Interspeech 2018
2-6 September 2018, Hyderabad

Analyzing Effect of Physical Expression on English Proficiency for Multimodal Computer-Assisted Language Learning

Haoran Wu[1], *Yuya Chiba*[1], *Takashi Nose*[1], *Akinori Ito*[1]

[1]Graduate School of Engineering, Tohoku University, Japan

{wu.haoran.r5@spcom.ecei, yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

Abstract

English proficiency is important for communication in English. Computer-Assisted Language Learning (CALL) systems are introduced to provide a convenient and low-cost language learning environment. Most of the conventional speech-based CALL systems concentrate on developing verbal fluency of the learners. However, actual English communication involves not only verbal expressions but also facial expressions and gestures, which could affect the perceived proficiency. The objective of our research is to develop a CALL system that can evaluate fluency of physical expressions as well as the verbal fluency of English. However, it is not clear how physical expressions affect the overall proficiency of English. Therefore, this study investigates the relationship between the proficiency of English and the fluency of the physical expression by analyzing the dialog data of the multimodal CALL system.

Index Terms: Computer-Assisted Language Leaning (CALL), physical expressions, multimodal interaction, para-linguistic information

1. Introduction

Globalization has spread English as a global language, boosting the population of English as a second language speaker (L2) to a higher level than ever. A Computer Assisted Language Learning (CALL) system is one of the most promising technologies for L2 learners. The conventional CALL systems focus on developing the speaking, listening, and writing skills, and some of them are put to practice [1, 2].

Not only the verbal expressions but also the physical expressions including gesture and facial expression play important roles in English communication, and could affect perceived English proficiency. Therefore, the CALL system that enables to evaluate fluency of the physical expressions as well as pronunciation and intonation can be effective for development of L2 communication skill. Currently, limited studies focus on the fluency of the physical expressions of the L2 learners, and it is still unclear how the physical expressions affect the perceived proficiency.

The goal of our study is to develop a multimodal CALL system that can evaluate fluency of the physical expressions in addition to the verbal fluency. In this paper, effect of the physical expressions on the perceived English proficiency is analyzed to investigate the possibility of applying a multimodal scheme to language proficiency evaluation. Here, we focus on spoken English by Japanese learners. Several corpora of spoken English by L2 speakers have been developed, such as CSLU Foreign Accented English corpus (FAE) [3] and English Read by Japanese corpus (ERJ) [4] to analyze the characteristics of the L2 speech. However, these corpora do not include the physical expressions associated with the speech. Thus, we collect a multimodal dialog corpus of the L2 speakers at first. Then, several analyses are conducted to investigate the relationship between the fluency of the physical expressions and the perceived proficiency of English.

2. Conventional studies on language and social skill trainers

2.1. Conventional CALL systems

The CALL system has attracted attention as one of the methods that enables convenient and low-cost language learning [5–7]. Early systems only provided learning methods based on listening or drills [5], but recent systems can evaluate the pronunciation and the intonation automatically [6]. For example, comparing the speeches of native speakers and L2 learners using the signal processing is the most typical one [8, 9]. In addition, not only an acoustic aspect of the speech, but also grammatical errors in utterances were focused on [10, 11]. These methods transcribed the L2 speech by the automatic speech recognition, and aligned transcriptions to reference sentences to detect the linguistic errors.

In addition to the pronunciation and composition training, interactive CALL systems [12–14] have been also studied to develop conversation skills of L2 speakers. For example, Suzuki et al. focused on difference of duration of the switching pause between the L2 speakers and the native speakers, and developed a system to learn appropriate turn-taking [15]. Interactive CALL systems often employ anthropomorphic robots and virtual agents as dialog partners [16–18]. The systems that employ the robots are called Robot Assisted Language Learning (RALL) system, and they are expected to improve learners' motivation to study.

As mentioned above, the conventional CALL systems for pronunciation training have mainly focused on how to evaluate the fluency of the speech or how to give feedback of the evaluation results to the learner. When performing exercises of interaction, appropriate physical expressions are as important as verbal fluency. However, few studies focus on the physical expressions of the L2 speakers. To develop CALL systems for conversation exercise, it seems to be important to consider comprehensive English proficiency including non-verbal aspects.

2.2. Assessment of social skills based on multimodal information processing

Social skill trainers [19] that aim to improve performance of public speaking [20] or job interviews [21] investigate the learners' physical expressions associated with the speech. For example, Batrinca et al. reported that gaze pattern affects largely to the perception of the presentation in addition to a flow of the speech [20]. Naim et al. estimated an objective interview rating by using audio, visual, and linguistic information [21]. They reported that contribution of the physical expressions was not

10.21437/Interspeech.2018-1425

Table 1: *Example of dialog scenario (shopping)*

Speaker	Utterance
System:	Can I help you?
Learner:	I want two lemons, three peaches, and two packs of cherries, please.
System:	Is that all?
Learner:	Yes.
System:	OK. That would be nine hundred and eighty yen.
Learner:	Here you are.
System:	Thank you very much. Here's your change.

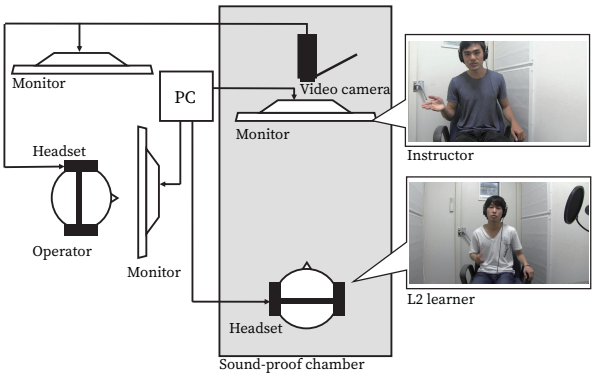


Figure 1: *Experimental environment for dialog collection.*

large because they only used limited kind of facial expressions and facial direction as the visual features, but indicated that the facial features were effective to estimate the mental state of interviewees. In addition to, several studies revealed that the multimodal information is efficient for the assessment of the public speaking [22–24].

These studies verified that the effectiveness of non-verbal information in addition to the verbal information in a domain of the social skill trainer. However, there is no study that aims to evaluate the physical expression of the L2 learner based on the relationship between the fluency of the physical expressions and the perceived language proficiency.

3. Collection of experimental data

3.1. Pilot system for constructing multimodal dialog corpus

This study assumes a situation that the learners practice the English by conducting the conversation with the system based on scenarios learned in advance [10]. This is because recognition of L2 speech is not an easy task even with recent speech recognition techniques [25]. This problem makes it difficult to achieve the CALL system that can conduct a domain-independent conversation. Although the learners cannot study how to craft sentences in our system, it is an appropriate setting for naive L2 learners because sentence construction is a high-mental load task and prevents them from focusing on practicing the pronunciation and the intonation [26]. Besides, some studies argued that repeatedly speaking the memorized sentences improves the English handling skill [27, 28]. In this paper, we employed two scenarios from an English textbook for Japanese students to construct the experimental system. The conversation scenarios in breakfast and shopping scenes were selected because they tend to occur in everyday conversation and consist of relatively easy sentences. Table 1 shows the shopping scenario. The total number of the learner's utterances contained in two scenarios is 7.

The experimental system displayed video of an instructor (a native speaker of English) on a monitor. The instructor spoke sentences of his turn with the natural gesture and facial expressions so that the learners can generate the physical expressions naturally. We employed one male native speaker of English as the instructor and recorded the video clips of the conversation. The instructor memorized two scenarios in advance and read the system's parts shown in the Table 1. The speech and motion of the instructor were recorded from the front by a video camera. An example of the recorded data is shown in Figure 1.

3.2. Experimental procedure

The experimental data were collected on the Wizard-of-Oz basis. Figure 1 shows the experimental environment. The experiments were conducted in a soundproof chamber. As shown in the figure, participants sat in front of the monitor that displayed the video of the instructor and talked with the system by following the scenarios. In the experiments, an operator out of the chamber observed the participant's responses and played back the video of the next speech when the participant's speech was completed. A video camera was placed above the monitor to record the participant's face and physical expressions. An experimental procedure is as follows:

Step 1: A participant memorizes the scenario.

Step 2: The operator confirms if the participant memorizes the scenario through a written test.

Step 3: The participant talks with the system along with the memorized scenario.

We took enough time for the participants to memorize the scenario, and started the experiments after confirming the participant memorize the scenario through the written tests, in which the participants transcribed the sentences of their turns. Thirteen undergraduate students in Graduate School of Engineering, Tohoku University (nine males and four females) participated in the experiment. The average and the standard deviation of TOEIC score is 665.3±98.4. All participants conducted two scenarios, and we collected 26 dialogs.

3.3. Evaluation of collected data

For the annotation of the data, we employed three male American English language teachers who have educational experience for the Japanese students. Hereafter, the respective evaluators are denoted as E1, E2, and E3. The evaluators annotated all dialog data with the following four criteria.

Physical expression: How natural the speaker's gestures and the facial expressions are.

Segmental: How native-like the pronunciation of the speech sounds.

Rhythm & intonation: How native-like the rhythm and the intonation of the speech sound.

Overall score: How natural the speaker's English is.

These criteria were selected by the reference to the previous study [4]. Each criterion was evaluated by 5-grade scale, one

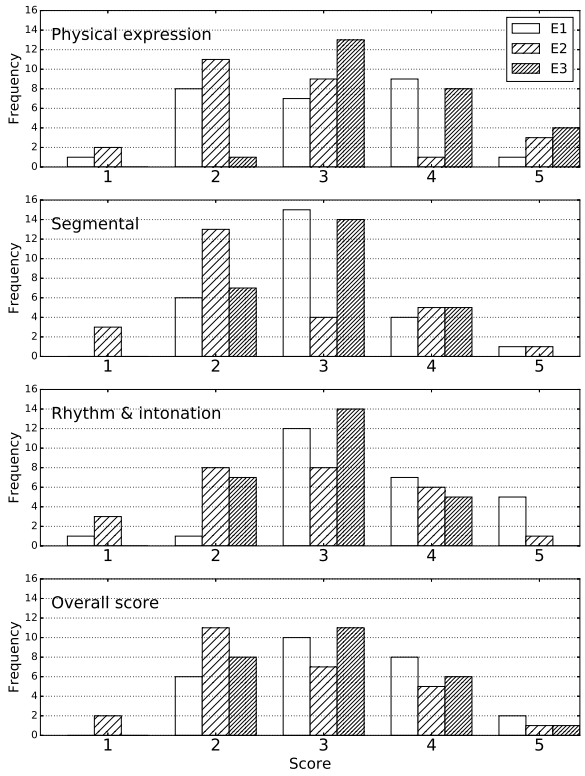


Figure 2: *Histogram of ratings given by three evaluators for each criterion.*

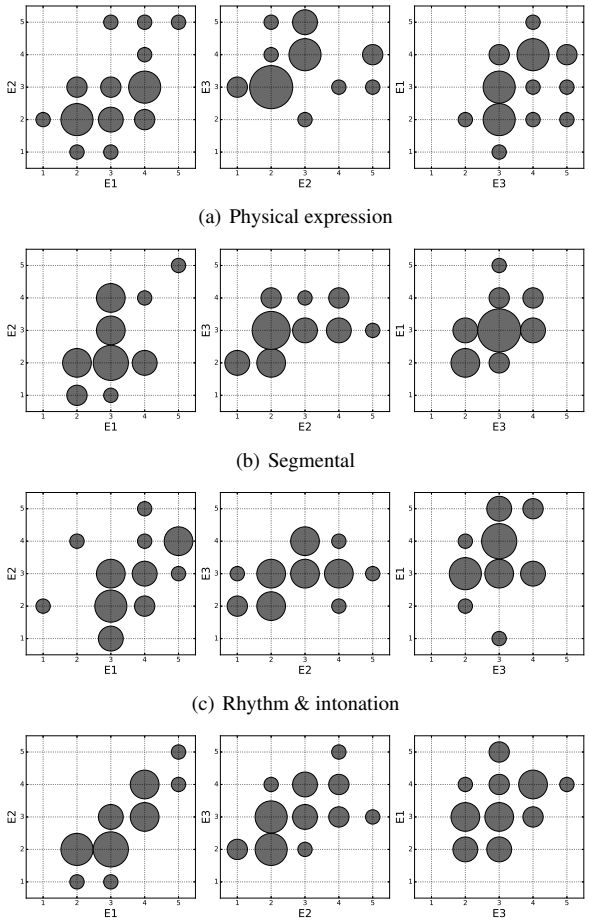
(not at all) to five (very much). We instructed the evaluators to give the scores so that the value of three was the average English level of the Japanese students the evaluator had taught.

4. Analysis of evaluation results

4.1. Concordance and correlation of ratings

First, we summarize the annotation results of the collected data. Figure 2 shows the histograms of the ratings given by the evaluators for each criterion. As shown in the figure, the ratings of two or three are the most frequent, while the ratings of one and five are relatively less frequent. Here, we calculated the correlation coefficients between the histograms of the average scores for the evaluators of the collected data and the ERJ corpus [4]. Because the ERJ corpus has the scores of the rhythm and the intonation separately, the average score was used for "Rhythm & intonation." The correlation coefficients between the corpora were 0.79 for "Segmental" and 0.52 for "Rhythm & intonation." Thus, our corpus is expected to have the similar distribution to the conventional corpus in terms of the perceived English proficiency.

Figure 3 shows the bubble charts between the ratings of the evaluators. The size of the circles represents the frequency of samples of the same ratings. As shown in the figure, the charts of the "Overall score" and "Segmental" show that the circles distribute diagonally. These results reflect that the concordance of the evaluation of the comprehensive proficiency and the pronunciation is relatively high. On the other hand, the circles of "Physical expression" and "Rhythm & intonation" are more



(a) Physical expression

(b) Segmental

(c) Rhythm & intonation

(d) Overall score

Figure 3: *Bubble charts of ratings between evaluators*

scattered than the other two criteria, which suggests that the perception of those items differ from evaluator to evaluator. We calculated the degree of the concentration [4, 29] of the ratings to compare the concordance with other dataset. The degree of the concentration is formulated as:

$$C_s = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} c_s(i,j) \tag{1}$$

$$c_s(i,j) = \frac{1}{N} \sum_{n=1}^{N} d_s^{(n)}(i,j) \times 100 \tag{2}$$

Here, N is the number of samples, K is the number of evaluators, $c_s(i,j)$ is the ratio of the samples that were rated similarly by both the evaluator i and j. For the calculation, we can consider two conditions: only considering matched samples (denoted as C_A), and allowing ± 1 gaps (denoted as C_B). Here, s denotes the condition of the calculation and $s \in \{A, B\}$. Therefore,

$$d_A^{(n)}(i,j) = \begin{cases} 1 & r_i^{(n)} = r_j^{(n)} \\ 0 & otherwise \end{cases} \tag{3}$$

$$d_B^{(n)}(i,j) = \begin{cases} 1 & |r_j^{(n)} - r_j^{(n)}| \leq 1 \\ 0 & otherwise \end{cases} \tag{4}$$

Table 2: *The degree of the concentration calculated for collected data*

	C_A	C_B
Physical expression	25.6%	79.5%
Segmental	43.6%	89.7%
Rhythm & intonation	24.4%	80.8%
Overall score	43.6%	92.3%

Table 3: *Correlation coefficients among evaluators*

	E1–E2	**E1–E3**	**E2–E3**
Physical expression	0.527	0.417	0.256
Segmental	0.501	0.466	0.494
Rhythm & intonation	0.500	0.298	0.409
Overall score	0.821	0.415	0.560

$r_{\hat{i}}^{(n)}$ is the rating of the sample n annotated by the evaluator $\hat{i}$. Table 2 shows the degree of the concentration calculated for the collected data. Minematsu et al. calculated these scores for the ERJ corpus, and obtained $C_A = 38.5\%$ and $C_B = 85.2\%$ for the segmental information, $C_A = 32.7\%$ and $C_B = 79.2\%$ for the rhythm, and $C_A = 21.3\%$; $C_B = 69.7\%$ for the intonation [4, 29]. They also calculated the scores for the FAE corpus and reported that they obtained $C_A = 46.5\%$ and $C_B = 80.3\%$ for the overall score. Therefore, the degree of the concentration of our dataset is almost the same or higher than these corpora. These results indicate that the ratings of the collected data are reliable enough to analyze. In addition, the physical expressions focused in this study showed that the degree of the concentration is comparable with "Rhythm & intonation." This result suggests that the fluency of the physical expression can be evaluated at the similar degree of reliability with the rhythm and the intonation of the speech.

Here, the degree of the concentration assumes that the evaluators annotate the data by the same criteria, and it is not appropriate for investigating the consistency of the ratings. Thus, we also calculated the correlation coefficients of the ratings. Table 3 shows the correlation coefficients between evaluators. As shown in the table, the correlations of the "Physical expression" between E1 and others are highly comparable to the "Segmental," but that between E2 and E3 shows the lowest value. These results suggest that the interpretation of the physical expressions depends on an evaluator, but is consistent between some evaluators.

4.2. Contribution of each criterion for overall score

Finally, we investigated the contribution of each criterion to the overall score by calculating the standardized partial regression. The standardized partial regression expressed the objective variable by the linear combination of the explanatory variables and formulated as:

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_0 \quad (5)$$

Here, y denotes the objective variable and x_1, x_2, and x_3 denote the standardized explanatory variables. b_1, b_2, and b_3 are standardized partial regression coefficients that can be regarded as the contribution of each explanatory variable. In this paper, we used the overall score as the objective variable, and the ratings of the other criteria as the explanatory variables to investigate the contribution of them. Table 4 shows the estimated standardized partial regression coefficients.

Table 4: *Standardized regression coefficients for overall score*

	E1	**E2**	**E3**	**All**
Physical expression	0.550	0.371	0.476	0.357
Segmental	0.102	0.578	0.513	0.430
Rhythm & intonation	0.202	0.247	0.505	0.324

From the table, we can find that "Physical expression" of E1 highly contributes to his "Overall score." In the case of E2, although the value of the standardized partial regression coefficient of the "Physical expression" is smaller than that of E1, it has a relatively large influence on "Overall score." Besides, in the case of E3, the value of the standardized partial regression coefficient is almost same among the criteria. From these results, it is suggested that "Physical expression" affects "Overall score" in all evaluators.

Then, we calculated the standardized partial regression coefficient by using all of the data. The result shows the "All" column of the table. The contribution of "Physical expression" is the second highest following the "Segmental," and almost equals to "Rhythm & intonation." This result suggests that "Physical expression" largely affects the perceived English proficiency although there are some individual differences. From these results, it is considered to be meaningful to construct the CALL system which can teach the appropriateness of the physical expressions in addition to the pronunciation and the intonation of the speech.

5. Conclusion

In this research, we investigated the effect of the fluency of the physical expressions on the perceived English proficiency to achieve the multimodal interactive CALL system. We constructed the multimodal dialog corpus recording the physical expressions of the L2 learners. Three evaluators annotated the dialog data about the proficiency of the English speech and the fluency of the physical expressions. The results of the analysis suggested that not only the acoustic aspect of the speech, but also the fluency of the physical expressions affects the overall score of the spoken English by the L2 speaker. In particular, the degree of the concentration and the coefficients of the standardized partial regression showed that the reliability and the importance of the physical expression are almost equal to those of the rhythm and the intonation. Therefore, construction of the multimodal interactive CALL system that can give advice on the physical expressions is considered to be highly meaningful.

In future work, the corpus is planned to be expanded by increasing the number of the participants and the evaluators. In addition, we will analyze the factors involved in the evaluator's judgment and examine estimation method of the ratings by using a machine learning.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP15H02720 and JP17H00823.

7. References

[1] Y. Tsubota, M. Dantsuji, and T. Kawahara, "An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors," *ReCALL*, vol. 16, no. 1, pp. 173–188, 2004.

[2] T. Kawahara, H. Wang, Y. Tsubota, and M. Dantsuji, "English and Japanese CALL systems developed at Kyoto university," in *Proc. APSIPA ASC*, 2010, pp. 804–810.

[3] T. Lander, "CSLU: Foreign accented english release 1.2," *Linguistic Data Consortium, Philadelphia*, 2007.

[4] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," in *Proc. ICA*, no. 1, 2004, pp. 557–560.

[5] U. O. Jung, "CALL: past, present and future–a bibliometric approach," *ReCALL*, vol. 17, no. 1, pp. 4–17, 2005.

[6] Y. Zhao, "Recent developments in technology and language learning: A literature review and meta-analysis," *CALICO Journal*, vol. 21, no. 1, pp. 7–27, 2003.

[7] M. Warschauer and D. Healey, "Computers and language learning: An overview," *Language Teaching*, vol. 31, no. 2, pp. 57–71, 1998.

[8] H. Meng, W.-K. Lo, A. M. Harrison, P. Lee, K.-H. Wong, W.-K. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The CUHK experience," in *Proc. APSIPA ASC*, 2010, pp. 811–820.

[9] S. Lee, H. Noh, J. Lee, K. Lee, and G. Lee, "POSTECH approaches for dialog-based English conversation tutoring," in *Proc. APSIPA ASC*, 2010, pp. 794–803.

[10] O.-p. Kweon, A. Ito, M. Suzuki, and S. Makino, "A grammatical error detection method for dialogue-based CALL system," *Journal of Natural Language Processing*, vol. 12, no. 4, pp. 137–156, 2005.

[11] T. Anzai and A. Ito, "Recognition of utterances with grammatical mistakes based on optimization of language model towards interactive CALL systems," in *Proc. APSIPA ASC*, no. 4 pages, 2012.

[12] S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," in *Proc. InSTIL/ICALL*, 2004, p. 4 pages.

[13] A. Ito, R. Tsutsui, S. Makino, and M. Suzuki, "Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system," in *Proc. INTERSPEECH*, 2008, pp. 2819–2822.

[14] P. Wik, A. Hjalmarson, and J. Brusk, "DEAL–A serious game for CALL practicing conversational skills in the trade domain," in *Proc. Speech and Language Technology in Education*, 2007, pp. 88–91.

[15] N. Suzuki, T. Nose, Y. Hiroi, and A. Ito, "Controlling switching pause using an AR agent for interactive CALL system," in *Proc. International Conference on Human-Computer Interaction*, 2014, pp. 588–593.

[16] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," in *Proc. InSTIL/ICALL*, 2004, pp. 147–150.

[17] S. Lee, H. Noh, J. Lee, K. Lee, G. G. Lee, S. Sagong, and M. Kim, "On the effectiveness of robot-assisted language learning," *ReCALL*, vol. 23, no. 1, pp. 25–58, 2011.

[18] J. Beskow, O. Engwall, B. Granstrom, and P. Wik, "Design strategies for a virtual language tutor," in *Proc. International Conference on Spoken Language Processing*, 2004, pp. 1693–1696.

[19] M. E. Hoque and R. W. Picard, "Rich nonverbal sensing technology for automated social skills training," *Computer*, vol. 47, no. 4, pp. 28–35, 2014.

[20] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero–Towards a multimodal virtual audience platform for public speaking training," in *Proc. International Workshop on Intelligent Virtual Agents*, 2013, pp. 116–128.

[21] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.

[22] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proc. INTERSPEECH*, 2005, pp. 513–516.

[23] M. I. Tanveer, J. Liu, and M. E. Hoque, "Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario," in *Proc. the 23rd ACM international conference on Multimedia*, 2015, pp. 863–866.

[24] R. Sharma, T. Guha, and G. Sharma, "Multichannel attention network for analyzing visual behavior in public speaking," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 476–484.

[25] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 1–13, 2009.

[26] P. Nation, "The role of the first language in foreign language learning," *Asian EFL Journal*, vol. 5, no. 2, pp. 1–8, 2003.

[27] H. D. Brown, *Principles of language learning and teaching*. Prentice Hall, 1994.

[28] M. A. Markell and S. L. Deno, "Effects of increasing oral reading: Generalization across reading tasks," *The Journal of Special Education*, vol. 31, no. 2, pp. 233–250, 1997.

[29] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese and Americans for CALL system development [in japanese]," *Japan Journal of Educational Technology*, vol. 27, no. 3, pp. 259–272, 2003.