Multimodal Analysis for Communication Skill and Self-Efficacy Level Estimation in Job Interview Scenario

Tomoya Ohba*
s2110034@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan

Candy Olivia Mawalim*
candylim@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan

Shun Katada
s2040005@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan

Haruki Kuroki
hkuroki0204@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan

Shogo Okada†
okada-s@jaist.ac.jp
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan

ABSTRACT

An interview for a job recruiting process requires applicants to demonstrate their communication skills. Interviewees sometimes become nervous about the interview because interviewees themselves do not know their assessed score. This study investigates the relationship between the communication skill (CS) and the self-efficacy level (SE) of interviewees through multimodal modeling. We also clarify the difference between effective features in the prediction of CS and SE labels. For this purpose, we collect a novel multimodal job interview data corpus by using a job interview agent system where users experience the interview using a virtual reality head-mounted display (VR-HMD). The data corpus includes annotations of CS by third-party experts and SE annotations by the interviewees. The data corpus also includes various kinds of multimodal data, including audio, biological (i.e., physiological), gaze, and language data. We present two types of regression models, linear regression and sequential-based regression models, to predict CS, SE, and the gap (GA) between skill and self-efficacy. Finally, we report that the model with acoustic, gaze, and linguistic features has the best regression accuracy in CS prediction (correlation coefficient $r = 0.637$). Furthermore, the regression model with biological features achieves the best accuracy in SE prediction ($r = 0.330$).

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

communication skill, self-efficacy, multimodal modeling, job interview, virtual agent

ACM Reference Format:
Tomoya Ohba, Candy Olivia Mawalim, Shun Katada, Haruki Kuroki, and Shogo Okada. 2022. Multimodal Analysis for Communication Skill and Self-Efficacy Level Estimation in Job Interview Scenario. In *21th International Conference on Mobile and Ubiquitous Multimedia (MUM 2022), November 27–30, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3568444.3568461

1 INTRODUCTION

Interpersonal communication skills are required for successful communication in our global and multicultural society. Accordingly, these skills are essential to many areas of life, including education, business management, and daily life. They are commonly considered in determining a candidate's compatibility during the job recruiting process. According to previous social science research [12], various types of skills, including nonverbal communication, message production, reception skills, and functional abilities, such as informing, explaining, arguing, and persuading, are required in specific communication situations. Several studies have focused on automatically assessing skills during job interviews using various features to develop automatic assessment models to predict the scores assigned by human expert raters. Automatic job interview systems support the hiring process in the human resource department and social skill training systems that provide automatic feedback. Hence, communication skill (CS) training systems are important for educating students and young business people.

External experts are generally hired to assess communication skill annotation for constructing automatic job interview systems. Consequently, interviewees themselves do not know their assessed score. Thus, even if an interviewee feels that an interview is successful, the actual rating might have a low score. As a result, this overestimation or underestimation of their skill level affects how the interviewee responds to feedback comments received from the interview training systems. This finding is considered to be useful for interviewees if systems automatically detected anxious or low self-efficacy user states during interviews [26].

[*] These two authors contributed equally to this work.
[†] Shogo Okada is the corresponding author.

MUM 2022, November 27–30, 2022, Lisbon, Portugal

ACM ISBN 978-1-4503-9820-6/22/11…$15.00
https://doi.org/10.1145/3568444.3568461

Self-efficacy is defined by Bandura [1] as "a person's particular set of beliefs that determine how well one can execute a plan of action in prospective situations". Due to its significant importance in job search and interview processes, several theoretical models have been proposed for self-efficacy [30, 50]. The person's belief in his efficacy could provide motivation, well-being, and numerous benefits in daily life, including his job interview process [22, 32]. Moreover, self-efficacy could help people overcome failure because they rely more on handling than worrying about it [1]. Consequently, building self-efficacy is essential, and research has shown a positive association between self-efficacy and feedback [1, 32, 57].

The automatic assessment of both the user's skill level and self-efficacy allows us to detect users with anxiety or low confidence. Subsequently, the system helped interviewees (trainees) calibrate discrepancies and gaps (GAs) between their skills and self-efficacy level (SE). Accordingly, the automatic assessment system is beneficial for generating adaptive feedback comments for users (e.g., a job interview feedback system) [45]. The job interview feedback system can provide more appropriate suggestions for users based on GA estimation. For instance, in the case of highly skilled users with low self-efficacy, the feedback system can suggest improving their self-efficacy, which could lead to a better impression or encouragement in the actual job interview.

This study presents a computational analysis of the relationship among multimodal features, the interviewee's skill and the interviewee's self-efficacy in the interview setting by using linear regression models and sequential deep neural regression models to predict the CS, SE, and GA between actual skill and self-efficacy. The benefit of this study on the employee's side in the job interview scenario is that it can be used to generate automatic evaluations within the interview training feedback system in the job search process. Meanwhile, on the employer's side, the system can automate low-level interviews (e.g., first-round interviews) to save time and reduce the blurring of ratings due to interviewer subjectivity. To address this challenge, we use a virtual agent system [55] with a virtual reality head-mounted display (VR-HMD) to collect a novel multimodal data corpus during interviews. The data corpus includes annotations of CS by third-party experts and SE by the interviewees. The data corpus also includes various kinds of multimodal data, including not only audio and language data, which have been used in previous works but also biological data and gaze data. Finally, we report the accuracy of the prediction model in estimating skill, self-efficacy and the GA.

The contributions of this study are summarized as follows:

(1) Multimodal interview data corpus including self-efficacy level and various types of nonverbal data: We develop a novel multimodal data corpus including human (interviewee)-agent (interviewer) interactions in a job interview training scenario. The data corpus includes CS labels and SE labels, which are annotated for each question-answer pair. Furthermore, the data corpus includes not only audio and language data but also biological data, which are obtained with wristband-type sensors, and gaze data, which are captured by eye trackers, to investigate the relation with the SE. These various types of multimodal data allow us to analyze the relationship among skill level, SE, and human behavior. The details of annotations are described in Section 3.

(2) Multimodal modeling communication skill and self-efficacy: Previous works have focused on predicting interview CS and hireability with external expert annotations. In contrast, we develop a computational model to estimate CS (third-party impression) and SE, as well as the GA between skill and self-efficacy. Feedback on this estimated GA can assist interviewees in determining their current skill level and calibrating their self-awareness. To the best of our knowledge, this study is the first investigation of multimodal features that display the skill, SE, and GA during job interview training. We report the accuracy of predicting the CS, SE and GA in Section 5.2.

(3) Analysis of the effectiveness of different features and regression performances: We analyze the prediction performance for each question and determine the question type for which the skill, SE, and GA are predicted by the trained model with better or worse accuracy in Section 6.2. In addition to the analysis for the question type, we clarify the specific feature group for contributing prediction tasks in Section 6.1.

2 RELATED WORK

The research on CS estimation can be classified based on different communication situations. Studies exist that focus on CS in a monologue situation, including public speaking [3, 46, 56], and social media [44]. Other directions for research studies include modeling CS in dyadic interactions, including job interview settings [38, 40], group interaction situations [20, 35, 43, 51], and human-computer (including robot and virtual humans) interactions [18, 52, 53]. Rasipuram et al. conducted a comprehensive survey of recent research progress on CS assessment technology [48]. The current research is related to the human-computer interaction in a job interview setting; thus, we focus on introducing works related to this topic in this section.

2.1 Automatic skill assessment

Nguyen et al. [40] extracted multimodality and interaction (relational) features rather than single-modality features, which included mutual gazing and speaking gestures that are predefined manually, to infer expert-coded hireability scores. They also analyzed thin slices of these interviews and demonstrated that thin slices were sufficient to predict the hireability scores [41]. Okada et al. [42] proposed a co-occurrence event-mining framework to explicitly extract the intermodal and interperson features (e.g., speaking with/without gestures) for multimodal dyadic data. The study also reported that the framework improved the classification accuracy of the hireability label on the data corpus in [40]. Naim et al. [37, 38] proposed a model with verbal and nonverbal audio-visual features for predicting job interview performance along with 16 different social traits, such as excitement and engagement. Li et al. [31] proposed a hierarchical coupled hidden Markov model to capture the synchronization of the facial expressions of two participants to infer conversation outcomes. The research shows that synchronized nonverbal templates contribute to predicting negotiation outcomes.

Rasipuram et al. [47] automatically assessed the CS of participants using verbal and nonverbal behavioral cues in asynchronous video interviews and face-to-face interviews. The study revealed that the assessment of CS could be performed with video interviews alone without human interviewers. Skanda et al. [36] presented a multimodal analysis for automatically assessing various social variables, such as professional skills, social skills, communication skills, and overall impression in hospitality, using nonverbal features. Chen et al. [5] automatically predicted the Big Five dimensions along with the overall hiring recommendations of participants giving video interviews. The automatic framework involved the extraction of verbal and nonverbal multimodal behavioral cues and manual annotation by experts. Subsequently, Chen et al. [4] also conducted a study to automatically assess job interview performance and oral presentation performance in monologue video interviews. Hemamou et al. [15] collected a corpus of more than 7000 candidates having asynchronous video job interviews for real positions. They proposed a hierarchical attention model called HireNet for predicting the hireability of candidates as evaluated by recruiters.

2.2 The virtual agent and robot as a job interviewer

CS assessment and training with a virtual agent are well studied. For instance, My Automated Conversational Coach (MACH) is a platform that provides social skills training with a virtual agent [18]. Other virtual agent systems are also proposed for social skill training [11, 52]. All these platforms provided a graphical representation by displaying sensing results of multimodal behaviors, and the behavioral qualities must be improved. A virtual interviewer should have interview techniques similar to those of a human interviewer who can cope with nonverbal information. For that reason, Sabouret et al. [49] developed a model that enables the virtual agent to adapt its social attitude during the interaction with the user in the context of a job interview. Meanwhile, Inoue et al. [19] developed an interview dialog system that generates follow-up questions based on the speech recognition of the interviewee's response.

2.3 The focus of this study

Most existing works have focused on automatically assessing impression scores by external coders. In other words, they have not focused on the self-reported inner state of interviewees, which has been found to be beneficial for social skills training. A few works on investigations of interviewees' inner states are as follows. Kimani et al. [26] focused on presenter anxiety in public speaking. They presented a virtual coach that uses cognitive behavioral therapy techniques to help presenters restructure irrational thoughts associated with public speaking anxiety and show the effectiveness through subjective experiments.

Furthermore, Kimani et al. [25] also proposed a virtual agent system to assist presenters in managing their anxiety in real time during presentations. They evaluated the automated real-time framework for detecting public speaking anxiety with the collected dataset. Inspired by this research that analyzed the psychological aspect of the job interview and public speaking setting, we investigate the relationship between skill, self-efficacy, and GA through multimodal regression modeling to predict each target. To the best of our knowledge, this is a pioneering work in that it explicitly investigates both skill scores and self-efficacy in a job interview setting.

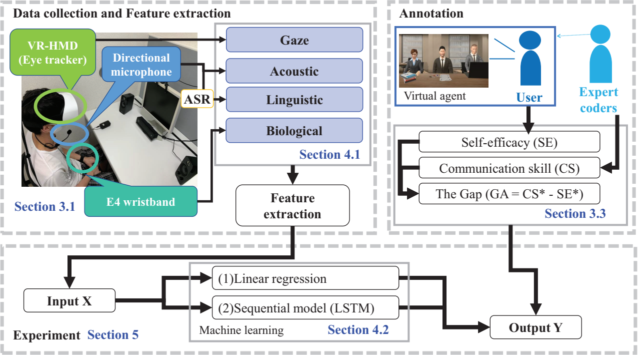


Figure 1: Overview of job interview agent system and multimodal modeling

3 DATASET

In this study, we collected a novel Japanese interview dataset, namely, the Multimodal Virtual Job Interview (MuViJI) corpus. This dataset includes audio, eye gaze, language, and video data to analyze the skills and SE of interviewees. In Section 3.1, we describe in detail the job interview agent system [55] used to develop the dataset. In Sections 3.2 and 3.3, we describe the design and structure of the dataset.

3.1 Job interview agent system

We developed a job interview system with multimodal sensing modules by using the Unity developmental platform. The virtual agent system was connected to a VR-HMD, namely, FOVE0[1], and interviewees virtually experienced the interview in 3D space. Such VR experiences are useful for providing realistic job interview experiences. The interview system is shown in Fig. 1.

3.1.1 Implementation of embodied conversational agent. We assumed that interviews with multiple interviewers were representative of a typical job interview situation. There were three interviewer agents in our interview system. In the interview session, only the male agent in the center spoke and asked questions.

Interview conversation: The agent asked each question based on a predefined question list. Specific voice data and the intonation of the agent were predefined for each question. A system operator manually determined the initial timing of the agent's utterance to prevent errors in the voice activity detection (VAD) of the interviewees. The development of a complete automatic dialog management module is a direction for future work. To implement natural conversation, a back-channel utterance (e.g., “Very well, now for our next question...”) was inserted before the next question.

Synthesis of lip motion: To improve the reality of the virtual agents in the proposed system, we also implemented a module to synthesize speaking lip animations (lip-sync) for virtual agents in real time using Oculus Lipsync Unity[2]. This module is not used on the data collecting process in this study.

[1] FOVE Co., Ltd, Minato-ku, Tokyo, Japan, https://fove-inc.com/
[2] https://developer.oculus.com/downloads/package/oculus-lipsync-unity/

3.1.2 Sensing environment. The proposed VR-HMD interview agent system has a multimodal sensing module. To investigate the skills and SE, we collected various types of multimodal data. First, clean audio data were captured by using a headset microphone (Shure). The participant's voice was recorded as 16-kHz Waveform Audio Format (WAV) files. Second, we recorded biological signals that captured involuntary changes related to the participant's inner state and SE by using an Empatica E4 wristband (Empatica Inc., Cambridge, MA, USA). The biological signal data included the heart rate (HR) and electrodermal activity (EDA) recorded with the E4 wristband. Third, we recorded the gaze data using an optical eye tracker that is included in the VR-HMD (FOVE0). Gaze activity can indicate nervous states in people [54].

3.2 Participants

The study included 41 participants (29 males and 12 females). The participants' ages ranged from 20 to 30 years. They were either about to start or had experienced an actual job interview. The recruited participants had fluent speaking Japanese proficiency because the question list was developed based on a job interview scenario in typical Japanese companies. The job interview scenario was designed by Japanese experts who are well experienced as job interviewers. A research ethics committee reviewed and approved the collection of data and the corresponding research using this dataset. The subjects agreed to the CS assessment and the experimental design before participating in the experiment. After the experiment, the subjects were asked to fill out a self-evaluation questionnaire using a 7-point Likert scale to indicate their SE when answering the 13 questions and during the overall interview. The 13 interview dialog questions are listed in Table 1. Each question was carefully designed to prevent a simple one-question/one-answer format and to ensure that each dialog could be evaluated as a sample. The content of the interview was created under the guidance of professional career counselors based on typical first interviews at various companies. Finally, 41 sessions of experimental data were obtained, averaging about 13 minutes and 35 seconds per session. The total video duration was approximately 9 hours and 16 minutes.

3.3 Annotations

The CS were annotated by expert coders, and the SE was annotated by interviewees themselves. The annotations are summarized in Table 2. In addition to the CS and SE, we analyzed the GA between the skill and self-efficacy. The skill and self-efficacy were annotated for each pair of questions and answers in each interview session. Because we collected 13 question-answers during 41 interview sessions, our dataset included a total of 533 samples (13 × 41) with annotations. The data distribution of the three annotation labels is shown in Fig. 2.

The Pearson correlation coefficient between the **CS** and the **SE** was $r = 0.124$; coefficients with a magnitude of less than 0.3 have little if any correlation.

This means that the skill level (**CS**) annotated by the experts deviates from self-efficacy (**SE**). From these results, we analyzed the differences in CS, SE, and GA by using multimodal features.

Table 1: Overview of the 13 questions used in the experiment.

ID	Summary of Questions
1	Could you introduce yourself in a duration of approximately 1 minute?
2	What did you put the most effort into during your college years ?
3	What have been your difficulties or challenges ? and how did you overcome them ?
4	What academic topic do you research at the university ?
5	Could you explain what you consider important in selecting your job ?
6	What industry would like to work in ?, and why did you want to work in the industry ?
7	What is your first choice of company ?, and why do you want to work for the company?
8	What type of job would you like to engage in at the company? What do you want to engage in on the job ?
9	Could you explain your strengths and the reason for them based on your own experience ?
10	Could you explain your weakness and the reason for them based on your own experience ?
11	How do people around you evaluate you and your personality ?, and could you explain the evidence ?
12	What role do you most often play in collaborative work ?, and could you explain the reason and any memorable episodes in your experience ?
13	What do you consider important in your interpersonal relationships ?

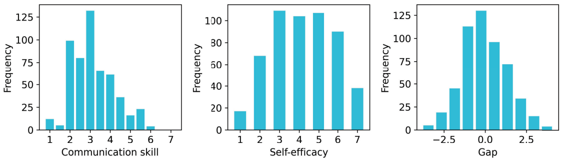


Figure 2: Data distribution

Communication skill: We asked two expert job interview trainers to assess the skills, and the two expert coders annotated the CS. The two expert coders had more than 10 years of experience in training university students in job interview training programs. Verbal and nonverbal skills are both required for successful job interviews, and these skills are composed of four items, including **engagement**, **voice**, **vocabulary**, and **logic**, which are defined in Table 2. The overall skill level (**Total**) was assessed based on the scores of the four indices, and we set the average of the total skill score by the two experts as the target variable in our machine learning model. The skill levels of each index were evaluated using a 7-point Likert scale.

We calculated the agreement between the two coders for CS labels using Krippendorff's alpha (α_k) [28] and Cronbach's alpha (α_c) in Table 3. The agreement values of α_k and α_c were more than 0.874 and 0.745, respectively. According to [6], $\alpha_c > 0.8$ is acceptable as an agreement level; thus, the score ($\alpha_c > 0.874$) denotes acceptable agreement.

Self-efficacy: We asked the participants to score their self-efficacy in answering each question in a postinterview session.

Table 2: Annotations for communication skills and self-efficacy.

Engagement	Interviewee responded to questions in a positive manner.
Voice	Interviewee spoke fluently with appropriate voice volume and speaking speed.
Vocabulary	Interviewee explains her (his) opinion with sufficient vocabulary.
Logic	Interviewee answered the question in a logical manner.
Total	Interviewee has sufficient communication skills. (Overall score by considering scale scores of engagement, voice, vocabulary and logic)
Self-efficacy	Rate your degree of confidence in answering each interview question.

Table 3: Agreement between annotators for communication skills' label (α_c and α_k denote Krippendorff's alpha and Cronbach's alpha, respectively.)

Item	α_c	α_k
Engagement	0.903	0.785
Voice	0.905	0.815
Vocabulary	0.923	0.848
Logical	0.942	0.858
Total	0.874	0.745

Self-efficacy indicates their belief in their capabilities in answering a question during the interview. Depending upon the task, there are numerous ways to assess self-efficacy. For instance, the General Self-Efficacy Scale (GSES) by Bandura [2] aims to assess self-efficacy based on four primary sources, including mastery experience, social modeling, social persuasion, and psychological responses. Prior research has used a traditional 0–100 scale or a Likert scale to assess self-efficacy. However, the study of Maurer and Andrews [34] shows that traditional, Likert, and even simplified scales for self-efficacy are strongly related, valid, and reliable with an adequate number of participants and purposes. Our purpose in this research is to obtain the participants' belief that they can perform well in answering an interview question (a high-level task). Accordingly, a 7-scale rating system of more than 35 participants is considerably adequate for this study [34]. The 7-scale self-efficacy rating that we utilized ranges from 1 (cannot do at all) to 7 (highly certain can do).

Gap between skill and self-efficacy: In the job interview scenario, subjects might tend to overestimate or underestimate their skills (a kind of cognitive bias). Cultural background has significant influences on confidence and self-efficacy [7, 33]. For instance, Heine [13, 14] found that individuals from East Asian, including Japanese, tend to underestimate or show hesitation in their skill level (low self-efficacy). Since this study utilized a Japanese job interview dataset, the GA between skill y_1 and self-efficacy y_2 of each interviewee was expected (we name it the "gap" y_3 in the following sections). Estimating the GA can further help the job interview feedback system provide more appropriate suggestions for promoting self-efficacy, such as through encouragement. We developed a prediction model of the GA that is capable of detecting the specific question-answer pair in which interviewees have overestimated or underestimated self-efficacy. To calculate the gap y_3, we normalized y_1 and y_2 of all samples such that y_1 and y_2 have a zero mean and one standard deviation (z score method). Next, we calculated y_3 as $y_3 = y_1^* - y_2^*$.

4 METHODS

The objective of this study is to investigate the relationship among multimodal features, skill, self-efficacy, and GA. To accomplish this objective, we developed multimodal regression models to predict labels including (y_1) the CS, (y_2) the SE, and (y_3) the GA between (y_1^*) and (y_2^*). By evaluating the regression accuracy of the trained models and the ablation test of each modality group, we can investigate effective multimodal features for predicting these labels.

4.1 Multimodal feature extraction

To develop regression models for predicting the CS and SE, we extracted both nonverbal features, such as acoustic features, biological features, and gaze features, and linguistic features, such as word-level features, based on the spoken utterances of the interviewees. Table 4 shows the summary of the extracted multimodal features in several subgroups. As preprocessing, we normalized all features using z score normalization so that the mean became 0 and the standard deviation became 1 for all samples.

4.1.1 *Acoustic feature extraction.*

OpenSMILE [10] was used to extract acoustic features. The "eGeMAPS" was used as the configuration file for the extracted features [9]. These acoustic features are often used as minimalistic standard features for speech analysis. The extracted acoustic features are comprised of frequency-related parameters, energy-related parameters, spectral parameters, temporal features, and some extended parameters. The acoustic features of each frame were calculated, and the frames were shifted from a frame width of 60 ms with a sliding width of 10 ms. Then, statistics, such as the averages, were calculated between frames, resulting in a total of 88 dimensions. The summary of acoustic features is shown in Table 4.

4.1.2 *Biological feature extraction.*

The biological features were extracted with the same procedure discussed in [23, 24]. In brief, the skin conductance (SC) signals recorded using the E4 wristband were decomposed into a tonic component (SC level) and phasic component. The tonic component was calculated by fitting the SC signals with a polynomial degree of 10, and the phasic component was calculated as the difference between the tonic component and the SC signal. PeakUtils[3] was used to detect the galvanic skin response (GSR) with an amplitude threshold of 0.3. The statistics of the SC signal and HR data for each question, such as the mean, standard deviation, and skewness, were calculated. Overall, 27 features were extracted as the biological features for each question.

[3] https://pypi.org/project/PeakUtils/

4.1.3 Gaze feature extraction. The gaze feature extraction process yielded a total of 18-dimensional feature values, including the mean and variance values of the x (right/left), y (up/down), and z (back/front) axes in the gaze direction of both the left and right eyes, which were obtained with FOVE0. In addition, we calculated the time length while the user looked at the agent's head and body. These values were normalized by the question duration so that a two-dimensional feature was obtained. Consequently, we obtained a total of 20-dimensional features as gaze features.

4.1.4 Linguistic feature extraction. We used the Speech-to-Text Application Programming Interface (API) of the Google Cloud Platform to obtain speech recognition results from the WAV files collected during the experiment. Based on these speech data, linguistic features were obtained using the methods. The extracted linguistic feature has a total of 780 dimensions.

Part-of-speech (PoS) and polarity features: The text was segmented into words using the Japanese morphological analysis tool MeCab [29]. The linguistic features extracted from each participant's utterances included the word frequencies (the number of words, nouns, proper nouns, verbs, conjunctions, adjectives, adverbs, interjection, and fillers) based on the word PoS frequencies. The polarities of each participant's utterances (positive, negative, or neutral) were analyzed using OSETI[4], which is a Japanese sentiment analyzer based on a sentiment polarity dictionary [16, 27]. The polarities were used to evaluate the participants' utterances, and the polarity score [-1, 1] and the number of positive and negative words were used as features. The total number of dimensions was 12.

Bidirectional encoder representations from transformers (BERT): BERT is a language representation model that has achieved state-of-the-art performance on various natural language processing (NLP) tasks [8]. We used the pretrained Japanese BERT model [5]. We also used this model for extracting BERT features. The speech sequences were first tokenized by MeCab and split into subwords with the WordPiece algorithm. Next, the activations were extracted from the second-to-last hidden layer of the BERT model, and the sequence was represented by average pooling, resulting in a single vector with a length of 768, as described in [8]. This vector was used as one group of linguistic features for the regression model.

4.2 Regression models

We prepared two types of regression models for training the collected dataset for question-answer (QA) pairs.

4.2.1 Linear regression models. In the first approach, we regarded 13 samples (corresponding to 13 QAs) observed from an interviewee as 13 independent samples, and we trained the independent samples using three typical linear models. As typical models, we used three regression models: ridge regression, least absolute shrinkage and selection operator (Lasso) regression, and support vector regression (SVR) for the regression task. The source of QA types is important information from which each sample is obtained. To make models distinguish specific features for each question type, we modeled the question sequence by concatenating the one-hot vector encoding of the question type to feature vectors. Subsequently, we trained the linear models with the concatenated feature vector.

Table 4: Summary of Multimodal Features.

Modality	Group Name	Description
Acoustic (**A**)	A1_Freq	Frequency-related parameters
	A2_Energy	Energy/amplitude-related parameters
	A3_Spectral	Spectral (balance) parameters
	A4_Temporal	Temporal features
	A5_Extended	Extended parameter set
Biological (**B**)	B1_EDA	Statistics from the SC signal and GSR number
	B2_HR	Statistics from the HR data
Gaze (**G**)	G1_Dir	Statistics of eye movement on the x-, y-, and z-axes
	G2_Count	The frequency when the user looked at the interview agents
Linguistic (**L**)	L1_BERT	Activation of hidden units of BERT
	L2_PoS	Word frequencies, such as number of nouns
	L3_Polarity	Polarity score ranges [-1,1] based on the number of positive/negative words

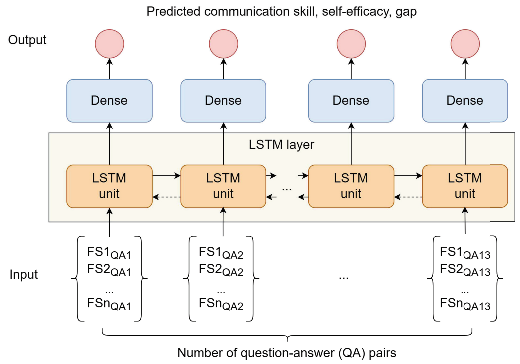


Figure 3: LSTM-based model architecture for communication skill (CS), self-efficacy level (SE), and gap (GA) prediction. The dashed line in the backward arrows shows the case when BiLSTM model is used.

4.2.2 Sequential models. In the second approach, we regarded 13 samples (corresponding to 13 QAs) observed from an interviewee as 13 sequence samples obtained from time-series QAs. We constructed sequential models for predicting the CS and SE. The sequential models are constructed based on long short-term memory (LSTM) models. Fig. 3 shows an illustration of our LSTM-based prediction model. These models should mimic the job interview question-answer process because the question-answer sequence is reflected in the LSTM sequence. Hence, we may obtain more information from earlier or later question-answer pairs to precisely predict the skill and SE of the present QA pair. Additionally, we compared the unidirectional and bidirectional LSTM (BiLSTM) models in the experiments. We considered BiLSTM because BiLSTM is an extended LSTM that enables the model to learn sequential patterns from both directions (backward and forward). In a job interview scenario, we predicted that the context in both the past and future would influence the skill and SE of each interviewee.

[4] https://github.com/ikegami-yukino/oseti
[5] https://github.com/cl-tohoku/bert-japanese

Table 5: Comparison of the best prediction accuracy among regression models (CS, SE, GA denote communication skill, self-efficacy level, and gap, respectively.)

Target	Lasso		Ridge		SVR		LSTM		BiLSTM	
	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE
CS	0.492	0.930	0.569	0.859	0.394	1.318	0.601	0.846	**0.637**	**0.808**
	A+L		A+G+L		A+G		A+G+L		A+G+L	
SE	0.295	1.580	0.169	1.641	0.211	1.779	0.296	1.638	**0.330**	**1.596**
	A+B		A+B+L		A+G+L		A+B+L		B	
GA	**0.386**	**1.255**	0.277	1.282	0.273	1.464	0.354	1.353	0.385	1.283
	A+L		L		B+G+L		B+L		A+L	

5 EXPERIMENTS

5.1 Experimental setting

5.1.1 Hyperparameter settings and evaluation. The Lasso parameter in the Lasso regression model was optimized using a 5-fold cross-validation scheme, with the alpha parameter values selected from [0.001, 0.01, 0.1, 1, 10, 100]. The ridge parameter in the ridge regression model was optimized using a 3-fold-cross-validation scheme, with the alpha parameter values selected from [0.001, 0.01, 0.1, 1, 10, 100]. The parameters of the linear SVR were optimized similarly using a nested cross-validation scheme, with the C and epsilon parameter values selected from [0.001, 0.01, 0.1, 1, 10, 100].

For the LSTM-based models, we utilized the adaptive moment estimation (Adam) stochastic optimizer with the mean square error (MSE) as the loss function. Instead of using specific time-step input data, we modified the time step as the question-answer sequence, resulting in a length of thirteen time steps. Furthermore, we controlled the parameters to improve the estimation results, such as the number of LSTM units $N_{LSTM} = 20$, and the batch size was set to 16. In the experiments, we used pipelining to determine which number of epochs led to the best performance, with a range at $N_{epochs} = \{100, 150, 200, 250\}$.

To evaluate the regression performance, we used the Pearson correlation coefficient (r) and root mean square error ($RMSE$) as performance criteria. In the experiments presented below, we used leave-one person (interviewee)-out cross-validation and report the average performance (r and $RMSE$) over all folds.

5.2 Results

In this section, we discuss the prediction results for all multimodal combinations of the four modalities (Acoustic (**A**), Biological (**B**), Gaze (**G**), Linguistic (**L**)). We show the comparison of regression accuracy in Table 5 and Table 6 for CS, SE, and GA. In some cases, the correlation coefficients were negative values ($r < 0$), which means that the ordinal relationship of the target score among samples could not be modeled well. We filled $r = 0.000$ in such failure cases of the regression modeling, as shown in Table 6. We mainly discuss the regression performance with correlation coefficients r by considering that the comparison of ordinal relationships is important to compare the CS, SE, and GA levels.

Table 5 shows the comparison of the best prediction accuracy among regression models. We discuss the best regression model for each prediction task and analyze the effectiveness of multimodal fusion in the best regression model by correlation (r). According to Table 5, the best and the second-best models for CS were BiLSTM (r = 0.637) and LSTM (r = 0.601), respectively. The best and the second-best model in correlation r for SE were also BiLSTM (r = 0.330) and LSTM (r = 0.296), respectively. The best and the second-best model in correlation r for GA were Lasso (r = 0.386) and BiLSTM (r = 0.385), respectively. From these results, LSTM-based sequence models are suitable for the prediction of CS and SE. However, Lasso regression is suitable for the prediction of GA. In the following sections, we analyze the relationship between prediction accuracy and multimodal or unimodal features in the best models. Table 6 shows the prediction accuracy obtained by the best sequential model and the best linear model in Table 5.

5.2.1 Prediction accuracy of communication skill. According to the comparison of unimodal models in Table 6 for CS (columns 3-6), in both the BiLSTM and ridge models, acoustic and linguistic features are effective in improving the accuracy in the four modalities. In particular, the linear regression model (ridge) outperformed BiLSTM in cases using acoustic features (r = 0.502) and linguistic features (r = 0.450). The linguistic usage of interviewees can display CS. The effectiveness of the linguistic features in estimating user's internal states is coincident with studies such as [17]. In the comparison of multimodal models (modals 2, 3, and 4 in Table 6), the BiLSTM model with A+G+L had the best accuracy (r = 0.637) for the prediction of CS. Ridge regression also achieved the best accuracy (r = 0.569) with equal multimodal features: A+G+L.

These results show that acoustic and gaze features effectively fuse with linguistic features. This finding is aligned with prior work that concluded that acoustic and linguistic features are effective in predicting skill [15]. Although annotators cannot observe eye movement of interviewees with VR-HMD, the results indicate that gaze activity in HMD is related to the skill. Hence, sensing gaze activity can effectively improve the skill in an interview setting with VR-HMD.

5.2.2 Prediction accuracy of self-efficacy. According to Table 6, the BiLSTM model with B had the best correlation (r = 0.330) for the SE. In the results of the Lasso regression model, the multimodal model (A+B) achieved the best correlation (r = 0.295). As a common effective modality, the biological feature set is most effective. Conversely, linguistic features are relatively ineffective, so the best feature set differs significantly from the CS. In particular, the best unimodal model was BiLSTM with the biological feature set (r = 0.330).

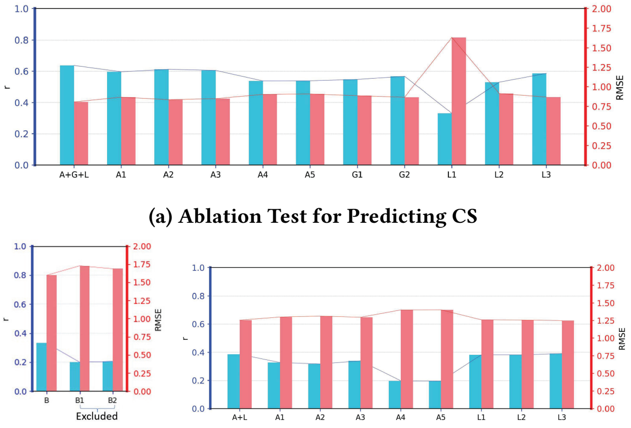
5.2.3 Prediction accuracy of the gap. According to the comparison of unimodal models in Table 6 for GA (columns 11-14), acoustic and linguistic features are effective in improving the accuracy in four modalities; i.e., the best correlation was achieved with A (Lasso:r = 0.355, BiLSTM:r = 0.286), and the second-best correlation was obtained with L (Lasso:r = 0.140, BiLSTM:r = 0.224). The results that A and L are effective are aligned with the results for CS. In the comparison of multimodal models, the Lasso model trained by the multimodal feature set A+L has the best accuracy (r = 0.386) for GA. In addition to Lasso, the best accuracy of BiLSTM is also consistently achieved with A+L (r = 0.385). GA can be predicted using acoustic and linguistic features, so the result is similar to that of CS.

6 DISCUSSION

In this section, we discuss the effective feature groups using ablation tests in Section 6.1, the prediction accuracy for questions in Section

Table 6: Prediction accuracy of unimodal and multimodal models (The results are obtained by the best ML methods in Table 5.)

Target prediction		CS				SE				GA			
Modality		Linear (Ridge)		Sequential (BiLSTM)		Linear (Lasso)		Sequential (BiLSTM)		Linear (Lasso)		Sequential (BiLSTM)	
		r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE
4 modals	A+B+G+L	0.562	0.866	0.583	0.864	0.166	1.682	0.117	1.782	0.285	1.345	0.327	1.312
3 modals	A+B+G	0.471	0.987	0.351	1.540	0.161	1.706	0.125	2.288	0.247	1.383	0.186	1.387
	A+B+L	0.537	0.884	0.523	0.910	0.267	1.608	0.207	1.722	0.375	1.267	0.339	1.323
	A+G+L	**0.569**	**0.859**	**0.637**	**0.808**	0.145	1.692	0.106	1.853	0.298	1.327	0.288	1.350
	B+G+L	0.435	0.941	0.394	1.029	0.710	1.595	0.033	1.881	0.061	1.340	0.339	1.307
2 modals	A+B	0.487	0.975	0.302	1.360	**0.295**	**1.580**	0.184	2.143	0.345	1.297	0.217	1.371
	A+G	0.501	0.947	0.317	1.741	0.147	1.704	0.118	2.348	0.246	1.377	0.078	1.469
	A+L	0.545	0.877	0.534	0.898	0.279	1.581	0.210	1.717	**0.386**	**1.255**	**0.385**	**1.283**
	B+G	0.000	1.091	0.000	1.356	0.000	1.595	0.087	1.962	0.000	1.371	0.003	1.555
	B+L	0.454	0.926	0.000	1.213	0.000	1.595	0.148	1.789	0.061	1.335	0.228	1.425
	G+L	0.432	0.941	0.403	1.016	0.000	1.622	0.000	2.084	0.011	1.366	0.231	1.487
unimodal	A	0.502	0.954	0.263	1.511	0.290	1.575	0.172	2.262	0.355	1.286	0.286	1.331
	B	0.000	1.055	0.000	1.205	0.000	1.595	**0.330**	**1.596**	0.000	1.344	0.145	1.430
	G	0.000	1.114	0.023	1.447	0.000	1.763	0.000	2.269	0.000	1.408	0.000	1.623
	L	0.450	0.926	0.423	1.003	0.000	1.596	0.029	1.915	0.140	1.359	0.224	1.457



(a) Ablation Test for Predicting CS

(b) Ablation Test for Predicting SE

(c) Ablation Test for Predicting GA

Figure 4: Results of Ablation Test. The most right bars show the baseline (the best prediction) model of each target prediction. The rest bars show the prediction results by excluding a specific feature group (as shown in Table 4)

6.2, and the impact of biological and gaze features on prediction tasks of CS and SE in Section 6.3.

6.1 Feature analysis with ablation test

In the ablation test, the regression model was trained using multimodal features by removing features of a specific feature group, and it was evaluated using the Pearson correlation coefficient (r). We analyzed the contribution of each feature group (as shown in Table 4) in the best model to predict the three targets using ablation tests. The contribution of the feature group was identified by comparing the classification accuracy of the best model with the accuracy of feature sets that exclude that specific feature group.

The excluded feature set was considered effective if the accuracy (correlation coefficient) was degraded. On the other hand, the excluded feature set was considered unnecessary if the accuracy improved. Fig. 4 shows the regression accuracies of the model that excludes specific features for CS, SE, and GA. In these tables, r denotes the accuracy (correlation coefficient) of the prediction in the test data.

Communication skill: The experimental results show that CS was the best predicted by the BiLSTM model with the feature set of A+G+L. The results of the ablation test in Fig. 4a show that the BERT-based feature (L1) in linguistic features is the most effective group in the prediction of CS. Compared to the model that used every feature, the model without L1 dropped its score (r) by 0.305. The next most influential feature group was the PoS group (L2), suggesting that features related to language were significant. There was no significant difference between the groups of features related to acoustics and gaze. Nevertheless, in light of the results for Table 6, it is possible to consider that those acoustic features contributed more than gaze features.

Self-efficacy: The results of the ablation test are summarized in Fig. 4b. We set a baseline as the BiLSTM model with feature B, which was the best predictor of SE in the experimental results. The ablation test shows that two feature groups composed of B (B1 and B2) contribute almost equally to the SE prediction. Combining both B1 and B2 can improve the prediction (from $r \approx 0.2$ to $r = 0.330$). The results suggest that features related to biological features are important.

Gap between skill and self-efficacy: The experimental results show that the Lasso regression model using feature A+L had the best GA accuracy. The results of the ablation test in Fig. 4c show that the temporal feature (A4) and extended feature (A5) in acoustic features are the most effective group in the prediction of SE. The difference in r between these models and the model with A+L is approximately 0.19. Fig. 4b and Fig. 4c show that linguistic features do not contribute to the prediction of SE and GA.

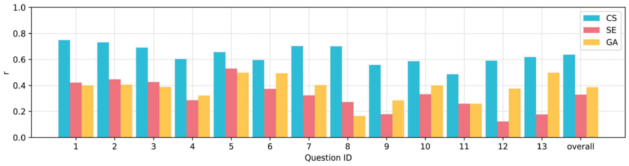


Figure 5: Prediction for question

6.2 Analysis of question type with better accuracy

We analyzed specific question types to determine which models achieved better accuracy. To develop an interview training system with detailed feedback comments, the results of the question analysis need to help interviewees understand which QAs were successful or failed. Fig. 5 shows the regression results for each question type on the best models in Table 6. Samples obtained in Question 1 (Self-introduction) were predicted with the highest accuracy in all samples for the CS (r = 0.749). Samples obtained in Question 5 (about the axis of job selection) were predicted with the highest accuracy in all samples for SE (r = 0.529). On the prediction task of the GA, the samples in Questions 5 and 13 (about the value of the interpersonal relationship) were predicted with the highest accuracy (r = 0.497). Although CS was predicted in all questions with better accuracy than 0.484, the regression accuracy of SE in Question 9 (about strength), 12 (about the role in a team), and 13 was relatively low (r is smaller than 0.179). These results imply that it was difficult to predict the SE for questions about relationships with others.

6.3 Impact of gaze and biological features

To investigate the skill judged by external experts and self-reported self-efficacy, we extracted gaze features and biological features. Although it has been found that gaze features [39] and biological features [24] are effective in social signal processing tasks, the effect of gaze and biological features in a job interview setting has not been well studied. As implied in Section 5.2.1, the gaze feature was effective in the multimodal model for predicting the CS. In the result of CS in Table 6, the difference between r = 0.637 of the best BiLSTM model with A+G+L for CS and r = 0.534 of the BiLSTM model with A+L (without G) was more than 0.1. The result shows the effectiveness of gaze features. Although annotators cannot observe eye movement of interviewees with VR-HMD, the results indicate that gaze activity in HMD is related to the skill. Hence, sensing gaze activity can effectively improve the skill in an interview setting with VR-HMD. In particular, biological features are effective for the prediction of the SE. The best unimodal model was BiLSTM with B (r = 0.330). SE was self-reported, and it is not always displayed from observable features such as linguistic features. On the other hand, the biological feature set potentially detects emotional changes by capturing biological changes derived from the autonomic nervous system (ANS). Since the ANS is involuntary, it cannot be controlled consciously. Such findings can partially explain why biological features were effective in predicting SE. These findings conclude that gaze and biological features are useful for analyzing CS and SE.

6.4 Limitation and Future Work

An important direction for future works is to explore appropriate feedback methods by using the estimated results of CS, SE, and GA. We plan to develop a job interview training system to validate the effect of the feedback based on estimated CS, SE, and GA on the interview training. A more detailed SE assessment is also worth exploring to provide a more appropriate reference for the feedback system. According to the review [21] for the effectiveness of virtual reality techniques with HMDs in education settings, many studies showed there are many situations where HMDs are useful for skills acquisition, including cognitive skills related to remembering and understanding spatial and visual information and knowledge; and affective skills related to controlling participant's emotional response to stressful or difficult situations. From these findings, we utilized the VR technique with HMD for acquiring communication skills on job interviewing. However, we did not investigate the difference in influence on user experience between the VR and an actual interview environment, so a crucial future work is to investigate influences on the data collection and machine learning process by conducting experiments on the VR environment.

The second direction for future works is to enhance the automatic assessment model by collecting more large-scale datasets. We believe that the 41 interview data sessions that we collected from a graduate school were sufficient for achieving the purpose of this work. However, a more diverse data corpus is expected for future works. We plan to recruit participants from numerous universities that majored in various kinds of fields (e.g., economics, business, engineering and science).

7 CONCLUSION

This study presented a multimodal analysis to estimate the CS, SE of interviewees, and GA. Through regression modeling, we found that the effective features are significantly different among CS, SE, and GA. For CS, the BiLSTM model with acoustic, linguistic, and gaze features yielded the best regression accuracy (correlation coefficient r = 0.637). Meanwhile, for SE, the BiLSTM model with biological features achieved the best accuracy in SE (r = 0.330). For GA, the Lasso model with acoustic and linguistic features obtained the best accuracy (r = 0.386). The results denote important findings that observable features, including acoustic and linguistic features, displayed CS assessed by external coders, and conversely (non-observable) biological features can capture self-reported self-efficacy. Appropriate feedback methods using the estimated results of CS, SE, and GA can be useful for developing a job interview training system. As future work, we explored the solution of the remaining issues in Section 6.4.

ACKNOWLEDGMENTS

This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Numbers 22H04860, 22K21304, and 22H00536) and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6). This work was also partially supported by the Research Program of "Five-star Alliance" in "NJRC Mater & Dev.".

REFERENCES

[1] Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84 (1977), 191–215.

[2] A Bandura. 2006. Guide for Constructing Self-Efficacy Scales (Revised). *Self-efficacy beliefs of adolescents* 5 (2006), 307–337.

[3] Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. 2014. Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*.

[4] Lei Chen, Gary Feng, Michelle Martin-Raugh, Chee Wee Leong, Christopher Kitchen, Su-Youn Yoon, Blair Lehman, Harrison Kell, and Chong Min Lee. 2016. Automatic Scoring of Monologue Video Interviews Using Multimodal Cues. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 32–36.

[5] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. 504–509.

[6] Jose M. Cortina. 1993. What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology* 78 (1993), 98–104.

[7] Tori DeANGELIS. 2003. Why we overestimate our competence. 34, Article 2 (2003).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing (HLT-NAACL)*.

[9] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM International Conference on Multimedia*. 1459–1462.

[11] Patrick Gebhard, Tanja Schneeberger, Elisabeth André, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius König, and Markus Langer. 2019. Serious Games for Training Social Skills in Job Interviews. *IEEE Transactions on Games* 11, 4 (2019), 340–351.

[12] John O Greene and Brant Raney Burleson. 2003. *Handbook of communication and social interaction skills*. Psychology Press.

[13] Steven Heine and Takeshi Hamamura. 2007. In Search of East Asian Self-Enhancement. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 11 (2007), 4–27.

[14] Steven Heine, Toshitake Takata, and Darrin Lehman. 2000. Beyond Self-Presentation: Evidence for Self-Criticism among Japanese. *Personality and Social Psychology Bulletin* 26 (2000), 71–78.

[15] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, Vol. 33. 573–581.

[16] Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Learning sentiment of nouns from selectional preferences of verbs and adjectives. In *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing*. 584–587.

[17] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 85–94.

[18] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 697–706.

[19] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. Job Interviewer Android with Elaborate Follow-up Question Generation. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 324–332.

[20] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2018. Analyzing Gaze Behavior and Dialogue Act During Turn-taking for Estimating Empathy Skill Level. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*.

[21] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 23, 4 (2018), 1515–1529.

[22] Ruth Kanfer, Connie Wanberg, and Tracy Kantrowitz. 2001. Job search and employment: A personality-motivational analysis and meta-analytic review. Journal of Applied Psychology, 86, 837-855. *The Journal of applied psychology* 86 (2001), 837–55.

[23] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 315–323.

[24] Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation. *IEEE Transactions on Affective Computing* (2022), 1–1.

[25] Everlyne Kimani and Timothy Bickmore. 2019. Addressing Public Speaking Anxiety in Real-Time Using a Virtual Public Speaking Coach and Physiological Sensors. In *Proc. ACM international conference on intelligent virtual agents (IVA)*. 260–263.

[26] Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. 2019. You'll be Great: Virtual Agent-based Cognitive Restructuring to Reduce Public Speaking Anxiety. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. 641–647.

[27] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proc. International Joint Conference on Natural Language Processing*. Springer, 596–605.

[28] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43/.

[29] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 230–237.

[30] Robert Lent and Steven Brown. 2013. Social Cognitive Model of Career Self-Management: Toward a Unifying View of Adaptive Career Behavior Across the Life Span. *Journal of counseling psychology* 60 (2013).

[31] Rui Li, Jared Curhan, and Mohammed Ehsan Hoque. 2015. Predicting video-conferencing conversation outcomes based on modeling facial expression synchronization. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–6.

[32] Songqi Liu, Jason Huang, and Mo Wang. 2014. Effectiveness of Job Search Interventions: A Meta-Analytic Review. *Psychological bulletin* 140 (2014).

[33] Mary Lundeberg, Paul Fox, Amy Brown, and Salman Elbedour. 2000. Cultural influences on confidence: Country and gender. *Journal of Educational Psychology* 92 (2000), 152–159.

[34] Todd Maurer and Kimberly Andrews. 2000. Traditional, Likert, and Simplified Measures of Self-Efficacy. *Educational and Psychological Measurement - EDUC PSYCHOL MEAS* 60 (2000), 965–973.

[35] Candy Olivia Mawalim, Shogo Okada, and Yukiko I. Nakano. 2021. Task-Independent Recognition of Communication Skills in Group Interaction Using Time-Series Modeling. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 4, Article 122 (2021).

[36] Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. 2016. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In *icmi*. 84–91.

[37] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–6.

[38] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2018. Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing* 9, 2 (2018), 191–204.

[39] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*. 139–148.

[40] Laurent Son Nguyen, Denis Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia* 16, 4 (2014), 1018–1031.

[41] Laurent Son Nguyen and Daniel Gatica-Perez. 2015. I Would Hire You in a Minute: Thin Slices of Nonverbal Behavior in Job Interviews. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 51–58.

[42] Shogo Okada, Laurent Son Nguyen, Oya Aran, and Daniel Gatica-Perez. 2019. Modeling Dyadic and Group Impressions with Intermodal and Interperson Features. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1s, Article 13 (2019), 30 pages.

[43] Shogo Okada, Yoshihiko Ohtake, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating Communication Skills Using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 169–176.

[44] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 50–57.

[45] Monica Pereira and Kate Hone. 2021. Communication Skills Training Intervention Based on Automated Recognition of Nonverbal Signals. In *Proc. ACM CHI Conference*. Article 742, 14 pages.

[46] Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. 2015. Evaluating Speech, Face, Emotion and Body Movement Time-series Features for Automated Multimodal Presentation Scoring. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 23–30.

[47] Sowmya Rasipuram, Pooja Rao S. B., and Dinesh Babu Jayagopi. 2016. Asynchronous Video Interviews vs. Face-to-Face Interviews for Communication Skill Measurement: A Systematic Study. In *icmi*. 370–377.

[48] Sowmya Rasipuram and Dinesh Babu Jayagopi. 2020. Automatic Multimodal Assessment of Soft Skills in Social Interactions: A Review. *Multimedia Tools and Applications* 79, 19–20 (2020), 13037–13060.

[49] Nicolas Sabouret, Hazaël Jones, Magalie Ochs, Mathieu Chollet, and Catherine Pelachaud. 2014. Expressing social attitudes in virtual agents for social training games.

[50] Alan M. Saks. 2005. Job Search Success: A Review and Integration of the Predictors, Behaviors, and Outcomes. *Career development and counseling: Putting theory and research to work* (2005), 155–179.

[51] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[52] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2018. Listening Skills Assessment Through Computer Agents. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 492–496.

[53] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated Social Skills Trainer. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*.

[54] Richard C. Tessler and Lisa Sushelsky. 1978. Effects of Eye Contact and Social Status on the Perception of a Job Applicant in an Employment Interviewing Situation. *Journal of Vocational Behavior* 13 (1978), 338–347.

[55] Fuhui Tian, Shogo Okada, and Katsumi Nitta. 2019. Analyzing Eye Movements in Interview Communication with Virtual Reality Agents. In *Proc. International Conference on Human-Agent Interaction*. Association for Computing Machinery, 3–10.

[56] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal Public Speaking Performance Assessment. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 43–50.

[57] Katarzyna M Zinken, Sue Cradock, and T Chas Skinner. 2008. Analysis system for self-efficacy training (ASSET): assessing treatment fidelity of self-management interventions. *Patient Education and Counseling* 72, 2 (2008), 186–193.