The Asian Journal of Applied Linguistics
Vol. 5 No. 1, 2018, pp. 117-129

AJAL

Multimodal analysis of public speaking performance by EFL learners: Applying deep learning to understanding how successful speakers use facial movement

Miharu Fuyuno
Faculty of Design, Kyushu University

Rinko Komiya
Department of Systems Design and Informatics, Kyushu Institute of Technology

Takeshi Saitoh
Faculty of Systems Design and Informatics, Kyushu Institute of Technology

Although multimodal corpus analysis has been widely practiced in the field of applied linguistics, few studies have investigated performance of English public speaking by EFL learners. Needs for effective public speaking are fundamental in the globalizing society; however, performing public speaking in English is challenging for EFL learners, and objective analysis on factors of eye contact and speech pauses still remain few though such information is crucial in efficient teaching. This study analyses public speaking performance by EFL learners based on data from a multimodal corpus. Data were collected in an annual speech contest at a Japanese high school. Speakers presented English speeches to an audience and judges. The data consist of video and digital audio recordings of performance, as well as speech scripts and evaluation scores by contest judges. Characteristics of speakers' facial movement patterns in regard to spoken contents and the correlation between facial movements and eye movements were examined. Facial and eye movements were detected with motion tracking and the deep learning method. The results indicated that facial direction changes were not synchronized with speech pauses among highly evaluated speakers. Furthermore, the facial direction changes tended to be synchronized with content words in the spoken utterance rather than function words.

Keywords: multimodal corpus; public speaking; EFL learners; utterance; non-verbal behaviour; Japan

Introduction

Public speaking is defined as speaking in front of an audience in a limited time on a certain topic (Breakey, 2005; Fukazawa & Kobayashi, 2012). In the globalized world, opportunities to disseminate one's opinions in English on various occasions have been increasing; online meetings as well as face-to-face communication with people from various backgrounds, including both English speaking countries and ESL/EFL countries, are common situations nowadays (Fuyuno, 2015).

However, performing public speaking in English is not an easy task for EFL learners. It is known to be a source of social phobia and many people experience anxiety and fear of public speaking regardless of nationality or generation (Amir, Weber, Beard, Bomyea, & Taylor, 2008; Batrinca, Stratou, Shapiro, Morency, & Scherer, 2013;

ISSN 2308-6262
http://caes.hku.hk/ajal

Kessler, Stein, & Berglund, 1998; Pertaub, Slater, & Barker, 2001). It is particularly problematic when EFL learners need to perform public speaking in their target language (Yaikhong & Usaha, 2012). MacIntyre, Dörnyei, Clément, and Noels (1998) point out that language learners' willingness to communicate in their L2 may not correlate with their proficiency levels, and when the situations of communication become more social and public, avoidance of L2 communication increases. In fact, a public speaking setting is pervasively used as one of the key factors when measuring EFL/SL learners' willingness to communicate and their communication anxiety (see, Yashima, 2002).

In Japan, EFL learners tend to have few opportunities to conduct public speaking in English during their school education. Kawachi (2012) surveyed first-grade university students' experience of public speaking at junior high school and senior high school and found that 36.4% of them had experienced public speaking in Japanese, whereas only 17.16% had performed it in English. As they have few opportunities to practice English public speaking even in school classrooms, it is easy to imagine that their anxiety about public speaking in real life may be relatively high.

There are also difficulties teaching English public speaking. Although various textbooks have focused on English public speaking for EFL/ESL learners, many of them have included both writing and speaking components, and most textbooks tend to focus on the construction of speeches rather than their effective delivery (see, for example, Jaffe, 2012; Sugita & Caraker, 2012). In addition, the quality of a public speech is affected by nonverbal factors in delivery (Griffin, 2011) although such factors are typically not described in EFL materials, or are treated insufficiently or ambiguously.

For example, many textbooks on public speaking explain the advantages of eye contact (change of facial direction), and they advise readers to make effective eye contact with their audiences, that is, not to stare down at their notes (see, for example, Hosoi & Fallon, 1996). However, the explanation of effective eye contact often remains unspecified in practical terms. To take another instance, Jaffe (2012) suggests public speakers look in at least three directions: to the front, the left, and the right. Griffin (2011), in a similar explanation, says speakers should look at more than one person in the audience to make effective eye contact. However, there is no detail about how often and how frequently that should be done. Livingston (2010) admits that good eye contact is difficult to define since it is a learned behaviour and may differ culturally. Nevertheless, learners need this skill.

Other previous studies suggest that video-taping of students' performance and self-reflection sessions with the video data can promote learners' awareness regarding their weakness in nonverbal factors, and that peer-reviews of public speaking performance are effective in facilitating better performance including nonverbal factors (Tsang & Wong, 2002; White, 2009). Instinctively the effectiveness of these methods seems likely. However, more explicit and general indexes based on quantitative performance data would be useful for promoting effective teaching and learning. If the essential elements of effective performance were extracted as concrete data for EFL teachers and learners, they could be the basis for new classroom materials and improved teaching methods. The goal of the research reported here was to identify these essential elements of effective performance using an evidence-based approach centred on data from a multimodal corpus.

Literature Review

A multimodal corpus is a database that includes various types of data input in contrast to a traditional text-based database (Adolphs & Carter, 2013). Corpora evolved gradually with the development of data recording and storage technology. In fact, nowadays, many corpora provide voice sound data (e.g., the Michigan Corpus of Academic Spoken English (MICASE) and the Santa Barbara Corpus of Spoken American English) and video data (e.g. the British Academic Spoken English Corpus (BASE)).

In analysis of discourse, new insights have been obtained by analysing features of speech (e.g. intonation, accent, voice volume, pause, etc.) and non-linguistic elements such as nodding, gesture, eye contact, and speaker-listener position (Adolphs & Carter, 2013; Knight, 2011; McCarthy, 1998; Tsuchiya, 2013). For example, Knight (2011) analysed multimodal corpus data of a face-to-face conversation by a male supervisor and a female supervisor in a university focusing on head nod movements in the conversation. They used a 2D motion tracking method on the video data to semi-automatize the process of head movement detection. Tsuchiya (2013) analysed listenership behaviours in English conversations between British-British and British-Japanese participants. One of the Japanese participants showed a significantly larger number of head nods than the British participant. This reveals a potential cross-cultural difference in communicative behaviour. Although the Japanese participant output a relatively small number of speech tokens in English, he might have been using a different strategy to show his participation in the conversation. As these results have shown, multimodal corpora analyses can provide information about communicative strategies in human interactions.

Reflecting the fact that multimodal corpora are relatively new in the field of corpus linguistics, the application of them to EFL research is also at an early stage. Nevertheless, we can find projects related to EFL such as developments of the Padova Multimodal English Corpus (Ackerley & Coccetta, 2007) and the Multimodal Academic and Spoken Language Corpus (Fortanet-Gómez & Querol-Julián, 2010). These two projects both include video data in their corpora, along with other types of data such as tagged scripts and audio-visual materials used during academic events.

Pedagogical applications based on the concept of multimodality have also started to appear; for instance, (Coccetta, 2018 in press) discusses the importance of developing multimodal communicative competence among L2 learners and provides two examples of classroom activities that utilize multimodal transcriptions. Students were led to awareness of functions of metalanguages and contexts through the activities. Their comments showed that the activities were useful in understanding the mechanisms played by different resources, not merely texts but also other visual elements that convey meaning in communication.

Multimodal corpora studies have been conducted with speakers of various mother tongues, and have been used for some EFL research. However, few studies have focused on characteristics of English public speaking performance by EFL learners, despite the fact that multimodal corpora are an excellent resource for such analysis.

The study reported here examines nonverbal factors objectively within an EFL context by collecting and analysing audio- and video-recorded data of authentic public speaking performances by EFL learners. Previously the physical characteristics of speech rate, pause duration and speech pause patterns from the viewpoint of the relation with the contents of the utterance, and the speakers' facial movement timings have been analysed (Fuyuno, Yamashita, & Nakajima, 2016; Fuyuno, Yamashita, Saitoh, & Nakajima, 2017). These previous analyses allowed the extraction of some common

characteristics of highly evaluated public speakers regarding physical factors (e.g. appropriate frequency and movement magnitude of eye contact). Furthermore, based on these earlier results, Fuyuno, Yamada, Yamashita, and Nakajima (2016) developed a student manual for public speaking and evaluated its effectiveness in reducing students' psychological nervousness during performance by conducting experimental lessons and pre/post evaluations.

Those previous studies did not cover cross relationships between gestures and uttered contents. In this research, we aim to deepen the analysis by conducting cross analysis between facial movements and the contents of utterance, and by comparing facial movements and eye movements in terms of their effects on performance evaluation especially for eye contact with the audience.

Research Questions

Although the majority of materials for EFL public speaking training emphasizes the importance of eye contact, few of them actually advise learners on appropriate timings of eye contact. The present analysis firstly approaches this issue by targeting the speakers' facial movement patterns in regard to uttered contents. By cross-analysing the timings of facial movements and contents of utterance, the tendencies of highly evaluated speakers may be shown, and this would be useful for EFL pedagogy. In this study, we categorize the uttered contents into content words, function words and speech pauses to see whether the timing of eye contact correlates with any of these categories.

Our second aim is to examine the relationship between facial movements and eye movements. In previous studies, it has been shown that facial movement magnitudes and facial movement ranges were higher in effective public speakers compared to lower evaluated speakers, suggesting the importance of dynamic facial movements in effective eye contact (see, for example, Fuyuno et al., 2017). However, to examine speakers' eye contact, it is important to consider not only facial movement but also eye movement. Therefore, the second research question focuses on the correlation between facial movement and eye movement. The research questions thus are:

RQ1. Do the EFL public speakers' facial movement timings have a characteristic correlation with spoken contents that are categorized into content words, functions words and speech pauses? Is there any common tendency among effective speakers?

RQ2. Do EFL public speakers' facial-movement and eye-movement correlate? What can be noted for the teaching of effective eye contact in public speaking?

Data

Although there are corpora including records of university lectures such as in BASE and MICASE, and it can be said that lectures are a kind of public speaking, the present research intends to analyse performance presented in an authentic public speaking environment under stable conditions in terms of the venue size, size of audience, speech environment, evaluation process and the length and types of speech. Therefore, we created our own multimodal corpus (for a fuller account, see, Fuyuno, Yamashita, et al., 2016).

Corpus data were recorded during an annual official recitation and speech contest held by a Japanese public high school in an authentic public speaking setting with a stage, podium and audience. The contest included both recitation and speech performances. To compare data under equal conditions, only datasets from the recitation were extracted for analysis. Nine Japanese contestants, all English majors, participated in the recitation part. The participants were offered three types of recitation assignments, and each contestant selected one assignment prior to their performance[1]. After preparation and rehearsal, the contestants performed English recitations in front of the official contest judges and an audience of more than 100 people.

The team of judges evaluated each performance using an evaluation sheet. The judges were three Japanese English teachers and two NSEs teachers; all were qualified EFL teachers and had English teaching experience in Japanese secondary schools. The evaluation sheet listed the nine evaluation items shown in Table 1. Each judge scored each performance manually. Then, the scores were collected and entered into a database. As the focus of our analysis is timings of eye contact, the scores for eye contact were extracted from the database and averaged.

Table 1. Evaluation items

Item	Full Score (each judge)	Description
Pronunciation	10	pronunciation
Intonation	10	intonation
Rhythm	10	speech rhythm
Speech Delivery	10	delivery / flow / pace
Volume	10	volume of voice
Gestures	10	gestures
Eye Contact	10	eye contact
Emotion	10	emotion / energy / passion
Memorization	20	memorization of assignment

The contest performances were audio- and video-recorded using a digital sound recorder (TEAC, DR-07) and a digital video camera (JVC, GZ-R70). The digital sound recorder was set at 44.1-kHz sampling and 16-bit linear quantization, and the video camera had a resolution of 854×480 pixels (frame rate: 29.97fps, average number of frames: 5179). The devices were set on stable tripods (Figure 1). After recording, the digital data were extracted and stored in a database (Figure 2). The basic descriptions of the datasets are summarized in Table 2. The average performance duration was approximately three minutes.

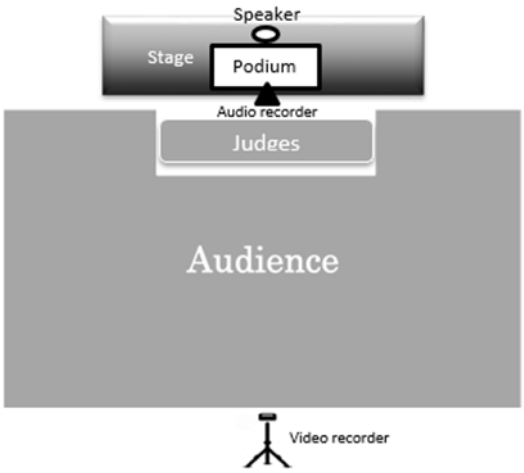

Figure 1. Arrangement for data recording

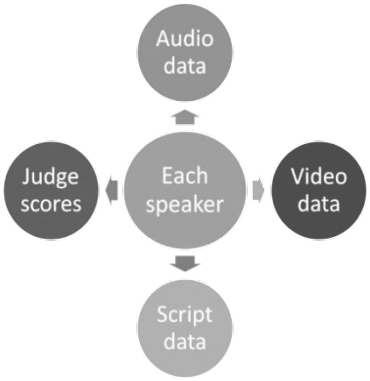

Figure 2. Datasets in our multimodal corpus

Table 2. Basic description of the data

Speaker (Anonymized)	Script Type	Average Score (Eye Contact) /100
S-01	A	68
S-02	C	60
S-03	A	60
S-04	C	88
S-05	A	74
S-06	B	92
S-07	A	78
S-08	C	64
S-09	A	70

Method

Facial movement patterns

To analyse the speakers' facial motion patterns, motion tracking was performed with a 2D CV-based original program for each speaker's video data. The program is based on the active appearance model (AAM) (Cootes, Edwards, & Taylor, 2001). This method allowed us to track pre-set feature points objectively and automatically (for details, see, Komiya, Saitoh, Fuyuno, Yamashita, & Nakajima, 2016). Forty-two feature points were set on each speaker's facial parts, i.e. jawline, eyebrows, eyes, nose and lips, as shown in Figure 3.

Figure 3. Locations of the 42 feature points for facial movement tracking

The speaker facial motions, including face roll degrees, were extracted as a series of numerical values by tracking these feature points automatically and calculating the coordinate values of the points. In this study, we focused on the speaker's facial roll magnitudes because these directly relate to eye contact movement to the sides. Firstly, coordinates of the centre of gravity of the face was obtained by posing a triangle that consists of the centres of both eyes (P_L, P_R) and the central point directly under the nose (at the top of the crevice above the lips) (P_N) (Figure 4). Facial roll degrees were defined as $roll = arctan\left(\frac{y_R - y_L}{x_{R-x_L}}\right)$. Figure 5 shows a sample motion tracking result where the speaker's face roll movement tracks to the right and left sides were obtained by the described process[2].

The centre of gravity of the left eye: PL = (xL, yL)
The centre of gravity of the right eye: PR = (xR, yR)
The centre of gravity under the nose: PN = (xN, yN)

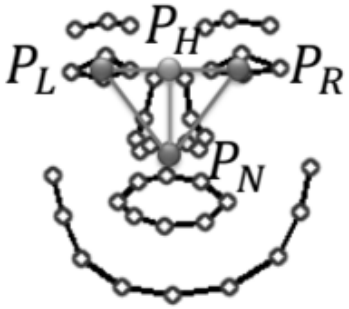

Figure 4. Calculation of the centre of gravity of the face

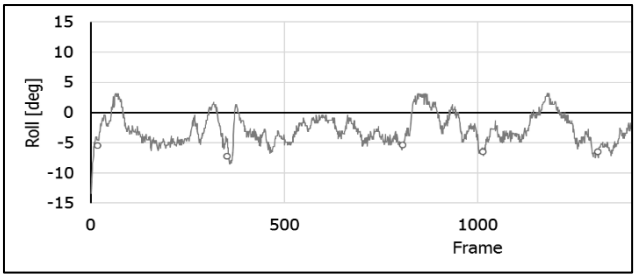

Figure 5. Sample motion tracking results of face roll movement

As we can observe in Figure 5, the result of waveform data contained minor noise peaks around the 0 degree, showing that the data includes a track of tiny movements. These noise peaks are considered to be caused mostly by physiological factors such as normal breathing by speakers and current technological limitations of feature point detection. In order to exclude these minor noise peaks, smoothing was performed on the original waveform data. The resultant smoothed peak values are shown in Figure 6. These peaks indicate the points where speakers faced rightmost/leftmost and started to change the facial direction towards reverse directions. The two illustrations of a face show sample images of facial directions in relation to specific moments during the motion tracking.

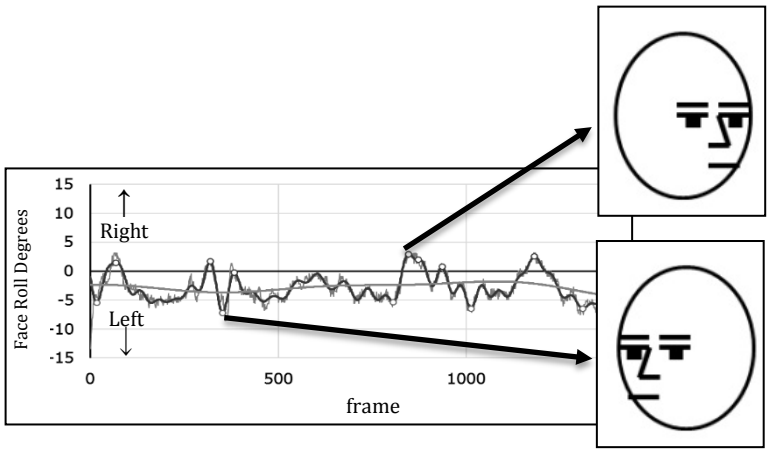

Figure 6. Sample of smoothed waveform and peak values

In order to analyse facial movements in relation to spoken utterances, we categorized the contents of utterance at the peak values of facial movement into the followings categories: content words, function words, and speech pauses. By comparing the tendencies of peak value distributions according to these categories, the characteristics between highly evaluated speakers and lower evaluated speakers are examined, with timings for eye contact direction changes in relation to uttered contents. Correlation analysis was used for evaluations.

Eye-movement patterns

The facial movement tracking shown in the previous section does not allow eye movement detection due to data processing sensibility. Eyes are much smaller than faces so tracking of pupils needs a more elaborate image recognition method. In order to track the movement of speakers' eyes accurately, deep learning for image recognition was adopted. Deep learning is a type of machine learning that utilizes multi-structured neural networks. Traditional machine learning required sets of features defined by humans; in contrast, a deep learning program learns features from an extensive learning data set (Krizhevsky, Sutskever, & Hinton, 2012).

First, the position of speakers' eyes is detected from facial motion tracking and generates images of the eyes (Figure 7). By applying the Convolutional Neural Network (CNN) approach, four feature points around the eyes and one feature point at the centre of the pupil are detected (Figure 8).

Figure 7. A sample image of an eye

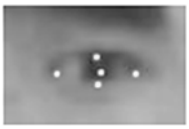

Figure 8. Locations of the five feature points for eye movement tracking

CNN requires a learning data set to learn patterns of eye movements. Data from SynthesEyes were used as input data for learning (see, Wood et al., 2015). SynthesEyes is eye image data especially developed at the University of Cambridge for eye-shape registration and gaze estimation in image recognition. It includes 11382 patterns of eye images. As a result, the 2D coordinate data of the five feature points were obtained. The movement magnitudes of eyes were then calculated (Figure 9).

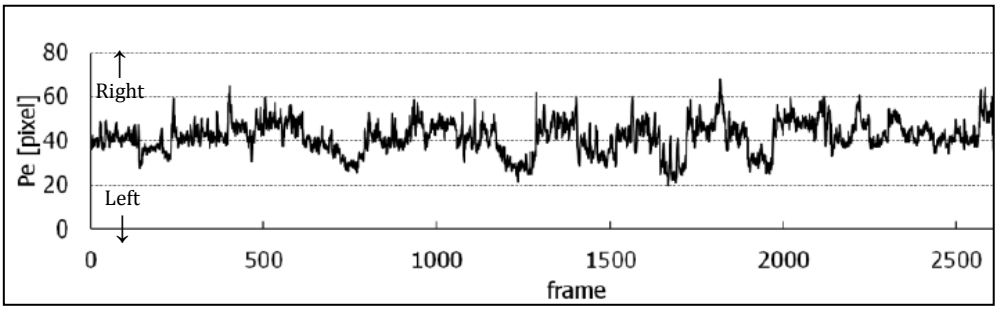

Figure 9. Sample motion tracking results of eye movement

Results and Implications

Cross analysis of facial movement patterns and spoken contents

The facial movement peak values were matched to the three categories in the speech script: content words, function words and speech pauses. All peaks in all speakers' data were labelled accordingly (see Table 3).

Table 3. Result of cross analysis between facial movement and speech content (ordered by eye contact score)

	Evaluation score for eye contact	Number of peaks	Content word (%)	Function word (%)	Speech pause (%)
R6	92	70	77.1	22.8	0
R4	88	69	57.9	36.2	5.7
R7	78	56	75	16	8.9
R5	74	83	59	32.5	8.4
R9	70	19	52.6	36.8	10.5
R1	68	84	72.6	20.2	7.1
R8	64	74	54	32.4	8.1
R2	60	33	33.3	42.4	24.2
R3	60	54	55.5	33.3	11.1

In order to analyse the relationship between facial movements and spoken contents with speaker evaluation, a correlation analysis was performed. The correlations between evaluation score and content word (%), evaluation score and function word (%), and evaluation score and speech pause (%) were analysed respectively. The coefficients of correlation were as follows; evaluation score and content word (%): r = .62, evaluation score and function word (%): r = -.38, evaluation score and speech pause (%): r = -.72.

The results show that facial direction changes tended not to synchronize with speech pauses in highly evaluated speakers' performance. This means they changed directions of eye contact while speaking, whereas lower evaluated speakers did not show the tendency. Considering the fact that there was a slight positive correlation between evaluation score and content word (%), facial direction changes may have tended to occur at points where highly evaluated speakers emphasized their contents.

It is also noticeable that the number of peaks of facial direction varies among speakers. To evaluate whether this is related to evaluation scores, a correlation analysis was performed. The coefficient of correlation was r = .31, showing low positive correlation. Since the r value was in a low range, it can be said that there were personal differences in the number of peaks in facial direction changes, however difference cannot be attributed to the performance proficiency.

The relation between facial movement and eye movement

From the facial movement tracking and eye movement tracking, the 2D coordinate numerical values from each tracking result were obtained. Correlation analysis was performed on the coordinate values. The coefficient of correlation was r = -.08, showing no correlation between the values. It was also confirmed that there was almost no correlation between evaluation scores and movement ranges of eye movements (r = .11).

A previous study showed that facial movement degrees and moving ranges were larger in highly evaluated speakers compared to lower scored speakers (Fuyuno et al.,

2017). The results of the present study and the previous study suggest that in teaching and training about eye contact in English public speaking EFL learners need to be encouraged to move their whole faces appropriately, not only focusing on eye movement.

Conclusion

This study analysed public speaking performance by Japanese EFL learners to extract useful information for efficient teaching and training. To handle multimodal data objectively and semi-automatically, methods of image recognition were applied on facial movement tracking and eye movement tracking. From the result of cross analysis with performance evaluation scores, it was noted that facial direction changes were not synchronized with speech pauses among highly evaluated speakers. Furthermore, the facial direction changes tended to be synchronized with content words in spoken utterances compared to function words.

When teaching EFL learners how to make English speeches and presentations, it has been difficult for teachers to explicitly show appropriate timings for eye contact. The results of the present study suggest that contents in speech scripts may be used as clues for planning appropriate timings for eye contact, even for students with relatively little experience in public speaking. For example, teachers can firstly encourage students to analyse their speech scripts and to decide where to put speech pauses, then make notes on timings so as to make appropriately timed eye contact by avoiding speech pauses.

Traditional textbooks and materials for EFL public speaking often depended on the authors' experience or subjective comments in providing tips for nonverbal factors such as eye contact and gestures. An evidence-based approach using the data of multimodal corpora can enhance objectivity and clarity in such materials. In our future research, more factors in public speaking, such as hand gestures and speakers' facial expressions, will be the targets of analysis with the goal of improving pedagogy.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science Kakenhi Grant numbers 15K12416 and 16H03079. The authors thank the teachers and participants at the high school for access to the research data, particularly Mr. Kazuhide Shinohara for his continued support and involvement in the research.

Notes

1. The three assignments are as follows: (A) an excerpt from 'The Principal's Address to the Graduates' by Tsuda Umeko (355 words); (B) an excerpt from Haruki Murakami's acceptance speech for the Jerusalem Award (362 words); and (C) an excerpt from 'The Little Prince' (English translation) by Antoine de Saint-Exupéry (328 words). The average number of the words was 348.33. Although the participants shared the same assignments, when they slightly changed the contents, including when they unintentionally repeated words or phrases, transcribed scripts were created for each speaker for the data analysis.
2. More elaborate discussions on the technological descriptions regarding motion tracking process and calculation for coordinate values are provided in our previous papers (Komiya et al., 2016; Komiya, Saitoh, Fuyuno, Yamashita, & Nakajima, 2017).

About the authors

Miharu Fuyuno is an assistant professor of Faculty of Design, Kyushu University, Japan. She has an MA in TESOL from the University of Nottingham, England, and a Ph.D. in linguistics from Seinan Gakuin University, Japan. Her research field includes multimodal corpora, English language teaching and technology enhanced language learning.

Rinko Komiya is a masters student at Kyushu Institute of Technology, Japan. She has a Bachelor of Information Engineering from Kyushu Institute of Technology, Japan. Her research field includes facial image processing.

Takeshi Saitoh holds the degrees of B.Eng., M.Eng. and Doctor in Engineering from Toyohashi University of Technology. He is an associate professor at Kyushu Institute of Technology. His research interests include image processing and pattern recognition.

References

Ackerley, K., & Coccetta, F. (2007). Enriching language learning through a multimedia corpus. *RECALL, 19*(3), 351-370

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. London: Routledge.

Amir, N., Weber, G., Beard, C., Bomyea, J., & Taylor, C. T. (2008). The effect of a single-session attention modification program on response to a public-speaking challenge in socially anxious individuals. *Journal of abnormal psychology, 117*(4), 860

Batrinca, L., Stratou, G., Shapiro, A., Morency, L. P., & Scherer, S. (2013). Cicero-Towards a multimodal virtual audience platform for public speaking training. In R. Aylett, B. Krenn, C. Pelachaud, & H. Shimodaira (Eds.), *International workshop on intelligent virtual agents* (pp. 116-128). Berlin: Springer.

Breakey, L. K. (2005). Fear of public speaking-the role of the SLP. *Seminars in speech and language, 26*(2), 107-117

Coccetta, F. (2018 in press). Developing university students' multimodal communicative competence: Field research into multimodal text studies in English. *System*. https://doi.org/10.1016/j.system.2018.01.004

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 23*(6), 681-685

Fortanet-Gómez, I., & Querol-Julián, M. (2010). The video corpus as a multimodal tool for teaching. In M. C. Campoy, M. C. C. Cubillo, B. Belles-Fortuno, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching corpus and discourse* (pp. 261-270). London & New York: Continuum.

Fukazawa, N., & Kobayashi, H. (2012). Components and development patterns of Japanese Shikiji speeches: Characteristics of one genre in Japanese public speaking. *Journal of Technical Japanese Education, 14*, 27-34

Fuyuno, M. (2015). Needs analysis of practical English skills in global business: Towards the development of Japanese global human resource [in Japanese]. *Studies in English Teaching and Learning in East Asia, 5*(13-27)

Fuyuno, M., Yamada, Y., Yamashita, Y., & Nakajima, Y. (2016). *Developing effective instructions to decrease Japanese speaker' s nervousness during English and Japanese public speeches: Evidence from psychological and physiological measurements* Paper presented at the 31st International Congress of Psychology 2016 (ICP2016).

Fuyuno, M., Yamashita, Y., & Nakajima, Y. (2016). Multimodal corpora of English public speaking by Asian learners: Analyses on speech rate, pause and head gesture. In F. A. Almeida, I. O. Barrera, E. Q. Toledo, & M. S. Cuervo (Eds.), *Input a word, analyse the world: Selected approaches to corpus linguistics*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Fuyuno, M., Yamashita, Y., Saitoh, T., & Nakajima, Y. (2017). Semantic structure, speech units and facial movements: Multimodal corpus analysis of English public speaking. *EPiC Series in Language and Linguistics, 1*, 447-461

Griffin, C. (2011). *Invitation for public speaking* (4th ed.). Connecticut: Cengage Learning.

Hosoi, K., & Fallon, C. R. (1996). *Introduction to English presentation*. Tokyo: Goken.

Jaffe, C. (2012). *Public speaking: Concepts and skills for a diverse society*. Connecticut: Cengage Learning.

Kawachi, T. (2012). The importance of incorporating student presentations in EFL listening courses. *Bulletin of Seikei Unviersity, 46*(4), 1-19

Kessler, R. C., Stein, M. B., & Berglund, P. (1998). Social phobia subtypes in the National Comorbidity Survey. *American Journal of Psychiatry, 155*, 613-619

Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.

Komiya, R., Saitoh, T., Fuyuno, M., Yamashita, T., & Nakajima, Y. (2016). Head pose estimation and movement analysis for speech scene. In *Proceedings of 15th IEEE/ACIS International Conference on Computer and Information Science* (pp. 1-5): IEEE.

Komiya, R., Saitoh, T., Fuyuno, M., Yamashita, T., & Nakajima, Y. (2017). Head pose estimation and motion analysis of public speaking videos. *International Journal of Software Innovation, 5*(1), 57-71

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105

Livingston, R. (2010). *Advanced public speaking: Dynamics and techniques*. Indiana, USA: Xlibris Corporation.

MacIntyre, P. D., Dörnyei, Z., Clément, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal, 82*(4), 545-562

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Pertaub, D. P., Slater, M., & Barker, C. (2001). An experiment on fear of public speaking in virtual reality. In *Studies in Health Technology and Informatics* (Vol. 81: Medicine Meets Virtual Reality, pp. 372-378).

Sugita, Y., & Caraker, R. R. (2012). *Writing for presentation in English*. Tokyo: Nan'un-do.

Tsang, W. K., & Wong, M. (2002). Conversational English: An interactive, collaborative and reflective approach. In J. C. Richards & W. Rendandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 212-224).

Tsuchiya, K. (2013). *Listenership behaviours in intercultural encounters: A time-aligned multimodal corpus analysis*. Amsterdam, Netherlands: John Benjamins.

White, E. (2009). Student perspectives of peer assessment for learning in a public speaking course. *Asian EFL Journal, 33*(1), 1-36

Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., & Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3756-3764).

Yaikhong, K., & Usaha, S. (2012). A measure of EFL public speaking class anxiety: Scale development and preliminary validation and reliability. *English Language Teaching, 5*(12), 23-35

Yashima, T. (2002). Willingness to communicate in a second language: the Japanese EFL context. *The Modern Language Journal, 86*(1), 54-66