significant main effects for stimulus type, \( F(2, 71) = 23.32, p < 0.001 \), and significant main effects for age \( F(2, 71) = 20.43, p < 0.001 \).

Because the significant interaction of age group with test-stimulus type meant that the tree groups changed differently over time, separate univariate ANOVAs were conducted for each test stimuli type. On the initials, there were no differences between the three age groups, \( F(2, 71) = 1.46, p =0.24 \). However, significant differences were found on both simplex finals, \( F(2, 71) = 26.46, p < 0.001 \), and compound finals, \( F(2, 71) = 24.55, p < 0.001 \). Post-hoc Scheffé’s tests showed significant differences on the accurate visual speech perception of compound finals between the 7-year olds and the two older age groups (\( p < 0.001 \) for both comparisons) but no differences between the two older age groups. On simplex finals, however, the patterns were a little different such that gaps between the 7-year olds and the two older age groups were still significant (\( p < 0.001 \) for both comparisons), whereas the gap between the 13-and the 16-year olds also approached significance (\( p < 0.001 \)).

**Response time**

The group means for response time are plotted in Figure 2. A separate 3 (age group) × 3 (test-stimulus type) ANOVA was used to investigate developmental differences in the time lapse prior to responding to the items of the visual speech perception task. Results revealed a significant main effect of age, \( F(2, 71) = 25.11, p < 0.001 \). Scheffé post-hoc
tests revealed that the 7-year olds exhibited significant longer response times ($p < 0.001$) than the 13- or the 16-year olds. Furthermore, the 16-year olds showed longer response times than the 13-year olds ($p < 0.001$).

Results also indicated a significant main effect of test stimuli type, $F(2, 71) = 27.35$, $p < 0.001$. Scheffé post-hoc tests showed that visual speech perception of initial consonants ($p < 0.001$) showed longer response times than both simplex finals and compound finals. No significant differences in the speed of visual speech perception were found between the simplex finals and compound finals ($p < 0.93$).

The age group by test stimuli type interaction was not significant, $F(4, 142) = 0.83$, $p = 0.51$.

**Discussion**

**Developmental patterns of visual speech perception**

This study represents the first to examine the development of visual speech perception in Chinese-speaking children. The research results in two major findings. First, for all three types of sounds, the middle school age group performed the best in both the accuracy and speed of visual speech perception, suggesting a somewhat U-shaped developmental trajectory. This pattern stands out particularly for the speed of visual speech perception. This developmental pattern is different from either the continuous, linear development in English reported in Hockley (1994) or the lack of such developmental change in French. In addition, it argues against the suggestion that visual speech perception performance becomes similar to adults ‘sometime after the child’s 6th year’ (Massaro et al., 1986; Ross et al., 2011). It is more consistent with the developmental patterns reported for English and Japanese in Sekiyama and Burnham (2008). Needless to say, it is not easy to compare the development of visual speech perception cross-linguistically. As Sekiyama and Burnham (2008) have pointed out, languages differ in the phonological systems, and thus, test stimuli and response alternatives should be carefully chosen for more meaningful cross-linguistic and developmental comparisons. However, the similar developmental patterns in the English, Japanese and Chinese, in contrast to the lack of developmental changes in French, may help us to understand the factors that may contribute to the development of visual speech perception. In particular, we suggest that more is involved in the lack or presence of developmental increase in visual speech perception than cross-
linguistic differences in phonological complexity and degrees of reliance on visual information. Specifically, compared to Japanese, English has a more complex syllable structure and phonetic inventory, and incorporate visually distinct consonant contrasts that don’t exist in Japanese. Chinese is similar to Japanese in terms of phonological complexity, but at the same time, it is a tone language which is suggested to rely more on auditory cues than visual cues from speech reading (Burnham et al., 2001; Sekiyama, 1997). According to Sekiyama and Burnham (2008), the phonological complexity and visual distinctiveness of English could provide pressure to rely more on visual information and consequently facilitate developmental increase in visual speech perception ability. However, as the similar developmental patterns in these three languages show, the expectation of Sekiyama and Burnham (2008) is not borne out. Accordingly, more is involved in the developmental increase in visual speech perception than cross-linguistic differences in phonological complexity and degrees of reliance on visual information. In their review of visual speech perception by hearing and hearing-impaired people, Woodhouse, Hickson and Dodd (2009) noted a range of factors contribute to visual speech perception performance, including intelligence, age, degree of hearing loss, gender, visual memory and possibly socio-economic status. For example, Woodhouse (2007) found that hearing girls performed better than hearing boys on a visual speech perception task (speech-reading alone) but there was no gender difference for the hearing-impaired group. Future studies investigating the possible contribution of these factors need to determine which of the factors are more significant, or the strength of their predictive power for the development of visual speech perception.

One interesting finding of the present study was that the 13-year-olds did better than the 16 years in terms of speed of visual speech perception as well as the accurate perception of simplex finals. A possible explanation may relate to their decreased reliance on the oral channel. Auer and Bernstein (2008), in a study estimating when and how words are acquired by deaf individuals, observed that their deaf participants who subjectively reported relying more on the spoken channel for acquisition also performed better on an objective measure of visual speech perception. To some extent, the younger children in the present study may have a greater demand for visual language communication and have consequently developed better exploitation of visual speech information (cf. Aparicio et al., 2012).

**Phoneme visibility effect on visual speech perception**

The second research finding was that across all age groups, the overall levels of visual speech perception accuracy rose whereas the response times fell for simplex finals, complex finals and initials. Clearly, any conclusions about the development of visual speech perception need to take into account the type of sounds being identified. One explanation for the less accurate identification of initials (consonants) in Chinese may be found in the acoustic contrast between consonants and vowels. Vowels are usually produced with much higher acoustic energy (Diehl et al., 1987), and tend to be higher in intensity, longer in duration and involve slower movement in the articulators than consonants (Kewley-Port et al., 2007). According to Ohman (1966), vowel information is distributed over a longer temporal interval, at least the entire length of a syllable and perhaps beyond. Vowels are more easily identified than consonants, and consonant
recognition is suggested to be vowel dependent (Diehl et al., 1987). Lesner et al. (1987) noted that people with high-frequency sensorineural hearing loss typically demonstrate consonant confusions but less difficulty with vowel identification.

Our finding regarding the more accurate visual speech perception for vowels is consistent with previous research in other languages on how phoneme visibility may affect speech reading performance. Essentially, the phonological ambiguities inherent in visually perceiving the (initial) consonants are greater than in the other two types of vowel sounds. For example, the phonemes /p/, /b/ and /m/ are visually identical on the lips when speech reading. To some extent, our finding regarding the effect of phoneme visibility provides some support for vowel-based, AV speech perception training, which according to Richie and Kewley-Port (2008) may allow individuals with hearing impairment to play to their strength with vowels, rather than focusing on their weakness with consonants.

In addition to the acoustic explanation of the consonant/vowel distinction observed in the present study, a related explanation resorts to the consonant/vowel distinction observed in the memory of visual speech. De Gelder and Vroomen (1994) examined native Dutch speakers’ serial recall of lipread and heard-plus-lipread lists containing vowel-varied or consonant-varied items. Among other things, results showed that lists of syllables varying in vowels were better remembered than lists consisting of syllables varying in consonants. This result about the vowel advantage, according to De Gelder and Vroonmen (1994), reveals the significance of the vowel/consonant contrast for visual speech memory.

**Limitation of the study**

A notable limitation of the present study is that we focused on speech reading of particular sound segments. In future studies, we need to investigate the development of visually perceiving lexical tone contrasts, as well as larger linguistic units such as words, sentences and discourse. Chen and Massaro (2008), for example, have shown that Chinese adults with normal hearing could be taught to use the visual information in the speaker’s neck and head movements to improve their visual tone identification. In the future, we will examine whether Chinese-speaking children may differ in tone identification abilities under visual-only conditions across ages. It should also be noted that segment-level visual speech perception performance may be highly related to word- or sentence-level visual speech perception performance. Auer (2010), for example, suggests that individuals with slightly improved ability in segment-level visual speech identification could reap large gains in word- and sentence-level recognition accuracy, and segment-level ability is crucial for understanding individual differences in visual speech perception.

Another limitation relates to the fact that the present study did not address the development of the visual influence or the integration of visual information in speech perception in general. In spite of these limitations, the present study contributes to our understanding of the complexities in the development of visual speech perception in general and such development in Chinese in particular.
Conclusion

The present study was one of the first systematic investigations of the development of visual speech perception in Chinese-speaking children. Several conclusions can be drawn from the present study. First, visual speech perception in Chinese is a developmental process that is acquired over time and is still being fine-tuned well into adolescence.

Furthermore, we found that the visibility of sounds being identified plays a significant role in the process of visual speech perception. On the one hand, it has been observed that across all age groups, the overall levels of visual speech perception accuracy rose whereas the response times fell for simplex finals, complex finals and initials. This result suggests that the development of visual speech perception needs to take the sounds to be identified into consideration. On the other hand, the perception of initial consonants differs from the perception of the two types of finals (vowels) both in accuracy and speed. This result may provide evidence for the vowel advantage in visual speech perception. This result also has some implications for visual speech perception training in children with hearing impairment. The vowel advantage in visual speech perception suggests a need to initially focus on the training of speech reading vowels for the children with hearing impairment, and gradually moving towards the training of consonants using multisensory support systems such as cued speech and kinaesthetic feedback.

Acknowledgments

We would like to thank the children and their families in the conduct of this research.

Declaration of interest

The authors report no conflict of interest. The authors alone are responsible for the content and writing of the paper.

Funding

This work was supported partially by the University of Georgia COE Faculty Summer Research Grant to the first author (grant number 13YJA740023) from the Ministry of Education of the People’s Republic of China and from Hubei Provincial Department of Education to the second author (grant number 14ZD005), as well as from the China National Social Science Foundation (grant number 15BYY069).

References


## Appendix

Test stimuli for visual speech perception (Chinese characters followed by Pinyin).

### Part I: Words for the identification of initial consonants

<table>
<thead>
<tr>
<th>Sound</th>
<th>Characters</th>
<th>Pinyin</th>
</tr>
</thead>
<tbody>
<tr>
<td>/b/</td>
<td>巴 (ba1) 马 (ma3) 布 (bu4) 伯 (bo2) 柏 (pa4) 末 (mo4) 父 (fu4) 笔 (bi3)</td>
<td></td>
</tr>
<tr>
<td>/z/</td>
<td>字 (zi4) 此 (ci3) 杂 (za2) 足 (zu2) 四 (si4) 主 (zhu3) 则 (ze1) 爪 (zhua3)</td>
<td></td>
</tr>
<tr>
<td>/tʰ/</td>
<td>他 (ta1) 大 (da4) 土 (tu3) 体 (ti3) 的 (de1) 那 (na4) 拉 (la1) 特 (te4)</td>
<td></td>
</tr>
<tr>
<td>/tʃʰ/</td>
<td>知 (zhi1) 是 (shi4) 日 (ri4) 差 (ch1) 出 (chu1) 子 (zi3) 齿 (ch1)</td>
<td></td>
</tr>
<tr>
<td>/ʨʰ/</td>
<td>且 (ji3) 七 (qi2) 去 (qu4) 具 (ju4) 喜 (xi3) 亲 (qi2) 习 (xi1) 求 (qi2)</td>
<td></td>
</tr>
<tr>
<td>/kʰ/</td>
<td>个 (ge4) 哭 (ku2) 口 (kou3) 谷 (gu1) 可 (ke1) 合 (he2) 卡 (ka2) 虎 (hu2)</td>
<td></td>
</tr>
</tbody>
</table>
Part II: Words for the identification of simplex finals

/a/ 阿(a1) 牙(ya2) 也(ye3) 打(da2) 我(wo3) 无(wu2) 巴(ba1) 雨(yu3)
/s/ 莫(mo4) 米(mi3) 伯(bo2) 佛(fo2) 五(wu1) 那(na4) 瓦(wa1) 坡(po1)
/e/ 乐(le4) 各(ge4) 母(mu3) 可(ke1) 它(ta1) 其(qi2) 和(he2) 莫(mo4)
/i/ 土(tu3) 气(qi4) 个(ge4) 日(ri4) 及(ji2) 她(ta1) 子(zi3) 我(wo3)
/o/ 体(ti2) 特(te4) 步(bu4) 坡(po2) 主(zhu3) 苦(ku2) 拉(la1) 出(chu2)
/y/ 具(ju4) 许(xu2) 足(zu2) 区(qu2) 住(zhu4) 如(ri2) 气(qi4) 鱼(yu2)

Part III: Words for the identification of complex finals

/ai/ 老(lao3) 爱(ai4) 百(bai3) 猫(mao1) 开(kai3) 看(kan4) 排(pai2) 王(wang1)
/o/ 罗(luo2) 括(kuo4) 路(lu4) 坡(po2) 国(guo2) 波(bo2) 火(huo3) 入(ri4)
/e/ 灭(mie4) 累(lei4) 铁(tie3) 节(jie2) 衣(yi1) 可(ke1) 丢(diu2) 也(ye2)
/an/ 旁(pang1) 兰(lan2) 看(kan4) 卡(ka3) 白(bai2) 安(an1) 当(dang1) 反(fan3)
/on/ 农(nong2) 羊(yang2) 中(zhong1) 空(kong1) 嗡(weng1) 康(kang1) 从(cong1) 朋(peng2)
/uo/ 干(gan1) 到(dao4) 多(duo2) 楼(lou2) 毛(mao2) 照(zhao4) 占(zhan1) 猫(mao1)