INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Scoring of Monologue Video Interviews Using Multimodal Cues

Lei Chen, Gary Feng, Michelle Martin-Raugh,
Chee Wee Leong, Christopher Kitchen, Su-Youn Yoon,
Blair Lehman, Harrison Kell, Chong Min Lee

Educational Testing Service (ETS)
660 Rosedale Rd
Princeton, NJ, 08541 USA
{LChen,GFeng,MMartin-raugh,CLeong,CKitchen,SYoon,BLehman,HKell,CLee001}@ets.org

Abstract

Job interviews are an important tool for employee selection. When making hiring decisions, a variety of information from interviewees, such as previous work experience, skills, and their verbal and nonverbal communication, are jointly considered. In recent years, Social Signal Processing (SSP), an emerging research area on enabling computers to sense and understand human social signals, is being used develop systems for the coaching and evaluation of job interview performance. However this research area is still in its infancy and lacks essential resources (e.g., adequate corpora). In this paper, we report on our efforts to create an automatic interview rating system for monologue-style video interviews, which have been widely used in today's job hiring market. We created the first multimodal corpus for such video interviews. Additionally, we conducted manual rating on the interviewee's personality and performance during 12 structured interview questions measuring different types of job-related skills. Finally, focusing on predicting overall interview performance, we explored a set of verbal and nonverbal features and several machine learning models. We found that using both verbal and nonverbal features provides more accurate predictions. Our initial results suggest that it is feasible to continue working in this newly formed area.

1. Introduction

Few interpersonal interactions carry more stakes than a job interview, in which the interviewee must demonstrate his or her employment qualifications and interpersonal skills under time and emotional stress [1]. These intense speech acts provide a rich playground for understanding how verbal and nonverbal components of speech are coordinated to achieve the speaker's intention. Furthermore, recent technological advancements make it possible to quantify workplace interactions such as job interviews [2]. Our research aims to evaluate the effectiveness of job interview performance using multimodal sensing technologies.

From the perspective of the interviewer, one important goal of an interview is to assess interviewees' knowledge, skills, abilities, and behavior in order to select the most suitable person for the job [1]. This puts much pressure on the interviewee, who must orchestrate his or her multimodal behaviors, such as speech content, prosody [3], and nonverbal cues to effectively communicate his or her qualifications in a limited amount of time [4, 5]. The success or failure of the interviewee's effort is traditionally assessed subjectively by the interviewer, either through a holistic impression or quantitative ratings. The validity and reliability of these assessments is subject to much research [6].

An alternative to the traditional human-only interview assessment model is to augment human judgment with automated assessment of interview performance. Social Signal Processing (SSP) [7] provides a general framework for using multimodal sensing and machine perception to analyze human communication. The workplace is a rapidly emerging field for the application of SSP, because effective human interaction is critical to productivity and because the accumulation of digital records can potentially be mined for insights [2]. However, SSP research in job interview performance is still in its infancy. One of the critical challenges it faces is the lack of high quality multimodal corpora. In this paper, we will describe (a) a new video interview corpus containing monologue online interview responses, which is missing in the existing data resources, (b) the inclusion of a job performance related task, giving public-speaking oral presentations, together with job interviews, and (c) our preliminary studies on the manual rating of job interviews and on automatic rating using rich multimodal cues.

The remainder of the paper is organized as follows: Section 2 reviews the previous research on interview coaching and evaluation using SSP; Section 3 describes our monologue video interview multimodal corpus; Section 4 describes the human ratings on the collected interviews; Section 5 describes a preliminary study on using a series of delivery features from both verbal and nonverbal channels to predict interview performance automatically; Lastly, Section 6 discuss our findings and plans for important next-step research.

2. Previous Research

The present study focuses on the automated scoring of interview videos in a task where the interviewee responds to a fixed set of standardized questions, also known as a structured interview (SI). Research from organizational psychology shows that structured interviews (SIs) tend to produce more valid results than unstructured interviews [8]. The structured nature of the interviews also facilitates the development of automated scoring algorithms. Our research is also motivated by the rapid growth of video-based interviews. In recent years, online video-based SIs have been increasingly used in hiring processes [9]. For example, HireVue[1] is a major vendor for hosting online video interviews and has been reportedly used by many Fortune 500 companies. Conducting online video-based interviews brings

[1] http://www.hirevue.com

http://dx.doi.org/10.21437/Interspeech.2016-1453

many benefits to both interviewers and interviewees, including the convenience of offline reviewing and decision making by human resources (HR) staff, which in turn enables HR staff to assess several job applicants in a short time window. It also opens the door for automated performance analyses to assist HR decision making and possibly reduce human biases.

There are research efforts in building automatic evaluation systems to judge interviewees' performance, such as [10, 11]. In [10], a multimodal corpus consisting of 62 interviews of candidates applying to a real temporary job was built. Each interview lasted approximately 11 minutes. Four interview questions measuring job-related skills on *communication*, *persuasion*, *conscientious works*, and *coping with stress*, were used. An organizational psychology Master's degree student rated each question and also the entire interview for a hiring recommendation. From both applicants and interviewers, various audio features (e.g., speaking activity, pauses, prosody, etc.) and visual behavior cues (e.g., head nods, smiling, etc.) were automatically extracted. Afterwards, these multimodal cues were used to predict five types of human rated scores by using different machine learning approaches, e.g., ridge regression.

[11] conducted research on the *MIT Interview Dataset*, which consists of 138 audio-video recordings of mock interviews from internship-seeking students at MIT. The total duration of the recorded interviews is about 10.5 hours. Counselors asked interview questions that were recommended by MIT Career Services to measure student applicant's behavioral and social skills. Likert scale questions (7-point) were used to rate interviewees' performance. Specifically, there were 16 assessment questions that included two questions about overall performance (*overall rating* and *recommended hiring*), with the remaining questions targeting behavioral dimensions (e.g., *presence of engagement*). The ratings were conducted by counselors and Amazon Mechanical Turk workers. The automatic analysis used the following multimodal cues: (a) facial expressions, language (e.g., word counts, topic modeling), and prosodic information. The ground truth ratings were obtained by a weighted average over the ratings from 9 Turkers. These multimodal cues were fed into a machine learning model (in regression mode), i.e., SVM regression or LASSO, to obtain automatic ratings.

To date, most research on automatic interview scoring focuses on dyadic interviews. We argue, however, that structured interviews may be a better place to begin. In addition, while most prior corpora include the interview performance only, we aim to build a richer assessment of job-related skills in which SI is a component. Thus, the goal of the present work includes (a) providing the first multimodal corpus of monologue online structured interviews, (b) complementing interview performance with interviewees' public-speaking presentations, and (c) building an initial automatic scoring model on our interview corpus.

3. Corpus

The present research involves a corpus with two types of multimodal, job-related tasks: (a) interview performance and (b) oral presentation performance. To investigate interview performance, we developed 12 past-focused behavioral interview questions that assess 4 different applied social skills, referring to the ability to function effectively in social situations at work [12]: (a) *communication skills*, (b) *interpersonal skills*, (c) *leadership*, and (d) *persuasion and negotiating*. Each of these four skills was based on previous research that suggests past-focused behavioral interview questions in which the applicant is asked about how he or she has handled work-related situations in the past, yield higher validity coefficients than future-oriented (or hypothetical) behavioral interview questions [13]. The questions were presented as slides in a PowerPoint presentation on a computer screen. Participants were given 1 minute to prepare and up to 2 minutes to respond to each question. The allocated response time was tracked on the computer screen. To measure presentation skills, we used two types of presentation tasks, including (a) an *informative presentation* and (b) a *persuasive talk*, following the research presented in [14].

A depth camera, Kinect for Microsoft Version 1, was used for recording body movements for the presentation task. An HD webcam (Logitech C615) was used for video recording in order to simulate the real scenario of online video interviews. Audio recording was done by concurrently recording audio responses using the webcam's microphone and the microphone array inside the Kinect device. Data streams from the different sensors i.e., Kinect and webcam, were time-synchronized in the MultiSense [15] framework.More details can be found in the (a) and (b) panels in Figure 1.

A total of 36 participants completed both tasks[2]. Most of the the participants were recruited from the authors' institution and did not receive compensation for participating (beyond their salaries). Participants from outside the organization were paid 60 USD for their participation. During each data collection session, participants first completed a personality survey. Then, they completed the interview task, which was followed by the presentation task. In total[3], we obtained 419 interview responses lasting about 753 minutes and 68 presentations lasting about 249 minutes.

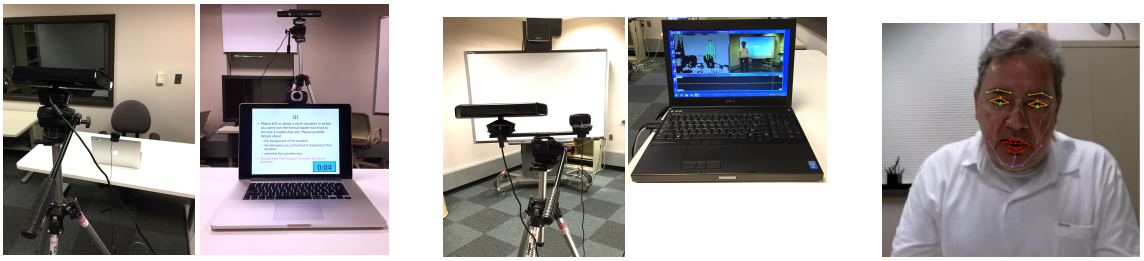
4. Human Rating of Interviews

We completed the human ratings of the interview videos by focusing on overall hiring recommendation and personality traits as suggested in [12]. Nine raters were recruited from our institution to rate the responses to each structured interview question individually. The entire rating process consisted of two rounds. The first was conducted in order to analyze inter-rater agreement. Four raters formed different rating pairs so that each response was rated by two different raters. Rating pairs were created such that all six possible combinations were included. Videos were assigned to rating pairs randomly within the type of question (i.e., communication, leadership, persuasion and negotiation, and teamwork). Later, in the second round, another five raters rated each individual response. As a result, on each response, for each rating dimension, we obtained 7 ratings.

All ratings involved raters indicating the degree to which they agreed with a particular statement about the participant's performance in the video on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Raters were asked to make their judgments based on their first impression of each video. In other words, raters were not trained on the features being rated and agreement between raters was not assessed prior to rating all of the videos. This rating procedure was consistent with those conducted in [16, 11]. Personality ratings were completed by providing the raters with statements including adjectives (e.g., *assertive*, *irresponsible*, *cooperative*) that corresponded to each of the Big Five personality traits (*extroversion*, *agreeableness*, *conscientiousness*, *emotional stability*, and *openness to experi-*

[2]Note that the job interview task was a simulated task, and was not associated with an actual hiring process.

[3]Data from some sessions were lost due to hardware/software issues.



(a) interview (b) presentation (c) Visage SDK screen shot

Figure 1: The data collection setups; In (a) the left panel shows the setup for collecting video interviews while the right panel shows the interview question prompt; In (b) the left panel shows the setup for collecting presentations while the right panel shows the MultiSense software running on a high-end laptop that shows a skeleton tracking; (c) shows the head and gaze tracking result using the Visage SDK.

ence), similar to the multiple-item measure used in [16]. Note that the personality factor-specific adjectives were selected for each type of interview question, resulting in 4 separate rating forms. In addition, raters were also asked to make a holistic judgment about hiring the participant for an entry-level position[4]. Table 1 reports on rating quality, including (a) Intraclass correlation coefficient (ICC) [17], a metric commonly used for measuring consistency among raters, (b) statistics with respect to the correlations of individual raters' scores to the final averaged scores, i.e., Min, Max, and Mean. Note that the ICC computation used the two-way random average measure of consistence, denoted as $ICC(2,k)$. ICC and R values suggest that the manual rating interview performance is a difficult task – it is hard to reach a high agreement between a pair of raters. It is generally a good practice to use the averaged scores of multiple raters.

Table 1: Analysis of human rating quality: ICC, Max, Min, and Mean of individual rater's scores' R to the averaged one

Category	ICC	Max	Min	Mean of R
Agreeableness	0.69	0.50	0.76	0.59
Conscientiousness	0.54	0.39	0.68	0.51
Emotion stability	0.65	0.51	0.71	0.58
Extraversion	0.69	0.49	0.71	0.60
Openness	0.54	0.32	0.64	0.53
Hiring recommendation	0.67	0.53	0.73	0.60

5. Automatic prediction

Inspired by [11], we used the tool Linguistic Inquiry Word Count (LIWC) [18] to extract lexical features related to interview content. The LIWC reports the counts of various psycholinguistic word categories that include words describing negative emotions (sad, angry, etc.), positive emotions (happy, kind, etc.), different function words (articles, pronouns, etc.), and various content categories (e.g., anxiety, insight).

As for measuring the speech delivery aspect, we mostly used the mature technology provided by the automated scoring (AS) research area. Speaking skills are comprised of multiple dimensions, including fluency, pronunciation, prosody, language usage, and so on. In the past two decades, automated scoring (AS) technology has been developing to provide objective and comprehensive measurements of these dimensions[19, 20, 21, 22]. As suggested in [8], various voice characteristic measurements related to speaking fluency, such as pitch range, speaking rate, pauses, etc., influence hiring decision making. Therefore, we utilized the technology that was originally developed for measuring speaking proficiency to provide voice cues. Note that the same technology has been utilized on rating oral communication performances [14, 11]. Therefore, following the feature extraction method described in [21, 23], we generated a series of features on the multiple dimensions of speaking skills using ETS's SpeechRater system. Note that manual transcriptions were used in this study rather than speech recognition outputs generally used in the speechRater system.

Keeping proper eye contact with interviewers is considered to be a proper nonverbal behavior during the interview process. Therefore, we extracted a set of features from head postures and eye gazes that were tracked from interview videos. Head postures are approximated using the *rotation* attribute (i.e., *pitch*, *yaw*, and *roll*) of the head through Visage's SDK FaceTrack[5], a robust head and face tracking engine. A screen shot of running this software on one collected interview video can be found in panel (c) of Figure 1. The tracking is activated if and only if the detector has detected a face in the current frame. The translation attribute can be represented using three coordinates X, Y and Z, corresponding to pitch (left to right), yaw (up to down), and roll (near to far), additionally, gaze directions are approximated through the *gazeDirectionGlobal* attribute of the Visage tracker SDK, which tracks gaze direction taking into account both head pose and eye rotation. Taking the camera as the reference point, gaze direction is estimated with three values determining the rotations around the three axes X,Y and Z in radians. For each interviewee's basic head pose measurements, (i.e., pitch, yaw, and roll) and gaze tracking measurements (i.e., X, Y, and Z) over the entire interview, we computed four types of statistics moments, i.e., emphmean, *SD*, *kurtosis*, and *skewness*. Additionally, a feature measuring extreme values' ratio (*ert*) is computed as follows: for each measurement, obtain the $10th$ percentile and $90th$ percentile from our entire data set as the lower-bound and the upper-bound. For each contour, we then use the proportion of the values beyond these two bounds as a feature.

[4]The position was defined as one that requires some work in groups, some solitary work, and some team leading responsibilities, but not managerial duties. The description of the position was broad because participants were not given a specific position to keep in mind while answering the questions.

[5]http://www.visagetechnologies.com/

Facial expressions from interviewees, e.g., adequate amount of smiles, also play important roles for behaving well during interviews. Therefore, we utilized an off-the-shelf emotion detection toolkit, Emotient's FACET SDK[6], to analyze facial expressions. FACET outputs the intensity (ranging from 0.0 to 1.0) and confidence values for seven primary emotions (i.e., *anger*, *contempt*, *disgust*, *joy*, *fear*, *sadness* and *surprise*), as well as three types of overall measurements on *neural*, *positive*, and *negative* emotions. Similar to head pose and eye gazes, for each emotion contour during an interview, we used one contour' various statistics moments, including mean, SD, kurtosis, and skewness to serve as interview-level emotion features following a standard way in multimodal research [24].

Table 2: The correlations between the selected multimodal features and human rated holistic scores ($N = 419$)

Feature	R
# Word	0.312
# Personal pronoun, e.g., *I*, *you*	−0.206
# Negation, e.g., *no*, *not*, *never*	−0.248
# Differentiation, e.g., *hasn't*, *but*, *else*	−0.204
3 features about responses' duration, e.g, # types	0.22 to 0.36
4 features about silence patterns, e.g, # silence per word	−0.21 to −0.43
4 features about stress patterns, e.g., mean distance of stressed syllables	−0.22 to −0.43
4 features about tone boundary patterns, e.g., mean distance of tone boundaries	0.22 to 0.35
3 features about pronunciation/dialect, e.g., shifts of phonemes' duration to standard	0.21 to 0.33
mean intensity of disgust	−0.170
mean intensity of neutral	−0.275
mean intensity of fear	−0.212
kurtosis, measuring "tailedness" of $gaze_Y$ distribution	0.174
extreme ratio of $gaze_Y$,	−0.192

For this paper, we focused on the tasks of using multimodal features to automatically predict hiring decision scores, which were obtained from averaging 7 raters' judgments. Since a large number of LIWC and SpeechRater features were available, for more accurate models, we selected verbal features according to the following three criteria: (a) the feature's inter-correlation values to others are not larger than a threshold (0.8 per our experience), (b) the Pearson correlation between a feature and the absolute value of the human-rated scores R is large enough (0.2 for verbal features), and (c) R's signs are consistent with our intuition. Regarding visual features, since their number is limited, we only applied steps (b) and (c) above. Finally, a set of multimodal features covering lexical, speech, and visual aspects were selected for our prediction experiment. The details of these features, including their definitions and R values can be found in Table 2. Note that for reasons of space, when describing SpeechRater features, we bundled the related ones into one row and reported the range of their R.

We applied a standard machine learning framework using the multimodal features to predict the interviews' holistic scores. In particular, we run a leave-one-interviewee-out cross-validation among all interviewees ($n = 36$). In each fold, interviews from 35 interviewees were used to train a regression model that was then applied to predict interview performance of the remaining interviewee. The conducted experiments are divided into three feature groups, namely (a) visual features (visual), (b) speech and lexical features (speech+lexical), and (c) the combination (multimodal). Three regression approaches widely employed in practice were utilized, with their implementations in the R **Caret** package [25]: (a) Support Vector Machine (SVM) using a linear kernel (svmLinear), (b) ridge regression, (c) lasso regression (LASSO). Hyper parameters of these machine learning models were automatically tuned by using an inner 10-fold cross-validation on the training set. The whole process was repeated for 36 times to obtain the machine predicted scores for all interviews. Table 3 reports on the correlation between the human-rated holistic scores and the machine-predicted scores.

Table 3: Using multimodal features to predict final holistic scores on the video interviews

Feature set	SVM (linear)	$ridgeLR$	LASSO
visual	0.344	0.335	0.324
speech+lexical	0.416	0.414	0.365
multimodal	0.446	0.458	0.452

The experimental results show that both verbal and nonverbal cues play roles in determining interview performance. Jointly using two types of cues provides more accurate prediction. Though various multimodal features were tried in our experiment, in this initial study stage, we didn't include any content related features. Based on comments made by industrial psychology researchers, it is important to address the lack of such features in order to substantially improve the overall prediction performance. In [11], a method based on topic modeling has been suggested. Additionally, other novel natural language processing methods, e.g., doc2vec [26], may be worth trying for providing important content related measurement.

6. Conclusions

In this paper, we report on a new multimodal corpus to support the use of Social Signal Processing (SSP) in a new area, the workplace. To our knowledge, this is the first research effort to collect structured video interview responses, which have become increasingly important in the online interview industry. Another novelty of this corpus is that beyond job interviews, the same interviewees' public speaking multimodal behaviors were recorded. This could provide useful evidence for job-related skills. We have finished rating personalities and hiring decisions by using a number of human raters. Using a set of multimodal cues, we conducted an experiment on predicting hiring decisions automatically. The prediction results show that both verbal and nonverbal cues are useful for more accurately rating interview performance. Also, the existing results suggest the importance of the inclusion of content features in the scoring model.

In the next steps of this research, we plan to enhance our existing features (particularly the visual features) and add content-related measurements. In addition, it will be interesting to include human-rated personality scores in our scoring model.

[6] http://www.emotient.com

7. References

[1] W. H. Wiesner and S. F. Cronshaw, "A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview*," *Journal of Occupational Psychology*, vol. 61, no. 4, pp. 275–290, 1988.

[2] D. Gatica-Perez, "Signal processing in the workplace [social sciences]," *Signal Processing Magazine, IEEE*, vol. 32, no. 1, pp. 121–125, Jan 2015.

[3] C. K. Parsons and R. C. Liden, "Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues." *Journal of Applied Psychology*, vol. 69, no. 4, p. 557, 1984.

[4] R. J. Forbes and P. R. Jackson, "Non-verbal behaviour and the outcome of selection interviews," *Journal of Occupational Psychology*, vol. 53, no. 1, pp. 65–72, 1980.

[5] A. S. Imada and M. D. Hakel, "Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews." *Journal of Applied Psychology*, vol. 62, no. 3, p. 295, 1977.

[6] I. Nikolaou and J. K. Oostrom, *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice*. Psychology Press, 2015.

[7] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5989788

[8] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *Journal of Business and Psychology*, vol. 24, no. 2, pp. 179–192, 2009. [Online]. Available: http://link.springer.com/article/10.1007/s10869-009-9098-0

[9] A. Hiemstra and E. Derous, "Video resumes portrayed: Findings and challenges," *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice*, pp. 44–60, 2015.

[10] L. Nguyen, D. Frauendorfer, M. Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *Multimedia, IEEE Transactions on*, vol. 16, no. 4, pp. 1018–1031, June 2014.

[11] I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque, "Automated prediction and analysis of job interview performance: The role of what you say and how you say it," in *Proc. of Automatic Face and Gesture Recognition (FG)*, 2015.

[12] A. I. Huffcutt, J. M. Conway, P. L. Roth, and N. J. Stone, "Identification and meta-analytic assessment of psychological constructs measured in employment interviews." *Journal of Applied Psychology*, vol. 86, no. 5, p. 897, 2001.

[13] H. T. Krajewski, R. D. Goffin, J. M. McCarthy, M. G. Rothstein, and N. Johnston, "Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future?" *Journal of Occupational and Organizational Psychology*, vol. 79, no. 3, pp. 411–432, 2006.

[14] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee, "Towards automated assessment of public speaking skills using multimodal cues," in *Proceedings of the 16th international conference on multimodal interfaces*. ACM, 2014.

[15] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 135–140.

[16] M. R. Barrick, G. K. Patton, and S. N. Haugland, "Accuracy of interviewer judgments of job applicant personality traits," *Personnel Psychology*, vol. 53, no. 4, pp. 925–951, 2000.

[17] G. G. Koch, "Intraclass correlation coefficient," *Encyclopedia of statistical sciences*, 1982.

[18] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[19] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[20] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, p. 401, 2010.

[21] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[22] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, p. 355, 2010.

[23] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[24] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 106–118, 2011.

[25] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[26] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1188–1196. [Online]. Available: http://www.jmlr.org/proceedings/papers/v32/le14.html