EPiC Series in Language and Linguistics
Volume 1, 2017, Pages 447–461
CILC2016. 8th International Conference on Corpus Linguistics

Semantic Structure, Speech Units and Facial Movements: Multimodal Corpus Analysis of English Public Speaking

Miharu Fuyuno[1], Yuko Yamashita[2], Takeshi Saitoh[3] and Yoshitaka Nakajima[1]
[1] Kyushu University, Fukuoka, JAPAN.
[2]Shibaura Institute of Technology, Tokyo, JAPAN.
[3]Kyushu Institute of Technology, Fukuoka, JAPAN.
m-fuyuno@design.kyushu-u.ac.jp, yama-y@shibaura-it.ac.jp,
saitoh@ces.kyutech.ac.jp, nakajima@design.kyushu-u.ac.jp

Abstract

This study examines connections between semantic structure and speech units and characteristics of facial movements in English as a Foreign Language (EFL) learners' public speech. The data were obtained from a multimodal corpus of English public speaking constructed from digital audio and video data from an English speech contest held at a Japanese high school. Evaluation data of contest judges were also included. For the audio data, speech pauses were extracted using acoustic analysis software. The spoken content (i.e. text) of each speech unit between two pauses was then annotated. The semantic structures of the speech units were analysed based on segmental chunks of clauses. Motion capturing was applied to the video data. 42 tracking points were set on the speaker's eyes, eyebrows, nose, lips and jawline. The results indicated: (1) Speakers with higher evaluations showed similar semantic structure patterns in speech units. Pause patterns and evaluation scores showed a strong correlation. (2) Face roll movement frequencies and the angles of face rolls for eye contact suggest that speakers with higher performance evaluations shared characteristic facial movement frequencies and degrees. These results may allow us to define model patterns for inserting pauses into public speech and develop facial movement criteria that effectively describe good eye contact patterns in public speaking.

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 447–461

1 Introduction

The ability to deliver effective presentations and speeches is considered an important professional skill in modern society. Public speaking skills can influence the outcome in various situations, such as professional meetings, conferences and job interviews. However, few people possess inherent public speaking skills and it takes practice and training for most people. In fact, many people fear public speaking, which is a common communication phobia for people across age groups (Kessler et al., 1998; Pertaub et al., 2001).

Furthermore, globalization has increased the importance of English public speaking skills (Fuyuno, 2015). Thus, developing significant proficiency in public speaking has become a common part of English as a Foreign Language (EFL) and English as a Second Language (ESL) courses. However, compared to speaking publicly in a speaker's native language, public speaking in English is difficult for EFL/ESL learners.

Although there have been various textbooks of English public speaking for EFL/ESL learners, many of them cover both writing and speaking together and the major part of textbooks tends to focus on the construction of speech rather than its effective delivery (cf. Jaffe, 2012). In addition to the content, the quality of a public speech is affected by nonverbal factors in delivery (Griffin, 2011; Batrinca et al., 2013). Despite this fact, nonverbal factors are typically not fully described in EFL materials, or even when considered, descriptions of such factors tend to be insufficient or ambiguous. For example, even when the importance of eye contact is addressed, the description of how to maintain effective eye contact tends to be vague (e.g. 'Look in at least three directions'.). More explicit examples that are based on concrete evidence are required for effective teaching.

Multimodal corpus analysis represents a possible way of improving EFL/ESL approaches. Indexes to set criteria for effective training and practice could be developed by analysing multiple factors in public speaking performance data. Recently, technological developments have improved data collection and analysis methods in corpus linguistics (cf. McCarthy, 1998; Rohrbach et al., 2012). Such developments have affected learner-corpus research and resulted in multimodal learner corpora, such as those developed by the International Corpus Network of Asian Learners of English (ICNALE) project and the Kyushu University Multimodal Corpus Analysis Project (KUMA Project). Learner-based multimodal corpora can store authentic audio and video data of learner speeches, which enables quantitative analysis of phonological and motion features. However, few studies have focused on public speaking.

The present study aims to apply a corpus-based approach to analyse and extract useful indexes of nonverbal factors in English public speaking for EFL learners. Our analysis focuses on speech pause insertion patterns and eye contact movement patterns from multimodal data of authentic English public speaking by Japanese EFL learners.

The subsequent parts of this paper are organised as follows. Section 2 provides an overview of the works related to public speaking performance analysis. Section 3 describes the data collection and analysis method. Section 4 presents results and discussions, and Section 5 provides conclusions and suggestions for future work.

2 Literature Review

Previous studies on the effectiveness of public speaking performances have been conducted in various fields (North et al, 1998; Amir et al., 2008; Batrinca et al., 2013; Fuyuno et al., 2016). Although analysis of nonverbal elements in public speaking performances has not frequently been addressed in corpus linguistics and English language teaching studies, the topic has attracted attention in other academic fields such as psychology, applied engineering and information processing.

For example, Bartica et al. (2013) analysed the performance of expert public speakers to extract factors that determine effective nonverbal behaviours during speeches from an information processing perspective. They audio-recorded two expert speakers' performances and analysed various parameters such as frequency of pauses and changes in voice intensity. They also video-recorded the performances and analysed the duration of speakers' eye contact using a virtual audience projected on a wide screen in front of the speaker. The results showed that the number of pauses and gazes towards the audience significantly affected the effectiveness of the speech.

Although the innovative analysis methods used by Bartica et al. (2013) revealed the importance of both phonological and motional factors in public speaking, the amount of data was limited, and the data came from professional public speakers speaking in their first language. Furthermore, the experimental setting was virtual; thus, the speakers were not addressing an actual audience.

Fuyuno et al. (2016) analysed EFL learners' English public speaking data from the perspective of multimodal corpora analysis and EFL pedagogy. They analysed a multimodal corpus of English public speaking by EFL learners to examine pedagogical implications for effective teaching of public speaking skills. The corpus data were collected in an authentic public speaking setting. The data included audio and video recordings of public speaking performances and the evaluation scores assigned by qualified judges.

Based on their results, Fuyuno et al. (2016) noted that the three speakers with the top evaluation scores shared similar speech pause duration patterns. Compared to speakers with lower scores, these speakers paused at commas and periods effectively and did not insert unnecessary short pauses. In addition, the highly evaluated speakers' relative cumulative frequencies of the duration of pauses in each category (i.e. comma/pause/others) were similar to those of native speakers of English (NSE).

Furthermore, horizontal facial movement patterns in 2D video data were analysed to examine effective eye contact movements. The results indicated that speakers who received higher evaluations shared similar characteristics relative to their head gesture patterns. Two speakers with high scores for movement aspects of the performance moved their heads horizontally to the left and right (in a 9-cm to 9-cm range approximately). Other speakers showed larger or smaller movement magnitudes. Fuyuno et al. (2016) suggested that a certain, appropriate amount of horizontal movement can enhance performance.

The results presented by Fuyuno et al. (2016) provide an objective reference for pedagogical applications. However, the analysis had various limitations. First, although the pause distribution patterns showed a correlation with speech evaluation scores, more cross-analysis between speech pauses and speech content is required to make the information practically applicable to classroom teaching. For example, learners need information about where and how often to place pauses in their speeches (scripts). Second, the facial movement analysis contained a technological limitation, i.e. the motion tracking method used by Fuyuno et al. (2016) did not distinguish between horizontal facial movement and face-rotation movement.

Considering these previous studies, the present study aims to analyse speech pause insertion patterns in terms of speech content and face-rotation patterns in eye-contact movement by focusing on EFL learners' public speaking performances. The research questions in this study are as follows.

1) Phonological aspect: What are the characteristic pause insertion patterns in English public speaking among high-scoring EFL speakers in terms of spoken content?
2) Movement aspect: What are the characteristic face roll movement patterns for eye contact among high-scoring EFL speakers?

3 Data and Method

3.1 Data

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 NSE 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 speech pause insertion patterns and facial movement patterns for eye contact, the scores for 'Rhythm', 'Speech Delivery' and 'Eye Contact' were extracted from the database and averaged (cf. Table 2). The two evaluation items for phonological analysis, i.e. 'Rhythm' and 'Speech Delivery', were chosen because speech rhythms and flows in delivery largely depend on pause patterns.

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

Table 1: Evaluation items

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. The devices were set on stable tripods (Fig. 1). After recording, the digital data were extracted and stored in a database (Fig. 2).

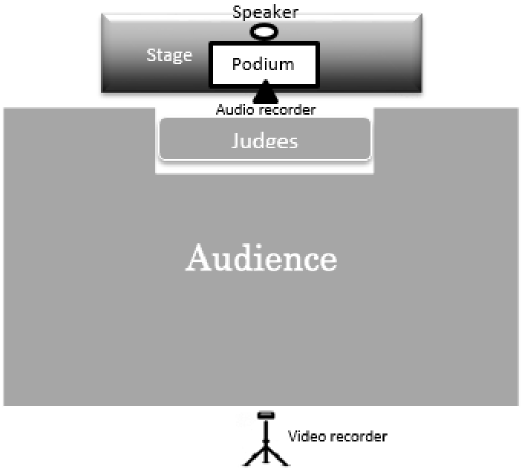

Figure 1: Arrangement for data recording

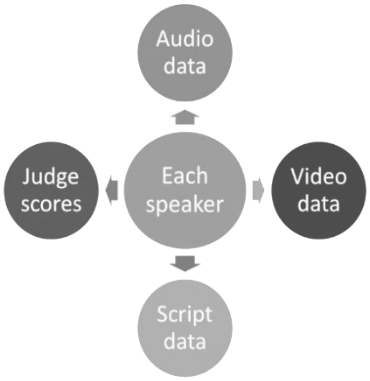

Figure 2: Datasets in our multimodal corpus

The basic descriptions of the datasets are summarized in Table 2. The average performance duration was approximately four minutes.

Speaker (Anonymized)	Script Type	Average Score (Rhythm and Delivery) /100	Average Score (Eye Contact) /100
S-01	A	59	68
S-02	C	65	60
S-03	A	63	60
S-04	C	80	88
S-05	A	73	74
S-06	B	89	92
S-07	A	65	78
S-08	C	74	64
S-09	A	68	70

Table 2: Basic description of the data

3.2 Method: Pause Insertion Patterns

Pause insertion patterns in public speaking can be described by various elements, such as pause duration, pause frequency, and the location of pauses. To analyse pause insertion patterns in terms of spoken content, criteria that describe such patterns are required.

In corpus analysis of natural utterances (e.g. free conversations between adult participants talking in their first language), basic exchanges are thought to consist of speech information units. Different ideas about speech units have been used in human communication studies; however, one of the most widely shared notions about speech units is the Intonation Unit (IU), which was suggested by Chafe (1987). An IU is a linguistic expression of information within utterances, and it plays the role of making speaker-listener communication smooth. An IU is normally defined as a single intonation contour in speech, but it is also typically separated by pauses. Therefore, this notion is applicable when considering speech pause insertion patterns and spoken content.

The IU is largely used in analysis and annotations of natural utterances. However, how can we define ‘good pauses’ in public speaking that differ from pauses in natural and spontaneous conversations? As mentioned previously, human utterances usually consist of speech units separated by pauses for smooth speaker-listener communication. In this sense, good pauses in public speaking may mark speech units that are semantically meaningful chunks in order to convey meaning clearly to the audience. Croft (1995) pointed out that nearly all IUs are also grammatical units, such as clauses or phrases. Because public speaking involves preparation and is intended to be more carefully delivered than natural utterances, ‘good pauses’ in effective public speaking may stably mark speech units that are semantically meaningful chunks.

Based on the hypothesis that high-scoring speakers may have more meaningful chunks in their speech units, the audio data were analysed using acoustic analysis software (Praat). First, randomly chosen 60-s pieces of audio data from the dataset of each speaker were extracted. Pauses were extracted automatically using the software. For this process, pauses were defined as speech intervals longer than 0.2 s (cf. Kendall, 2013). Next, each speech unit separated by pauses was annotated with spoken script. Figure 3 shows a screenshot of the Praat work screen.

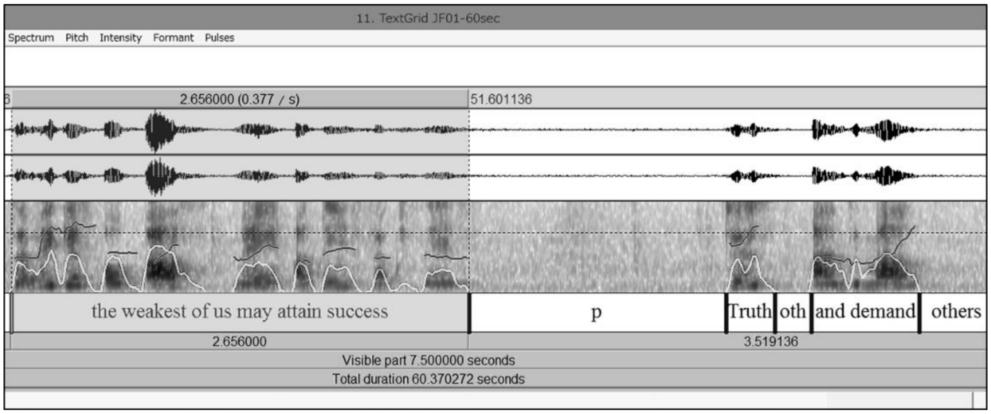
11. TextGrid JF01-60sec
Spectrum Pitch Intensity Formant Pulses
2.656000 (0.377 / s)
51.601136
the weakest of us may attain success
p
Truth
oth
and demand
others
2.656000
3.519136
Visible part 7.500000 seconds
Total duration 60.370272 seconds

Figure 3: Screenshot of automatic pause extraction and speech annotation using Praat

Figure 4 shows examples of the annotated data of two speakers. The black dots represent speech units and the white dots represent pauses. The sequences (A) and (B) show the two speakers’ results, respectively.

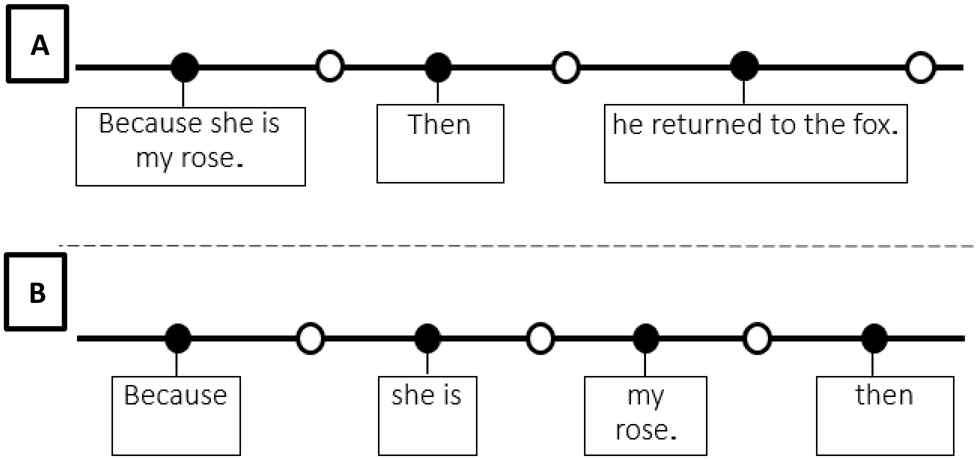

Figure 4: Examples of annotated data

After automatic pause separation and annotation, the speech units without clauses or punctuation in all speaker data were counted. We refer to such units as *semantically incomplete units*. For example, in the data of the two speakers shown in Figure 5, the speech units indicated by thick rectangles are semantically incomplete because they do not include clauses or punctuation.

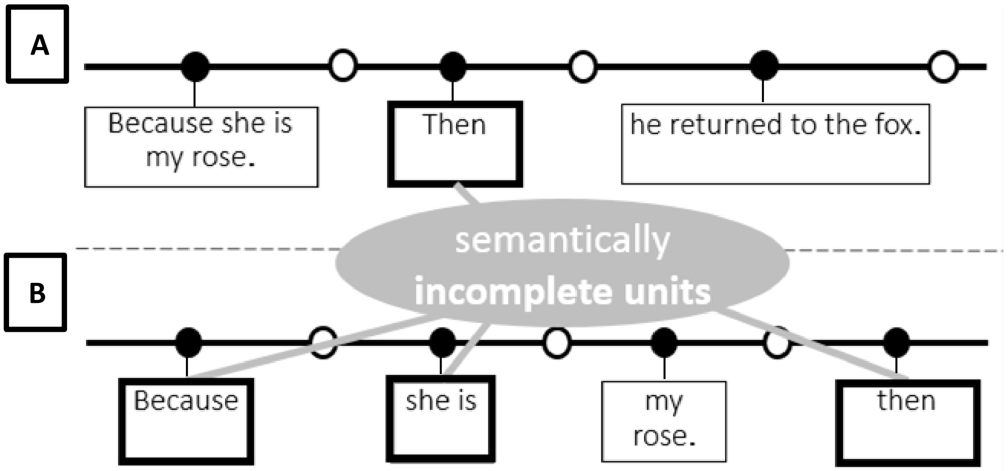

Figure 5: Examples of semantically incomplete units

Finally, the ratio of incomplete units to all speech units was calculated for all speakers. The results are discussed in Section 4.1.

3.3 Method: Eye Contact Movement Patterns

To analyse the speakers' facial motion patterns, motion capturing was performed with a CV-based original program for each speaker's video data. The program is based on the active appearance model (AAM) (Cootes et al., 2001). This method allowed us to track pre-set feature points objectively and automatically (cf. Adolphs & Carter, 2013; Fuyuno et al., 2016; Komiya et al., 2016a; 2016b). Forty-two feature points were set on each speaker's facial parts, i.e. jawline, eyebrows, eyes, nose and lips, as shown in Figure 6.

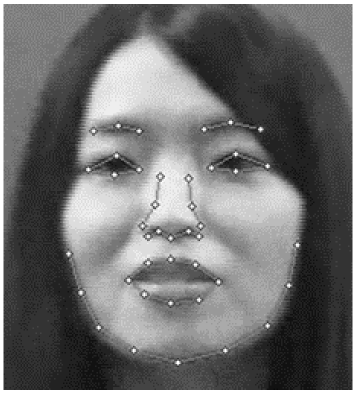

Figure 6: Locations of feature points

The speaker facial motions, including face roll degrees, were extracted as a series of numerical values by tracking these feature points automatically[2]. In this study, we focused on the speaker's facial roll frequencies and degrees rolled because these directly relate to eye contact movement. Figure 7 shows a sample motion tracking result. The speaker's face roll movement tracks to the right and left sides were obtained by the illustrated process.

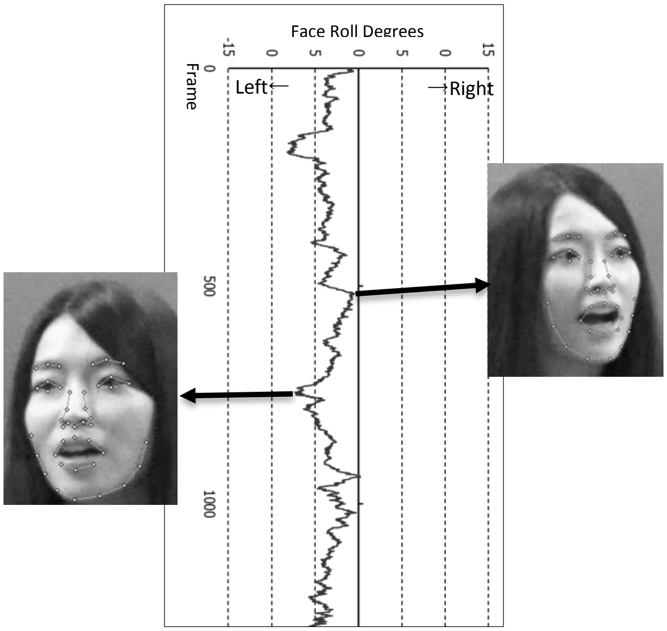

Figure 7: Sample motion tracking results of face roll movement

The Fourier transform and power spectrum were calculated from the motion track data. Based on the maximum frequency in each dataset, the speakers' face roll frequencies per minute were obtained. The results are discussed in Section 4.2.

4 Results and Discussion

4.1 Result: Pause Insertion Patterns

Based on the analysis of pause insertion patterns in the audio data, the ratios of semantically incomplete units to all speech units were obtained, as shown in Table 3, which lists the results in order of average score (high to low). It seems that speakers with high evaluation scores for their speech rhythms demonstrate a relatively lower incomplete unit ratio.

Rank	Speaker	Script Type	Average Score (Rhythm and Delivery) /100	Incomplete Unit Ratio (%)
1	S-06	B	89	18.5
2	S-04	C	80	13.6
3	S-08	C	74	19.2
4	S-05	A	73	21.7
5	S-09	A	68	24.0
6	S-02	C	65	24.1
6	S-07	A	65	20.8
8	S-03	A	63	27.2
9	S-01	A	59	29.6

Table 3: Results of incomplete unit ratio

To compare the results to NSE samples, two NSE datasets were recorded. Two speakers were handed a speech script (the content was the same as assignment C[1]). The speakers had five minutes to read the script and practice. After preparation, the speakers performed the speeches in front of an audience of three people. The speeches were recorded with the same recording device used for compilation of the multimodal corpus. The results of the NSE samples are shown in Table 4. The incomplete unit ratios in the NSE sample data, i.e. the ratios of incomplete units in all speech units, were both quite low, and they are considered lower than the normal utterances of NSEs (cf. Chafe, 1994). This reflects the fact that public speaking speech is the result of certain preparations and that the speech script itself is pre-written and revised for speaking. The results indicate that fluency in public speaking performance and incomplete unit ratio are related to an extent.

Speaker	Script Type	Incomplete Unit Ratio (%)
NSE-01	C	0
NSE-02	C	7.7

Table 4: Incomplete unit ratio of NSE samples

Using the results shown in Table 3, Spearman's rank-order correlation coefficients were computed by comparing the scores and the incomplete unit ratios. The results of the correlation confirmed a strong negative correlation (r = −.91). It was confirmed significant (P < .01). Figure 8 shows the correlation between the scores and the incomplete unit ratios.

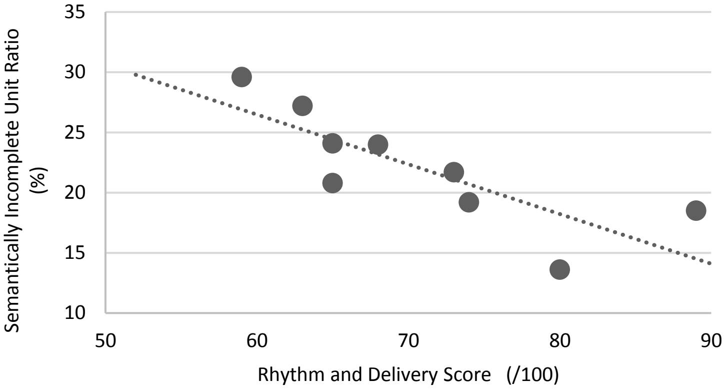

Figure 8: Correlation between Rhythm and Delivery scores and semantically incomplete unit ratios

The results suggest that speakers with high evaluation scores for speech rhythms tend to pause at semantic boundaries, i.e. between clausal units. In contrast, other speakers tend to pause at boundaries without clauses. Several possibilities could explain this tendency: (1) speakers are not sufficiently fluent (e.g. memorization of the script and/or practice was insufficient); (2) speakers emphasize the meaning of words by placing pauses around them intentionally; or (3) speakers do not pay attention to semantic units in the speech script. Regarding the second point, the two sequences shown in Figure 5 indicate this pattern. Sequence A in Figure 5 is from the results of a speaker with a high score, and sequence B is from a relatively lower scoring speaker. The speaker of sequence B appeared to place pauses after ‘because’ and ‘she is’ intentionally to emphasise the sentence dramatically.

In summary, speakers with high evaluations tend to pause speech at semantic boundaries, such as punctuation marks and between clausal units, while other speakers made speech pauses that marked semantically incomplete units. Thus, the incomplete unit ratio to the total speech units of a public speaker can be considered an objective index to assess performance.

4.2 Results: Eye Contact Movement Patterns

From the motion tracking results, face roll frequency per minute and face roll degree for each speaker were obtained as shown in Tables 5 and 6. Both categories were sorted in order of average score for eye contact (high to low).

Rank	Speaker	Average Score (Eye Contact) /100	Face Roll Frequency per Minute
1	S-06	92	7
2	S-04	88	8
3	S-07	78	9
4	S-05	74	15
5	S-09	70	7
6	S-01	68	22
7	S-08	64	4
8	S-02	60	7
8	S-03	60	7

Table 5: Result of face roll frequency per minute

Rank	Speaker	Average Score (Eye Contact) /100	Average Face Roll Degree
1	S-06	92	5.20
2	S-04	88	5.46
3	S-07	78	3.24
4	S-05	74	2.13
5	S-09	70	2.02
6	S-01	68	1.57
7	S-08	64	4.91
8	S-02	60	2.73
8	S-03	60	3.75

Table 6: Results of average face roll degrees

The results of face roll frequency per minute show how many times a speaker changed face roll direction (to the right or left) on average in one minute. As can be seen in Table 5, the frequencies of the top three speakers indicate similar numbers; they changed face roll direction eight times per minute on average (i.e. once every seven seconds). Compared to these three, the lower ranked two speakers (i.e. S-05 and S-01) changed face direction more frequently. These may have been considered too frequent, thereby resulting in a negative impression, as reflected by the evaluation scores. An F-test was performed to compare the difference in variances between the top three speakers and other speakers. The results indicate a significant difference (F (5, 2) = 46.2, p < .05). However, the two lowest scoring speakers (i.e. S-02 and S-03) showed frequencies similar to the top three speakers. Why would this happen?

The answer seems to lie in the actual amount of speaker movement. If a speaker rolled their face to larger degrees, the speaker could make eye contact with a wider audience to both the right and left edges. The average face roll degrees shown in Table 6 indicate that the top three speakers tended to move their faces at greater degrees. In fact, the average face roll degrees of the top three speakers were greater than those of the other speakers (top three: 4.63; others: 2.85). There was a marginal difference in the degrees for the top three speakers (SD=0.99) and the others (SD=1.14) in a two-tailed t-test; t (7)=2.02; p=0.08. Figure 9 illustrates box-and-whisker plots of the comparison.

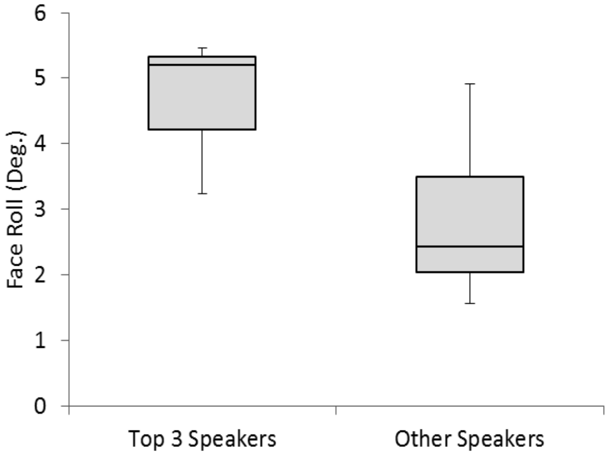

Figure 9: Box-plots of the face roll degrees of the top three speakers and the others

The results of the motion track graphs describe these differences clearly. The graphs of two speakers were extracted from the results as an example. In Figure 10, the graph on the left shows the face roll motion track of S-04 (evaluation score = 88/100) and the graph on the right shows the face roll motion track for S-02 (evaluation score = 60/100).

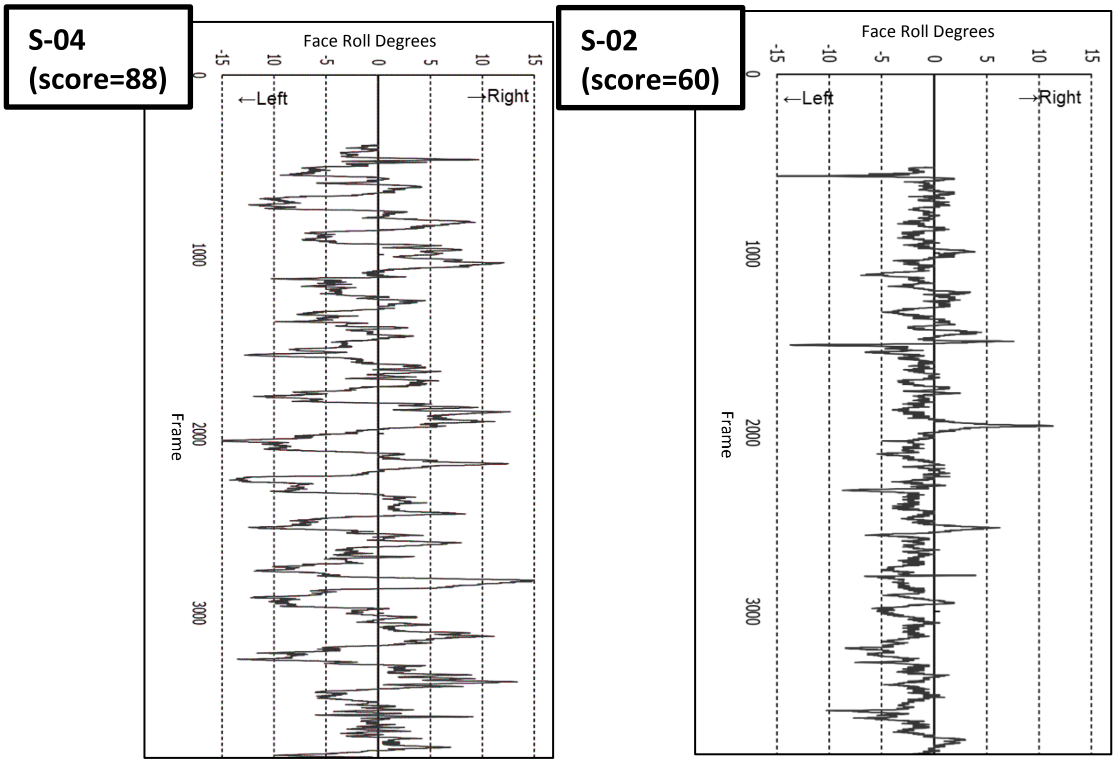

Figure 10: Examples of motion track results

As can be observed in the graphs, S-04 rolled his/her face more dynamically to both sides. In fact, the average coverage to the two sides was 13.4 degrees for S-04 and 8.96 for S-02. To maintain eye contact with a wide range of the audience, speakers may need to move their faces to wider degrees rather than simply moving their eyes.

Figure 11 shows a scatter plot of the average face roll frequencies and degrees.

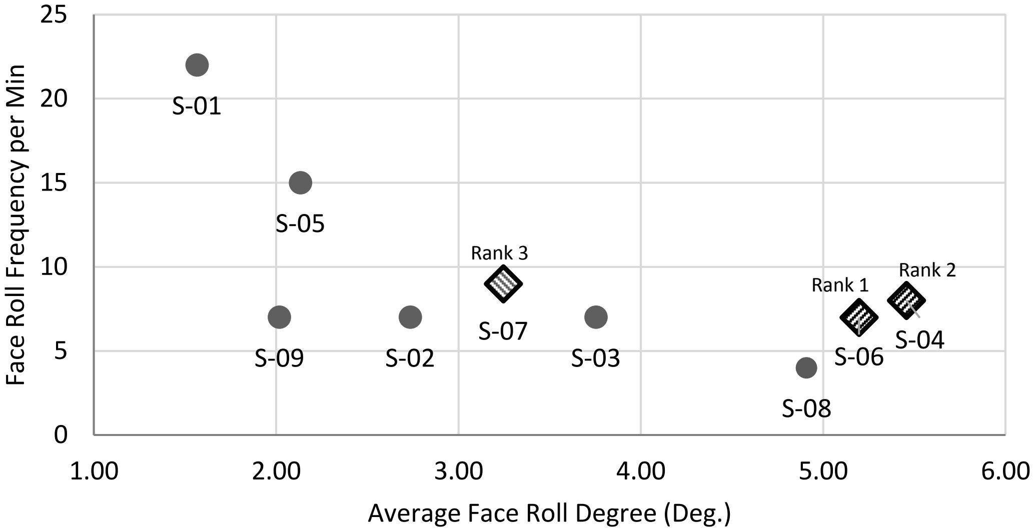

Figure 11: Scatter plot of face roll frequencies and face roll degrees

As can be seen, the top two speakers, i.e. S-06 and S-04, are plotted in the same area. A tendency for effective facial movement to maintain eye contact can be observed from these results. The top two speakers show larger face roll degrees compared to the other speakers, with a frequency of approximately eight times per minute. These results demonstrate an example of adequate eye contact movement for EFL learners.

5 Conclusions

In this study, pause insertion patterns and face movement patterns in EFL learners' public speaking have been analysed using data from a multimodal corpus. The results of cross analyses between pause insertion patterns and performance evaluation scores and between facial movement patterns and performance evaluation scores indicated that highly-evaluated public speakers demonstrate similar tendencies. These results could be used as evidence-based examples for teaching public speaking in English and performance assessment.

Our quantitative analysis results suggest various future prospects. First, there were connections between speech content and speech pauses and a strong correlation between these and speech evaluations. The timing of speech pauses seems to be crucial to determining the quality of public speaking. However, in addition to the speech voice and spoken content, other behavioural factors, such as eye contact, hand gestures and facial expressions, interact simultaneously in public speaking.

In future studies, the relationships among all phonological, content and behavioural factors should be examined.

In addition, this study has focused on eye contact facial movement using movement analysis. Different types of motion factors, e.g. eye movement, hand gestures and postures, could provide useful clues for pedagogical application. In fact, high scoring speakers in our multimodal corpora were observed to use these factors effectively. Methods to handle and analyse the corpus data to elucidate the characteristics of such factors will be the target of future studies.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science Kakenhi Grant numbers 15K12416 and 16H03079, and the Kyushu University Projects in Research Development (Type: Tsubasa, Number: 27817).

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; (B) an excerpt from Haruki Murakami's acceptance speech for the Jerusalem Award; and (C) an excerpt from 'The Little Prince' (English translation) by Antoine de Saint-Exupéry.

2) For more detailed information about the technological descriptions, please refer to Komiya et al. (2016b).

References

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), p. 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 *Intelligent Virtual Agents* (pp. 116-128). Berlin: Springer Berlin Heidelberg.

Chafe, W. (1987). Cognitive constraints on information flow. *Coherence and grounding in discourse, 11*, pp. 21-51.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

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

Croft, W. (1995). Intonation units and grammatical structure. *Linguistics, 33*(5), pp. 839-882.

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*, pp. 13-27.

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.

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

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

Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: studies in corpus sociophonetics.* Palgrave Macmillan.

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

Komiya, R., Saitoh, T., Fuyuno, M., Yamashita, Y., & Nakajima, Y. (2016a). *Feature Points based Head Motion Analysis for Public Speech Guidance* [in Japanese]. Paper presented in IEICE General Conference 2016 at Kyushu University, D-15-32, p.21.

Komiya, R., Saitoh, T., Fuyuno, M., Yamashita, T., & Nakajima, Y. (2016b). Head pose estimation and movement analysis for speech scene. *Proc. of 15th IEEE/ACIS International Conference on Computer and Information Science* (to be appeared).

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

North, M. M., North, S. M., & Coble, J. R. (1998). Virtual reality therapy: an effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, *3*(2), pp. 2-6.

Pertaub, D. P., Slater, M., & Barker, C. (2001). An experiment on fear of public speaking in virtual reality. *Studies in health technology and informatics*, pp. 372-378.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, M. (2012). A database for fine grained activity detection of cooking activities. *Proceedings of Computer Vision and Pattern Recognition*, pp. 1194-1201.