ANALYSIS OF MULTIMODAL FEATURES FOR SPEAKING PROFICIENCY SCORING IN AN INTERVIEW DIALOGUE

Mao Saeki[1], *Yoichi Matsuyama*[1], *Satoshi Kobashikawa*[2], *Tetsuji Ogawa*[1], *Tetsunori Kobayashi*[1]

[1]Department of Communications and Computer Engineering, Waseda University, Japan
[2]NTT Media Intelligence Laboratories, NTT Corporation, Japan
{saeki,matsuyama,ogawa}@pcl.cs.waseda.ac.jp
satoshi.kobashikawa.he@hco.ntt.co.jp, kobas@waseda.jp

ABSTRACT

This paper analyzes the effectiveness of different modalities in automated speaking proficiency scoring in an online dialogue task of non-native speakers. Conversational competence of a language learner can be assessed through the use of multimodal behaviors such as speech content, prosody, and visual cues. Although lexical and acoustic features have been widely studied, there has been no study on the usage of visual features, such as facial expressions and eye gaze. To build an automated speaking proficiency scoring system using multimodal features, we first constructed an online video interview dataset of 210 Japanese English-learners with annotations of their speaking proficiency. We then examined two approaches for incorporating visual features and compared the effectiveness of each modality. Results show the end-to-end approach with deep neural networks achieves a higher correlation with human scoring than one with handcrafted features. Modalities are effective in the order of lexical, acoustic, and visual features.

***Index Terms*—** Speaking proficiency assessment, multimodal machine learning, BERT (Bidirectional Encoder Representations from Transformers)

1. INTRODUCTION

In the context of the growing need for online education prompted particularly by the transmission of the novel coronavirus infection in 2020, automated speaking skill assessment is becoming increasingly important to place them in proper classes and adaptively provide learning materials and feedback. In recent years, much work on automated speaking proficiency assessment for non-native speakers has been conducted with the aim of providing reliable tests and generating feedback for effective learning [1]. However, most research of speaking proficiency test focus on the scoring of monologue speech, which fails to capture the conversational aspect such as interactional, sociolinguistic, inter-cultural communication competences [2, 3], besides competencies of pronunciation, vocabulary and grammar. Although dialogue systems allow to holistically assess speaking proficiency, very little work has been conducted up to now, because of their technical limitations. Responses in a dialogue have larger variance than in a monologue assessment, making the speech recognition more challenging. Litman et al. predicted human expert rating of a human-machine dialogue using audio and fluency features [4]. Ramanarayanan et al. predicted interactional aspect of a human-machine dialogue using text features [5]. Interlocutors exchange further non-verbal cues, such as eye gaze, in their dialogues [6], however, previous work on dialogue-based proficiency scoring system have only focused on limited modalities. Another challenge with speech proficiency scoring in dialogues is there being few dialogue dataset of language learner from a range of proficiency.

This paper investigates the usage of multiple modalities for holistically assessing speaking proficiency of non-native speakers in a dialogue task where conversational competencies can be sufficiently observed. The contributions of this paper are twofold: 1) we collect approximately 30 hours of online video interviews with 210 Japanese English-learners, scored for their holistic speaking proficiency, as one of the largest dialogue-based proficiency assessment datasets. Parts of the dataset are further scored by five raters for six qualitative features; 2) we propose the first neural model that incorporates lexical, acoustic and visual cues for scoring speaking proficiency in a dialogue task, and compare effectiveness of different modalities.

The rest of the paper is organized as follows. Section 2 briefly reviews previous work on automated language assessment to place this work in the context of dialogue-based speaking proficiency scoring. Section 3 explains the collection of the interview dataset and human scoring. Section 4 describes our proposed end-to-end dialogue scoring model as well as conventional feature engineered model we use as a baseline to evaluates the effectiveness of the proposed approach and different modalities. Finally, section 5 draws conclusion and discuss future research directions.

2. RELATED WORK

While speech proficiency scoring has been much studied in recent years, most of them focus on monologue speech. These include tasks such as reading aloud and answering written/spoken prompts. In the case where experts response to a prompt can be obtained, it is possible to score by comparing the learner's response to the expert's [7]. However collecting expert response can be impractical for tests where questions must be different each time. Conventional method has been to train linear regression models on set of hand crafted speech features extracted from the response [8]. Neural approach is gaining popularity in recent years for it allows learning of important feature instead of hand crafting them. [9] proposed a model that scores monologues in an end-to-end fashion using lexical and acoustic cues. They showed Bidirectional LSTM (BLSTM) with attention mechanism to outperform conventional method using hand crafted features. As a prior work on scoring dialogue response, [10] used acoustic features (F0, power) and fluency features (silence, disfluency, word and phone count) to score non-native's proficiency in a human-machine dialogue. [5] predicted the interactional aspect of a human-machine dialogue using text features. They combined a model using hand crafted features and an End to End Memory Network [11] based model and showed automatic rating close to human. [12] used an attention based BLSTM to score content, delivery and language use respectively, which were concatenated and fed into a dense layer to predict the holistic score. The inputs used were lexical and acoustic features. There has been few large scale dialogue dataset of non-native speakers collected over the years. The Trinity Lancaster Corpus includes speech and transcript of 2,053 interviews, and classified into three levels based on CEFR [13]. The ICNALE Spoken Dialogue (ICNALE SD) collected video interviews of 405 Asian learners, and classified into four levls also based on CEFR [14]. However they are both carried out in a face-to-face setting which differs in characteristic to online interviews.

Our brief survey has shown limitation of prior research. Firstly, previous research have only used limited modalities, and have not investigated the combination of modalities. In addition, there are no multimodal interview dataset of non-native speaker conducted in online video-conference setting. Therefore, in this paper we address the issues mentioned above.

3. INTERVIEW DATASET

3.1. Data Collection

Traditional speaking test often involves an interactive dialogue between human examiner and test taker [15]. Therefore it would be a natural to use an interview dialogue for automated scoring. We used an interview test used for an actual medium-stake class placement test at an university communicative English course. The test is normally carried out face-to-face, however we conducted it as an online video interview for easier collection. It is also likely that with recent global situation and rapid deployment of online teaching, more and more language assessment will be conducted in online space. Though face-to-face and online conversation differ in nature, research shows that online assessment is possible with high reliability [16]. The interview consists of ten oral questions ranging from ice breaking questions to topic interviewees were less familiar with, lasting about 10 minutes. Fig 1 shows a scenario of the interview dialogue. In the early phase, an interviewer asks a few questions to assess the interviewee's approximate CEFR levels (A-level, B-level, and C-level), and then assess fine-grained levels by asking additional questions. Although the questions were fixed, the interviewer may ask follow up questions depending on the length and content of the answer. We used the video conferencing tool Zoom [1] for the experiment. Interview were carried out by experienced English teachers. Interviews of 210 Japanese English learner were collected. Learners consisted of undergraduate and graduate university students, their conversational proficiency ranging from beginner to native level. Audio and video of both the interviewer and interviewee were collected.

3.2. Human Scoring

To build the dataset with speaking proficiency scores, interviews were scored by human raters using the Common European Framework of Reference for Language (CEFR), an international standard for measuring language proficiency that provides a set of common reference level along with its descriptors [2]. We adopted the scale for ”communicative language competence”, and will refer to this as the overall CEFR. The overall CEFR that is the holistic speaking proficiency is further broken down into subcategories of six qualitative features: lexical range (**Range**), grammatical accuracy (**Accuracy**), fluency of speech (**Fluency**), goodness of pronunciation (**Phonology**), interactional competence (**Interaction**) and sentence coherence (**Coherence**). Although we aim to automatically score the overall CEFR level, we had the raters score part of the interviews for the whole subcategory to see what aspect of speaking proficiency would be easier or harder to score. All CEFR categories consists of 6 levels: A1, A2, B1, B2, C1 and C2. A1 represents the lowest proficiency, and C2 the highest. Learner in the C2 range were very rare, therefore we scored them in the same level as C1.

We used five raters in total, each rater having extensive experience in English teaching. Raters were instructed to watch the whole interview recording and score the proficiency. In order to calculate the inter-rater reliability, 21 interviews were annotated by all five raters for the CEFR level and its subcat-

[1] https://zoom.us/

Level Check		
Hello my name is….. What's your name? Nice to meet you, _________. Please try to speak as much as you can in this interview.		
2a) Where do you live? Do you live in a house or apartment? 2b) Tell me about it (your house or apartment).		
3a) What is your favorite season? 3b) What do you like to do in _________?		
A1 to A2	**B1 to B2**	**C1 to C2**
4) What did you eat for breakfast this morning?	4 How are your weekdays different from your weekends/holidays?	4) What are some difficulties English learners face? How can they overcome these difficulties?
5) What will you do tomorrow?	5) Tell me about a movie or a TV program that you watched recently.	5) What are some things you would like to change about your university?
6) What do you like to do in your free time? (How long have you been doing that?)	6) Which country would you like to visit in the future? What would you like to do there?	6) Do you think students should be asked to evaluate their teachers?
7) Have you ever been to a foreign country? No - Which countries do you want to go to in the future? Yes - Tell me about your trip.	7) What are some good and bad points about social networking sites like Facebook and Twitter?	7) What are some advantages and disadvantages of globalization?
8) Tell me about your favorite place in Tokyo.	8) Tell me about your best friend. (Why do you think he/she is _____?)	8) How do you think Japan will change in the next 10 years?
9) Now, please ask me some questions.	9) Now, please ask me some questions.	9) Now, please ask me some questions.
Thank you. This is the end of the interview.		

Fig. 1. Scenario of the interview dialogue. In the early phase (level check), an interviewer asks a few questions to assess the interviewee's approximate CEFR levels (A-level, B-level, and C-level), and then assess fine-grained levels by asking additional questions.

egories. The inter-rater reliability of each category calculated in Krippendorff's α [17] are as follows: overall CEFR 0.72, Range 0.72, Accuracy 0.75, Fluency 0.66, Phonology 0.55, Interaction 0.62, and Coherence 0.56. The overall CEFR, Range and Accuracy shows high agreement while the other categories show medium agreement. This is an encouraging result since it shows that all dimension of conversational proficiency can be measured through an online video interview. The rest of the interviews were scored for overall CEFR by a single rater, since high reliability was confirmed.

The number of interviewees in A1 to C2 level was 45, 35, 50, 45, 30, 5, respectively. Since there were only five C2 level students and greatly imbalanced with other levels, we aggregated C1 and C2 levels as C1+.

During the annotation we also collected the raters comments on the use of visual cues. One notable finding was that facial expression and eye movement help understand the person's mental load when formulating a speech, which can be a marker for their proficiency. Also cultural background can be visually recognised. For example it is typical for a Japanese speaker to nod their head at the end of a statement, which also happens for native English speaker but less often. Some gestures affected by the native language could confuse English speakers, and the inability to suppress them can be accounted for low proficiency. These observations support the use of visual features.

4. EXPERIMENTS

This section compares two approaches for automatically scoring conversational proficiency for the collected interview dialogue. We first explain the conventional model which use hand crafted features, which have been widely used for speech scoring [18]. We then explain our proposed end-to-end model that incorporates lexical, acoustic and visual cues. The machine scoring models are then evaluated by comparing the prediction with human scoring. We also analyze the effectiveness of different modalities by using different combination of modalities for the neural model.

4.1. Conventional Model

We chose set of features used in many previous automatic speech scoring models [5, 18]. The features are listed in Table 1, categorised in the qualitative features of the overall CEFR. All features were extracted from the whole leaner

Table 1. Features used for the conventional model

Qualitative Feature	Features
Range	Word n-gram (n=1, 2) count Word complexity
Accuracy	Language model posterior
Fluency	Speech rate Pause frequency Utterance duration
Phonology	Acoustic model posterior

Interaction	Frequency of transition mode;
	Overlap (under 0 sec);
	Latch (between 0 and 0.5 sec);
	After gap (after 0.5 sec)

Coherence	Discourse marker count

utterance during the interview. Words and acoustic/language model posterior was obtained using an ASR system developed in [19]. Word complexity was estimated using the CEFR-J Wordlist[2], which contains the pseudo-CEFR level for each word. Fluency related features were extracted using Praat[3].

Using 17 interviews as a development set, we compared Multi layer perceptron (MLP), Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT), and found MLP to perform best. We will hereafter use MLP as the classifier of the conventional model.

4.2. Neural Model

Fig. 2 depicts the overall architecture of our proposed neural model. The model consists of three submodels that encodes lexical, acoustic and visual features respectively. Encoded features for n dialogue turns are then concatenated and passed to a full connection and a softmax layer, which outputs the probability of each CEFR level. Note that the model can work with a single or combinations of submodels since they are fusioned by a simple concatenation. We use n turns instead of the while dialogue because the neural model requires a lot of data to train, and using n turn allows the extraction of more data in a sliding window fashion. Encoding is done for each turn so as to align features of different modalities. The model is trained for every n dialogue turns, and during testing we compute the score of the leaner as the median of the scores predicted for the whole interview dialogue. Using the same development set as section 4.1, we found n = 4 to perform well while being small enough to be able to extract more training data, meaning that the model will process four dialogue turns in a sliding window to predict the overall CEFR. We next describe each submodels.

Lexical Model: The lexical model takes the text prompt and response pair as input. We use both the interviewee and interviewer utterance to capture the appropriateness of responses, an important factor for assessing speaking proficiency. Bidirectional Encoder Representations from Transformers (BERT) [20] was used for word embedding. BERT extracts contextual word representations that can be easily fine tuned for downstream tasks, and has been shown to achieve state-of-the-art results in many natural language understanding tasks. We used the BERT base model with 768 units with pretrained weights obtained at HuggingFaces Transformers[4]. The word embeddings are average pooled and passes though a linear layer to obtain lexical features.

Acoustic Model: For input, we use pitch and power extracted every 10ms. Features are extracted using and BLSTM with attention mechanism. BLSTMs allow the processing of sequential input and learn long term forward and backward dependencies. Attention mechanism allows the model to focus on important points of the input, and together with BLSTM have been shown to outperform other models in speech scoring tasks [9, 12].

Visual Model: Facial features, head movements and eye gaze was observed in section 3 to be helpful for speaking proficiency scoring. We used Facial Action Unit (AU) to capture facial features. AU encodes facial muscle movement, and is a main building block for facial expression analysis [21]. We used OpenFace [22] to extract eye gaze direction, head tilt and intensity of 18 AUs every 20ms. BLSTM with attention mechanism was used to extract visual features.

4.3. Results and Discussion

PyTorch[5] was used to implement all models, and cross entropy loss was used. We used a 10-fold cross-validation on the remaining 193 interviews to evaluate all the models. For the neural model, we experimented with the combination of submodels or when using a single submodel to see the effectiveness of different modalities. Specifically, we used the combination of all submodels (multimodal model), lexical and acoustic model, lexical and visual model. We computed the unweighted accuracy (UA), weighted accuracy (WA) and Pearson's r between the model prediction and human scoring over three runs, and report the average and standard deviation in Table 2. WA is the classification accuracy of all interviews and UA is the average of individual class accuracies.

The combination of lexical and acoustic feature achieved the highest UA and WA of 0.484 and 0.478 respectively, outperforming the conventional method. This result proves the advantage of training feature extraction and scoring model in an end-to-end fashion rather than designing sophisticated features which requires extensive knowledge in language education and assessment. The confusion matrix of the neural model using lexical + acoustic features is shown in Fig 3.

Comparison of the neural model using different submodels show that modalities are effective in the order of lexical,

[2] http://www.cefr-j.org/download.html

[3] http://www.praat.org/

[4] https://github.com/huggingface/transformers

[5] https://pytorch.org/

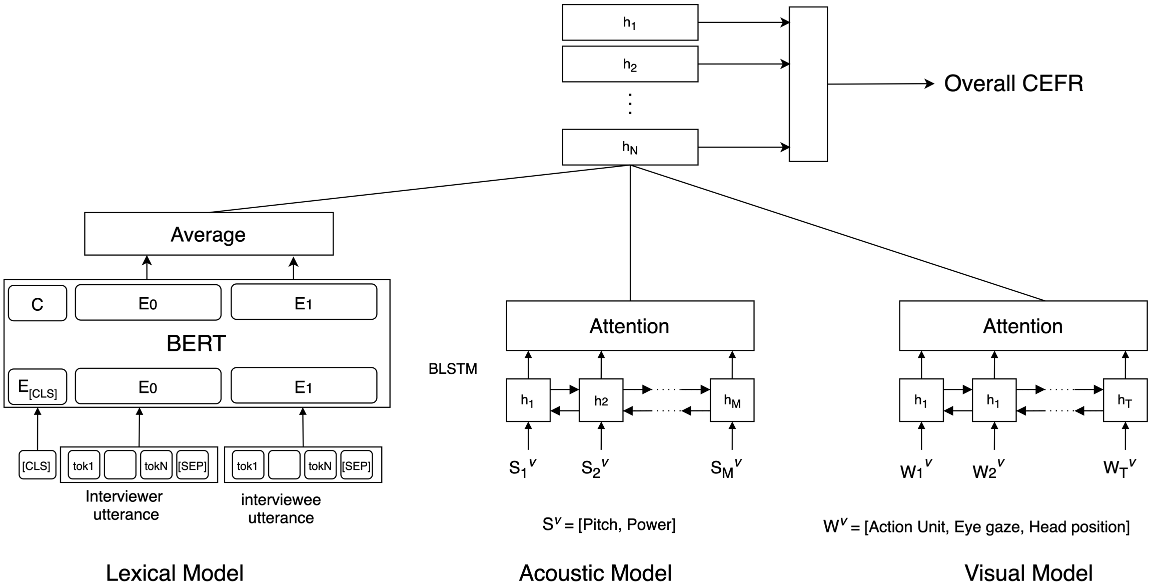


Fig. 2. Multimodal speaking proficiency scoring model architecture

Table 2. Comparison of unweighted average (UA), weighted average (WA) and correlation between human scoring and prediction of machine scoring models

Model	UA	WA	Pearson's r
conventional model	0.421 ± 0.015	0.414 ± 0.015	0.676 ± 0.008
multimodal model	0.462 ± 0.008	0.451 ± 0.016	$\mathbf{0.799 \pm 0.018}$
lexical+ acoustic model	$\mathbf{0.484 \pm 0.017}$	$\mathbf{0.478 \pm 0.018}$	0.786 ± 0.005
lexical+ visual model	0.442 ± 0.033	0.426 ± 0.033	0.786 ± 0.025
lexical model	0.442 ± 0.017	0.436 ± 0.018	0.777 ± 0.014
acoustic model	0.223 ± 0.024	0.229 ± 0.018	0.165 ± 0.026
visual model	0.189 ± 0.016	0.203 ± 0.006	0.030 ± 0.030

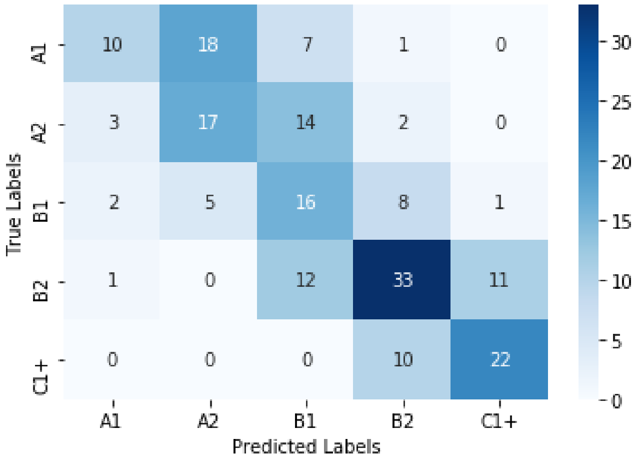
acoustic and visual. Despite the importance of visual cues in interaction, using visual cues alone showed very low accuracy. One challenge of using visual features is the large individuality of the use of visual cues. For example, learners showed a variety of gestures while thinking such as frowning, eye rolling and freezing. Also there is no explicit mention of verbal cues in the overall CEFR rubric, meaning raters use non-verbal cues as an indirect measurement of the level. For these reasons it is possible that it was difficult to learn any useful visual features from the amount of data used in this paper. In the future we could identify specific non-verbal cues that contribute to the classification, and explicitly extract them to input to the model.

5. CONCLUSIONS

This paper has investigated the use of lexical, acoustic and visual cue for speaking proficiency assessment. First we have collected an online video interview dataset of Japanese English-learner. We have observed that humans raters are able to score CEFR level and its subcategories at a medium to high inter-rater reliability in an online interview setting. We then proposed a speaking proficiency scoring model that incorporates lexical, acoustic and visual features, and analysed the effective modality by using each features individually. Results showed that the modalities were effective in the order of lexical, acoustic and visual. Combination of lexical and acoustic features are able to show high accuracy, outperforming conventional method using hand crafted features. Many challenges remain for scoring speaking proficiency in dialogues. More studied needs to be made in the future to identify the key visual cues for proficiency scoring.

6. ACKNOWLEDGEMENTS

This work was partially supported by the joint research with Waseda University Academic Solutions Corporation ("Automatic Tutorial English Class Placement System Based on



True Labels \ Predicted Labels	A1	A2	B1	B2	C1+
A1	10	18	7	1	0
A2	3	17	14	2	0
B1	2	5	16	8	1
B2	1	0	12	33	11
C1+	0	0	0	10	22

Fig. 3. Confusion matrix of the overall CEFR scoring using lexical + acoustic model

Speaking Proficiency").

7. REFERENCES

[1] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] Council of Europe, "Common European Framework of Reference For Languages: Learning, Teaching, Assessment," *Cambridge University Press*, 2018.

[3] Maria Elena Oliveri and Richard J. Tannenbaum, "Are We Teaching and Assessing the English Skills Needed to Succeed in the Global Workplace?," *The Wiley Handbook of Global Workplace Learning*, pp. 343–354, 2019.

[4] Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke, "Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Los Angeles, Sept. 2016, pp. 270–275, Association for Computational Linguistics.

[5] Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian, "Scoring interactional aspects of human-machine dialog for language learning and assessment using text features," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, Sept. 2019, pp. 103–109, Association for Computational Linguistics.

[6] Kristiina Jokinen, "Non-verbal signals for turn-taking and feedback," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010, European Language Resources Association (ELRA).

[7] Su-Youn Yoon and Chong Min Lee, "Content Modeling for Automated Oral Proficiency Scoring System," in *Workshop on Innovative Use of NLP for Building Educational Applications*, Florence, Italy, aug 2019, pp. 394–401, Association for Computational Linguistics.

[8] Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov, "Using exemplar responses for training and evaluating automated speech scoring systems," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, June 2018, pp. 1–12, Association for Computational Linguistics.

[9] Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian, "End-to-end Neural Network Based Automated Speech Scoring," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[10] Diane Litman, Helmer Strik, and Gad S. Lim, "Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities," in *Language Assessment Quarterly*. 2018, vol. 15, pp. 294–309, Routledge.

[11] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus, "End-to-end memory networks," *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 2440–2448, 2015.

[12] Y. Qian, P. Lange, K. Evanini, R. Pugh, R. Ubale, M. Mulholland, and X. Wang, "Neural approaches to automated speech scoring of monologue and dialogue responses," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8112–8116.

[13] Dana Gablasova, Vaclav Brezina, and Tony McEnery, "The trinity lancaster corpus: Development, description and application," *International Journal of Learner Corpus Research*, vol. 5, pp. 126–158, 01 2019.

[14] Shin'ichiro Ishikawa, "The icnale spoken dialogue: A new dataset for the study of asian learners' performance in l2 english interviews," *ENGLISH TEACHING*, vol. 74, pp. 153–177, 12 2019.

[15] Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, "The NICT JLE Corpus," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[16] Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry, and Evelina Galaczi, "Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language: A Mixed-Methods Study," *Language Assessment Quarterly*, vol. 14, no. 1, pp. 1–18, 2017.

[17] Andrew F. Hayes and Klaus Krippendorff, "Answering the Call for a Standard Reliability Measure for Coding Data," *Communication Methods and Measures*, vol. 1, no. 1, pp. 77–89, 2007.

[18] Klause Zechner and Keelan Evanini, *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, Routledge, 2020.

[19] Satoshi Kobashikawa, Atushi Odakura, Takao Nakamura, Takeshi Mori, Kimitaka Endo, Takafumi Moriya, Ryo Masumura, Yushi Aono, and Nobuaki Minematsu, "Does Speaking Training Application with Speech Recognition Motivate Junior High School Students in Actual Classroom? – A Case Study," in *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 2019, pp. 119–123.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[21] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[22] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face and Gesture Recognition*.