PERTANIKA PROCEEDINGS

Journal homepage: http://www.pertanika.upm.edu.my/

Is Feedback from a 3D Avatar as Effective as Feedback from a Real Person?

Junko Tanaka[1]*, Enqi Wei[1,2], and Hajime Murao[1]

[1]Graduate School of Intercultural Studies, Kobe University, 1-2-1 Tsurukabuto, 657-8501 Kobe, Japan
[2]Institute of General Education, Otemon Gakuin University, Ibaraki, 567-0013 Osaka, Japan

ABSTRACT

Research on speech perception suggests that directing attention to a speaker's mouth can strengthen visual–auditory associations that support comprehension. This study investigates whether such benefits extend to online second-language (L2) learning by examining the effectiveness of feedback delivered through 3D avatars that reproduce articulatory movements. Sixteen Chinese-speaking learners of Japanese (beginner to intermediate levels) were assigned to either a 3D-avatar feedback condition or an audio-only condition within an online training environment. Participants completed a pre-test, a feedback session, and immediate and delayed post-tests targeting the acquisition of Japanese particles. The 3D avatars, created with Unity and DiDiMo, were presented as short video clips, while the audio-only group received the sound extracted from those same clips. Descriptive analyses showed that although the 3D-avatar group did not outperform the audio-only group at any testing point, they exhibited larger gains from pre-test to both post-tests. This pattern may reflect the 3D-avatar group's lower initial proficiency; nonetheless, the consistent improvement suggests that visually enriched feedback may support L2 development even when it does not lead to higher absolute performance. Further research with larger samples and methods such as eye-tracking is needed to determine whether increased attention to articulatory cues drives these effects and to clarify the potential pedagogical value of 3D-modeled facial movement in online L2 instruction.

Keywords: Avatar, feedback, L2 Japanese, mouth-movements

ARTICLE INFO

Article history:
Received: 30 December 2025
Published: 06 March 2026

DOI: https://doi.org/10.47836/pp.2.1.022

E-mail addresses:
jtanaka@kobe-u.ac.jp (Junko Tanaka)
i-onki@otemon.ac.jp (Enqi Wei)
hjmr@opal.kobe-u.ac.jp (Hajime Murao)
* Corresponding author

INTRODUCTION

With the rise of online communication tools such as Zoom, especially accelerated by COVID-19, online language learning has become a natural choice. During online communication, some participants choose to hide their faces, which is their right. However, showing one's face provides valuable paralinguistic cues, such as facial expressions, but there is more to showing one's face: it allows the interlocutor to understand the speaker better. It is known that people tend to pay attention to the mouth area of their interlocutor, especially when speech is difficult to hear due to noise, the speaker's non-native accent, or when the listener is an L2 speaker.

The present study is part of a larger study that examines the effect of negative feedback (FB) on L2 learning, particularly on difficult linguistic items such as Japanese particles, when L2 learners are able or unable to look at the face of the FB giver. Based on previous research, we hypothesise that learners who can see the speaker's face, especially the mouth area during FB, will understand the FB better and thus learn the target language more effectively than those who receive auditory (audio-only) FB. Our research question is: Does audiovisual FB that shows the mouth area lead to better L2 learning than auditory FB?

RELATED LITERATURE

Recent neuroscience research has investigated whether attention to the face of the interlocutor, particularly the mouth, enhances audiovisual speech processing (Aller et al., 2022). Chandrasekaran et al. (2009) describe speech communication as a multisensory event in which the speaker's signals align with the listener's neural processing. Birulés et al. (2020) found that even proficient L2 speakers rely on the speaker's mouth when processing speech. Similarly, Grüter et al. (2023) found that L2 proficiency modulates attention to the speaker's mouth. These studies suggest that showing the speaker's face during communication facilitates comprehension more than audio-only interactions. Thus, we decided to experimentally test whether FB delivered via 3D avatars would aid in the learning of difficult, non-linguistically salient particles that L2 learners find hard to acquire through input alone.

METHODS

Sixteen L1 Chinese speakers, most of whom were university students, with intermediate Japanese proficiency responded to a call for participation via SNS. Eight were assigned to a 3D avatar (audio-visual) FB condition and another eight to an auditory (audio-only) FB condition based on self-reported Japanese Language Proficiency Test (JLPT) levels.

The 3D avatars were used instead of human interlocutors to avoid variables such as inconsistent facial expressions in live interactions. The avatars were created using DiDiMo, loaded into Unity, and animated using Unity Face Capture to reflect real-time facial movements. We created videos for each FB (30 items in total), focusing on the Japanese particles *o*, *de*, *ni*, with *kara* as a distractor.

Participants first completed a background questionnaire and a vocabulary test. They then completed a pre-test (30 items), a learning session with FB, an immediate post-test, and a delayed post-test one month later. Eye-tracking was planned to assess attention to the avatar's mouth movements, but these data are not yet available.

RESULTS AND DISCUSSION

Table 1 presents the descriptive statistics of the results. At pre-test, the auditory FB group demonstrated higher proficiency (i.e., higher knowledge); during the FB session, the 3D avatar group had a mean score of 65.00 (*SD*=16.81), while the auditory FB group had a mean score of 73.33 (*SD*=12.97). In the immediate post-test, the 3D avatar group scored 70.42 (*SD*=12.90) and the auditory FB group scored 74.58 (*SD*=12.72). On the delayed post-test, the 3D avatar group scored 71.67 (*SD*=9.76), while the auditory FB group scored 73.33 (*SD*=11.95).

We also used Quade's nonparametric ANCOVA (alpha = .0167) due to non-normality of the data and unequal variances. The results indicated no statistically significant differences between the two groups at any of the sessions: FB, immediate post-test, or delayed post-test.

Contrary to our prediction, however, the 3D avatar group never outperformed the auditory group in any session, which could be partly explained by the fact that the 3D avatar group's proficiency level was 10.45 points lower than the auditory group's at the beginning of the experiment. Although the 3D avatar group never outperformed the auditory group, the former showed a bigger improvement in their gain between the pre-test and both the immediate and delayed post-tests. This means that the FB provided by the 3D avatar had a lasting effect, as evidenced by the greater gain between the pre-test and two post-tests.

Table 1
Descriptive statistics results

		Pre-test	**FB Session**	**Post-test**	**Delayed Post-test**	**Short-term Gain**	**Long-term Gain**
		(T1)	**(T2)**	**(T3)**	**(T4)**	**(T3-T1)**	**(T4-T1)**
3D Avatar FB	Mean	51.25	65.00	70.42	71.67	19.17	20.42
	(SD)	(9.25)	(16.81)	(12.90)	(9.76)	(3.66)	(0.51)
(n = 8)	Min	36.67	30.00	43.33	60.00	6.67	23.33
	Max	60.00	86.67	83.33	90.00	23.33	30.00
Auditory FB	Mean	61.67	73.33	74.58	73.33	12.927	11.67
	(SD)	(9.43)	(12.97)	(12.72)	(11.95)	(3.29)	(2.52)
(n = 8)	Min	46.67	60.00	53.33	50.00	6.67	3.33
	Max	73.33	100.00	93.33	86.67	20.00	13.33

The larger gain for the 3D avatar FB group is consistent with previous research. Birulés et al. (2020) found that listeners paid more attention to the speaker's mouth when the speech was difficult to understand (e.g., speech with background noise, speech by non-native speakers of a language). This means that combining visual information with auditory information could help listeners understand speech better. The FB with the 3D avatar, which provides not only auditory information but also visual information about the mouth movements of "a speaker", must have helped the L2 learners to understand the content of the FB better and thus promoted the learning of the target language items more than the auditory FB group.

To validate the above interpretation, further research with more participants assigned to groups with similar prior knowledge is needed to increase the stability and reliability of the results. In addition, more sophisticated research methods, such as eye-tracking, are needed to investigate whether such an increase in L2 learning was brought about by attention to a 3D avatar that mimics the mouth movements of a human speaker.

With eye-tracking data analyses to follow in the next phase of this research, we should be able to further investigate what in visual information really works in understanding auditory information and how: whether our participants are attending to mouth movements while listening to auditory information, as Birulés et al. (2020) found, or whether the gains were the result of multiple modalities conveying FB information.

This study is unique in that it uses a 3D avatar movie that mimics human mouth movements as FB in L2 learning. If the presence of a 3D avatar helps learners attune to what is conveyed in the FB movie, then it would justify the use of 3D avatars in on-demand or online learning courses instead of live-action movies of human teachers, including those for L2 learning.

CONCLUSION

This experimental study investigated the effect of these two types of FB on the learning of L2 Japanese particles by 16 L1 Chinese learners of L2 Japanese in two groups: one 3D avatar FB group ($n = 8$), in which the 3D avatar mimicked human mouth movements along with auditory information, and another auditory FB group ($n = 8$), in which only audio information was provided. Contrary to our expectation that the 3D avatar FB group would outperform the auditory FB group, there was no statistical difference between the two groups at any point in the experiment. However, the gains between the pre-test and the two post-tests were greater for the 3D avatar FB group.

These preliminary results suggest that the 3D avatar FB may help L2 learners retain rules about Japanese particles better than the auditory FB. This may indicate that 3D avatars, which closely mimic human mouth movements, support L2 learning.

However, it is also possible that the improved performance is simply due to the presence of a human-like head on the screen, and not necessarily due to focusing on the mouth area.

To confirm whether learners do indeed focus on the mouth area, we need further research with more participants and more sophisticated methods using eye-tracking data. If it is found that viewers do focus on the mouth area of 3D avatars, and if this helps them to tune in to the message conveyed by the avatars, then the use of 3D avatars in online or on-demand learning courses will be supported.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their insightful comments and constructive suggestions, which greatly improved the clarity and quality of our manuscript.

REFERENCES

Aller, M., Solberg Økland, H., MacGregor, L. J., Blank, H., & Davis, M. H. (2022). Differential auditory and visual phase-locking are observed during Audio-Visual benefit and silent lip-reading for speech perception. *Journal of Neuroscience, 42,* 6108-6120. https://doi.org/10.1523/JNEUROSCI.2476-21.2022

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language Cognition and Neuroscience, 35,* 1314-1325. https://doi.org/10.1080/23273798.2020.1762905

Chandrasekaran, C., Trubanova, A., Stillittano. S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLOS Computational Biology, 5*(7), Article e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Grüter, T., Kim, J., Nishizawa, H., Wang, J., Alzahrani, R., Chang, Y. T., Nguyen, H., Nuesser, M., Ohba, A., Roos, S., & Yusa, M. (2023). Language proficiency modulates listeners' selective attention to a talker's mouth: A conceptual replication of Birulés et al. (2020). *Studies in Second Language Acquisition, 45*(4),1074-1089. https://doi.org/10.1017/S0272263123000086