研究論文

Effects of Visual Facial Information on English Pronunciation Improvement: A Comparative Study of Young Children and University Students

YANAGITA, Nene
(Japan Immigration Office in Fukuoka)

NAKANISHI, Hiroshi
(Seinan Gakuin University)

1. Introduction

Speech perception is a multisensory process that integrates both auditory and visual information, particularly the movements of a speaker's mouth. This integration can lead to fascinating perceptual illusions, such as the McGurk effect, first reported by McGurk and MacDonald in 1976. In this effect, when we hear the sound /ba/ (produced with the lips coming together and then parting rapidly) but simultaneously see a speaker's lips forming the sound /ga/ (produced with the lips parted and the back of the tongue briefly contacting the soft palate before releasing), our brain often interprets this conflicting information as a fusion of both, typically perceiving /da/. This perceived sound is essentially an integration of the auditory and visual inputs. The McGurk effect demonstrates that our brain doesn't process auditory and visual speech information separately, but integrates them to construct our overall perception of speech, highlighting the complex nature of phonetic processing and the significant role of visual cues in speech perception.

Research on English-speaking infants has demonstrated that at 18-24 weeks of age, infants fixate significantly longer on speaker videos with mouth shapes matching the sounds /a/ or /i/ compared to those with mismatched mouth shapes (Kuhl & Meltzoff, 1982). This suggests that even newborns can recognize the correspondence between a speaker's mouth movements and speech sounds.

Around six months of age, infants begin to integrate audiovisual information more effectively. Kushnerenko et al. (2008), through EEG studies, demonstrated that when presented with the sound /ba/ alongside a speaker's mouth shape for /ga/, five-month-old infants recognize it as an integrated sound (McGurk effect). Conversely, when presented with the sound /ga/ alongside a speaker's mouth shape for /ba/, infants can detect the discrepancy due to the distinctive nature of the lip-closing motion for /ba/. This research highlights the crucial role of visual information (mouth shape) in infant speech perception (Altvater-Mackensen & Grossmann, 2016).

As infants age, their focus on the speaker's mouth increases. Lewkowicz and

Hansen-Tift (2012) revealed that while 4-month-olds fixated longer on the speaker's eyes, from 6 to 10 months, infants began to fixate longer on the mouth than the eyes. This indicates an increase in attention to the speaker's mouth during the period when humans rapidly acquire phonological knowledge of their native language. A study on Japanese-speaking infants showed that while 6-7 and 9-10-month-old Japanese infants tend to look at the speaker's eyes more than the nose or mouth, 12-13-month-old infants primarily focus on the mouth (Morisawa et al., 2013).

The importance of mouth movements extends beyond perception to imitation in infants and young children. Studies on English-speaking infants have demonstrated that 3-4-month-olds imitate sounds more frequently when audiovisual information is congruent (Legerstee, 1990). Research on Japanese-speaking infants has also shown that 6-month-olds who focus on the speaker's mouth engage in more vocal imitation (Imafuku et al., 2009). Furthermore, a study on German-speaking infants revealed that 6-month-olds who attended longer to the speaker's mouth exhibited stronger activation in the speech motor planning area (Broca's area) (Altvater-Mackensen & Grossmann, 2016).

This focus on the mouth has been linked to language acquisition. Research on English-speaking infants has reported that 6-month-olds who look more at the mouth demonstrate higher language production skills (vocabulary comprehension and production) at 24 months (Young et al., 2009). A study on Japanese-speaking infants revealed that infants who spent more time looking at the speaker's mouth at 6 months understood more vocabularies at 12 months (Imafuku & Myowa, 2016). These findings suggest that infants may enhance their speech acquisition by observing articulatory movements (Morisawa et al., 2013), potentially contributing to language development (Imafuku, 2019).

These previous studies demonstrate that infants and young children focus on the speaker's mouth shape to perceive native language sounds, and that mouth information promotes imitation and contributes to language acquisition. The present study aims to investigate whether pronunciation practice utilizing speaker's mouth movements is also effective in second language speech acquisition.

2. Purpose of This Study

This study investigates the effectiveness of visual information in the acquisition of English sounds by Japanese-speaking children (5-year-olds) and university students. The target English sounds in this research are the vowels /æ/, /ɑ/, and /ʌ/, which are known to be difficult for Japanese English learners to distinguish (Shinohara et al., 2019). Using English words (/cap/, /cop/, /cup/) containing these target vowels, we examine whether learners' English pronunciation improves more effectively when they can clearly see the model speaker's mouth movements during video-based pronunciation practice, compared to when these movements are not visible.

3. Method

3.1 Participants

A total of 18 five-year-old children (7 boys, 11 girls) and 36 university students (9 males, 27 females) participated in this study. Data that did not meet the experimenter's instructions or contained unclear audio were excluded from the analysis. The final sample consisted of 17 five-year-old children (6 boys, 11 girls) and 34 university students (8 males, 26 females). The English proficiency of the university students, as measured by TOEIC scores, ranged from 305 to 770 (mean = 539.41, standard deviation = 117.35).

3.2 Procedure

All participants completed pre- and post-tests, as well as an English pronunciation training session (which involved repeating English word sounds while watching videos). The experiments with children were conducted individually in a separate room. In contrast, university students participated in group sessions in a Computer Assisted Language Learning (CALL) classroom.

3.2.1 Pre-test and Post-test

Participants were instructed to pronounce the English words (/cap/, /cop/, /cup/), containing the target vowels (/æ/, /ɑ/, /ʌ/), after listening to a model sound. The participants' pronunciations were recorded for analysis.

3.2.2 Training Sessions

Participants were instructed to pronounce the English words while watching videos of a native English speaker (American male) producing the pronunciations. The model speaker repeated each target word (/cap/, /cop/, /cup/) 15 times. All pronunciations made by the participants were recorded. Two conditions were established (see Figure 1):

Condition A (Full-face Video): Participants could refer to the model speaker's mouth movements while pronouncing.

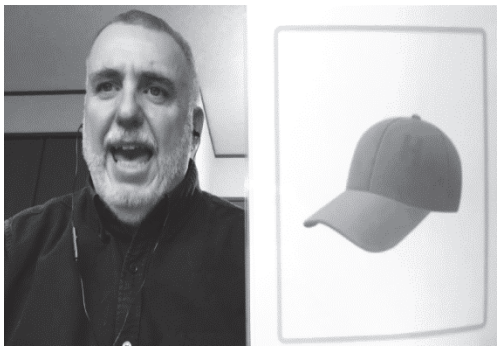
Condition B (Mosaic Video): The mouth area of the model speaker was obscured by mosaic processing, preventing participants from referring to mouth movements during pronunciation.

For the pronunciation training, 7 children and 20 university students were assigned to Condition A, while 10 children and 14 university students were assigned to Condition B. It was confirmed that there was no significant difference in English proficiency between the groups of university students ($t(32) = 0.14$, ns).

Figure 1

Video Conditions in the Training Session

Condition A: Full-face Video



Condition B: Mosaic Video

3.3 Analysis and Evaluation Method

The first (F1) and second (F2) formant values of the target vowels (/æ/, /ɑ/, /ʌ/) pronounced by the learners (children and university students) during the pre-test and post-test were measured using the acoustic analysis software Praat. To evaluate how closely the learners' vowel pronunciations approached the F1 and F2 values of the native English speaker model after pronunciation training, we used the Euclidean distance as an indicator.

The Euclidean distance is calculated on a two-dimensional plane as the difference between each participant's vowel formant frequencies (F1, F2) and the model voice's formant frequencies (F1, F2). This method is widely used in previous studies (Flege et al., 2003; Iverson & Evans, 2007).

Euclidean distance = $\sqrt{}$((F1 participant - F1 model)^2 + (F2 participant - F2 model)^2)

In this formula, F1 participant and F2 participant represent the participant's vowel formant frequencies, while F1 model and F2 model represent the model voice's formant frequencies. If the Euclidean distance in the post-test minus the Euclidean distance in the pre-test becomes shorter (i.e., a negative value), it indicates improvement in the learner's pronunciation.

In the following analysis, we conducted a two-way ANOVA using the change in Euclidean distance as the dependent variable, and age group (children vs. university students) and training condition (Condition A vs. Condition B) as independent variables.

4. Results and Discussion

Table 1 presents the mean values of the change in Euclidean distance (post-test minus pre-test) for each condition. Standard deviations are shown in parentheses.

Table 1

Change in Euclidean Distance from the Model Sound

Word	Age Group	Condition A (Full-face)	Condition B (Mosaic)
CAP	Children	125.50(215.50)	-61.61(212.60)
	Adults	-46.41(347.03)	59.97(332.06)
COP	Children	38.11(309.28)	-103.35(229.90)
	Adults	-8.32(470.99)	166.18(508.49)
CUP	Children	54.31(180.83)	128.58(460.02)
	Adults	-13.75(287.42)	-1.85 (196.17)

For CAP, although the interaction between age group and training condition was not significant, a large effect size was observed ($F(1, 51) = 2.590$, ns, $\eta p^2 = .052$). This result suggests that the effect of visual information (model speaker's mouth movements) on improving learners' English pronunciation may differ depending on the learners' age. Specifically, for children, the mosaic condition (-61.61) showed a shorter Euclidean distance than the full-face condition (125.50), whereas for university students, the full-face condition (-46.41) showed a shorter Euclidean distance than the mosaic condition (59.97). This implies that for children, pronunciation improvement is more likely when visual information (model speaker's mouth movements) is not presented, while for university students, visual information contributes to pronunciation improvement.

Similarly for COP, although the interaction between age group and training condition was not significant, a medium effect size was observed ($F(1, 51) = 1.517$, ns, $\eta p^2 = .031$). This result suggests that the effect of visual information (model speaker's mouth movements) on learners' pronunciation improvement may tend to vary by age group. Specifically, for children, the mosaic condition (-103.35) showed a shorter Euclidean distance than the full-face condition (38.11), whereas for university students, the full-face condition (-8.32) showed a shorter Euclidean distance than the mosaic condition (166.18). This suggests that for children, pronunciation improvement is more likely without visual information (model speaker's mouth movements), while for university students, visual information contributes to pronunciation improvement.

For CUP, although the interaction between age group and training condition was not significant, a trivial effect size was observed ($F(1, 51) = .130$, ns, $\eta p^2 = .003$). Both children and university students showed shorter Euclidean distances in the full-face condition. However, for children, although the full-face condition (54.31) showed a shorter Euclidean distance than the mosaic condition (128.58), both conditions showed positive values, indicating deterioration in the post-test compared to the pre-test. In contrast, for university students, the full-face condition (-35.33) showed a shorter Euclidean distance than the mosaic condition (29.84) and a negative value, indicating pronunciation improvement. This suggests that while visual information (model

speaker's mouth movements) contributes to pronunciation improvement for university students, children's pronunciation is difficult to improve regardless of the presence or absence of visual information.

These results suggest that the effect of visual information (model speaker's mouth movements) may vary depending on the task and age group. While visual information contributes to pronunciation improvement for university students, its effect is limited for children, and focusing on auditory stimuli without visual information might be more beneficial for their pronunciation improvement. Research on first language (L1) acquisition has shown that sufficient perceptual and auditory experience is necessary to understand speech from visually presented mouth movements (Desjardins et al., 1997; Mugitani et al., 2004). It is presumed that matching unfamiliar English sounds with corresponding mouth shapes is very difficult for the children in this study due to their limited perceptual and auditory experience with English.

Additionally, previous second language (L2) research has shown that adult Japanese native speakers exhibit a stronger McGurk effect for English audiovisual stimuli compared to Japanese stimuli. This is because Japanese has a simple phonological structure (e.g., five vowels), resulting in less reliance on visual information for native Japanese stimuli. However, when hearing unfamiliar English phonemes that differ from Japanese phonemic categories, listeners tend to utilize visual cues (Sekiyama & Tohkura, 1993). Furthermore, it has been noted that Chinese learners of Japanese with longer periods of stay in Japan show a stronger McGurk effect. This suggests that learners are acquiring the relationship between speech sounds and mouth movements during the process of foreign language acquisition (Sekiyama, 1997).

5. Limitation and Further Study

This study offers valuable insights into the application of visual information (model speaker's mouth movements) in pronunciation education, highlighting the need for age-appropriate instructional methods. Our findings suggest that for university students, pronunciation training incorporating visual information may be particularly effective. In contrast, for young children, it may be more beneficial to focus primarily on auditory information while gradually introducing visual cues as their developmental stage progresses.

However, it should be noted that our study did not reveal a significant interaction between age and video condition. Possible reasons for this limitation include:

1. Regardless of whether the model speaker's mouth was visible or not during pronunciation training, young children may have found it challenging to acquire three English sounds (/æ/, /ɑ/, /ʌ/) within a short period, particularly as these sounds do not require distinction in Japanese.

2. University students might have benefited from explicit instruction or feedback on their pronunciation. For example, Saito (2007) demonstrated that adult Japanese learners of English improved their pronunciation of /æ/, a sound typically challenging for Japanese speakers (Tsujimura, 1997; Ohata, 2004), not only immediately after receiving explicit phonetic instruction and feedback but also in a test conducted one week later. Incorporating such explicit instruction and feedback, particularly using facial videos that display the model speaker's mouth movements, could potentially lead to even greater improvements in learners' English pronunciation.

Additionally, while not directly related to the lack of significant interaction, it's worth noting that the training protocol (15 repetitions per word) may have been insufficient to induce substantial improvements in learners' English vowel pronunciation across all conditions.

Future research should focus on examining the role of visual information (i.e., facial cues that display the model speaker's mouth movements) in improving L2 learners' pronunciation. This investigation should consider the effects of training frequency, the role of explicit instruction and feedback, and the relationship between age and visual information processing ability. Through these studies, we aim to shed light on the audiovisual integration mechanisms underlying L2 learners' speech perception and production processes.

Acknowledgment

This study was partially supported by JSPS Grant-in-Aid for Scientific Research (C), (PI: Hiroshi Nakanishi, No.23K00786). We would like to express our sincere gratitude to the children and teachers of Samidori Kodomono Sono for their cooperation in this experiment.

References

Altvater-Mackensen, N., & Grossmann, T. (2016). The role of left inferior frontal cortex during audiovisual speech perception in infants. *NeuroImage, 133*, 14–20.

Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40*(4), 467–491.

Imafuku, M. (2019). *Akachan no kokoro wa dono yō ni sodatsu no ka: Shakaisei to kotoba no hattatsu o kagaku suru [How do babies' minds develop?: Exploring the science of social and language development]*. Minerva Shobo.

Imafuku, M., & Myowa, M. (2016). Developmental change in sensitivity to audiovisual speech congruency and its relation to language in infants. *Psychologia, 59*(4), 163–172.

Imafuku, M., Kanakogi, Y., Butler, D., & Myowa, M. (2019). Demystifying infant vocal imitation: The roles of mouth looking and speaker's gaze. *Developmental Science, 22*(6), e12825.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America, 122*(5), 2842–2854.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences, 105*(32), 11442–11445.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences, 109*(5), 1431–1436.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.

Mugitani, R., Kobayashi, T., & Hiraki, K. (2004). Audiovisual lip-voice matching on vowels by Japanese infants. *Journal of the Phonetic Society of Japan, 8*(1), 85–95.

Morisawa, N., Hakuno, Y., & Minagawa, Y. (2013). Developmental changes in infants' use of audio-visual information for speech produced by mother and stranger. *Journal of the Phonetic Society of Japan, 17*, 77–83.

Ohata, K. (2004). Phonological differences between Japanese and English: Several potentially problematic areas of pronunciation for Japanese ESL/EFL learners. *Asian EFL Journal, 6*(4) 1–19.

Saito, K. (2007). The influence of explicit phonetic instruction on pronunciation teaching in EFL settings: The case of English vowels and Japanese learners of English. *The Linguistics Journal, 3*(3), 16–40.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*, 73–80.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444.

Shinohara, Y., Han, C., & Hestvik, A. (2019). Effects of perceptual assimilation: The perception of English /æ/, /ʌ/, and /ɑ/ by Japanese speakers. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), Proceedings of the 19th International Congress of Phonetic Sciences (pp.1013–1017). Australasian Speech Science and Technology Association Inc.

Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Blackwell.

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science, 12*(5), 798–814.