MAP: Multimodal Assessment Platform for Interactive Communication Competency

Saad M. Khan
Educational Testing Service
skhan002@ets.org

David Suendermann-Oeft, Keelan Evanini, David M. Williamson, Scott Paris, Yao Qian, Yuchi Huang, Phillip Bosch*, Sidney D'Mello*, Anastassia Loukina, Lawrence Davis
Educational Testing Service
* University of Notre Dame
suendermann-oeft@ets.org; kevanini@ets.org; dmwilliamson@ets.org; sparis@ets.org; yqian@ets.org; yhuang001@ets.org; pbosch1@nd.edu; sdmello@nd.edu ; aloukina@ets.org; ldavis@ets.org;

ABSTRACT: In this paper, we describe a prototype system for automated, interactive human communication assessment. The system is able to process multimodal data captured in a variety of human-human and human-computer interactions. These data can be analyzed along many dimensions of verbal and non-verbal communication competencies including speech delivery, language use, social affect and engagement. The system also integrates speech and face recognition-based biometric capabilities and has design elements to enable ranking and indexing of large collections of assessment content.

Keywords: Multimodal Analytics, Interactive Communication Assessment, English Language Learning and Assessment, Biometrics, Learning Systems

1 INTRODUCTION

Assessment of English language communication competence is difficult because of the complex skills that underlie the competence and the technical difficulties of measuring dynamic speaking behaviors. Current large-scale assessments of non-native English speaking proficiency (such as TOEFL iBT[1] and Pearson Test of English Academic[2]) typically contain brief test questions and prompts that elicit monologues from the test taker. Since they do not elicit interactive conversations from the test taker or measure non-verbal communication, these assessments are incomplete evaluations of human communication skills. These tools do not capture the patterns of speech and behavior that are critical to effective human interactions (Pentland 2008). Other assessments (such as IELTS[3]) assess interactive speech by using one-on-one

1 http://www.ets.org/toefl

2 http://pearsonpte.com/

3 http://www.ielts.org

interviews with human examiners. However, this approach is difficult to standardize due to subjective aspects of individual interviewers.

In recent years some exciting work has been done to create learning/training systems that combine virtual agent, multimodal analyses of user behaviors, and dialogue management (Devault et al. 2014, Hoque et al. 2013). A key feature of these systems is the ability to perform real-time tracking of both verbal and non-verbal behavior to create immersive simulation experiences. However, these tools are not designed to create holistic assessments of conversational English communication skills and have not been validated for assessments of speaking proficiency of non-native English speakers.

In this paper we present the Multimodal Assessment Platform (MAP); a system that can analyze audio-visual data from non-native English speakers engaged in interactive conversations with other humans or automated agents as well as structured prompt questions and response scenarios. The system captures and analyzes data in authentic settings using rich audio-visual interfaces. MAP provides data about subjects along three dimensions of speech and non-verbal behavior: Delivery, Language Use, and Social Engagement. In addition, the system includes a Biometric Security functionality for identity verification based on facial recognition and speech-based speaker identification. MAP has the potential to be a useful tool to better gauge interactive English communication ability in a variety of domains including university admissions offices, workforce hiring non-native speakers and English language learning systems. In the following we provide details of the MAP system starting with a brief description of the service-oriented architecture underlying the back-end core.

2 TECHNICAL APPROACH

The MAP front-end user interface is designed to review, and verify analytical results as well as to demonstrate system functionality. The interface can be used to (a) play back processed audio-visual data, (b) annotate processed data with automatically scored metrics and behavioral events, and (c) view score reports and summary statistics. Figure 1 below shows the user interface when browsing the assessment results from the processing of a test video. The interface has a tabbed structure that allows users to easily browse through the details of each of the three assessment categories: Delivery, Language Use and Social Engagement. In addition, the Security tab displays results from speech biometrics and face recognition. Finally, the Dashboard tab provides an overview of performance results from each of the three constructs of interactive communication assessment and biometric authentication using color codes: threshold (yellow), above threshold (green) and below threshold (red). The threshold values are configurable by the system administrator and can be determined empirically. A customized version of the interface can be used to identify classification errors in analytics and provide data to improve and refine the underlying algorithms.

The MAP back-end has been designed using a service-oriented architecture that can accept processing job requests submitted over the web. The core consists of a multimodal data processing framework analyzing audio and visual data for automated detection of verbal and non-verbal communication. Computer vision, speech, and natural language processing algorithms, are utilized to extract low-level visual features including facial action units, head orientation, silhouette contours, and acoustic features

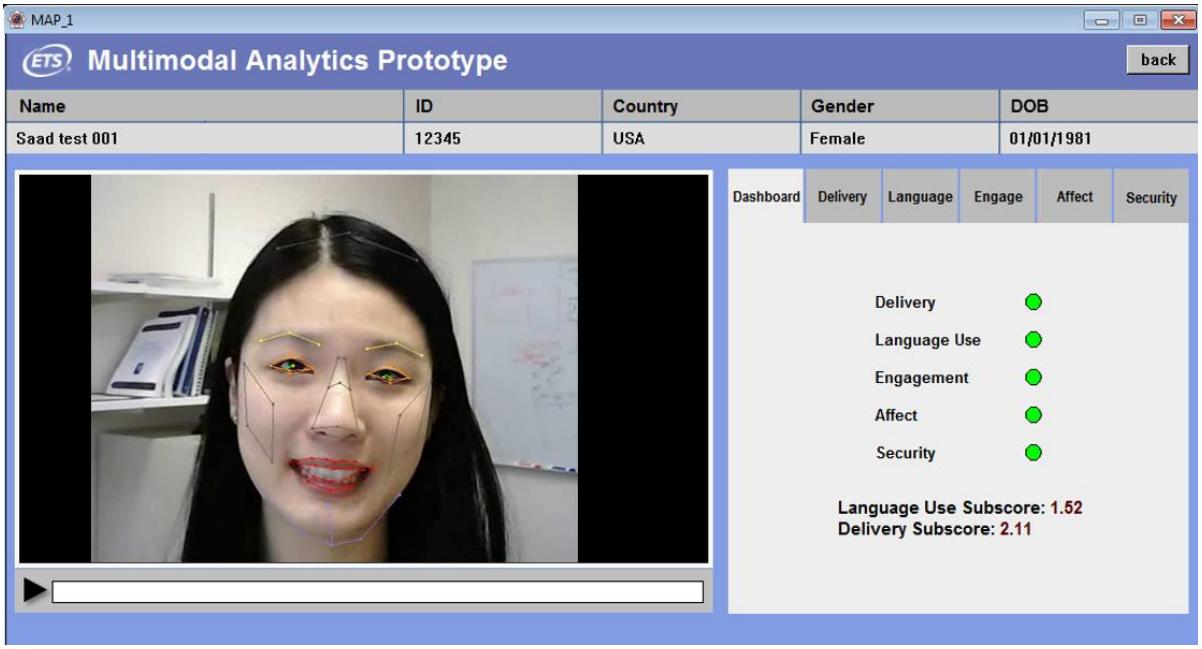

Figure 1: Front-end interface of the system. Users can click on tabs to obtain detailed assessment results on biometrics security, social affect, delivery, and language use. The dashboard tab shows a bird's eye view of the analysis results with color codes for threshold (yellow), above threshold (green) and below threshold (red).

such as Mel-frequency cepstral coefficients (MFCCs); mid-level features such as speaking rate and word tokens; and higher level behavioral classifications including facial expression/affect, prosody and vocabulary richness, among others. This core multimodal data processing framework is composed of four functional modules corresponding to the three aspects of interactive communication delivery, language use and engagement, plus biometrics used for speaker identification. In the following subsections we provide technical details of these functional modules.

2.1 Delivery and Language Use

In order to provide information about the speaker's English speaking ability, MAP provides feedback about two different aspects of the speaker's responses captured in the interactive conversation: delivery and language use. Delivery encompasses fluency and pronunciation whereas language use consists of grammar and vocabulary. While these two aspects of speaking proficiency are often correlated in a given non-native speaker's English, they represent different types of skills that are learned and taught in different ways. This is reflected by the fact that standardized assessments of English speaking proficiency often incorporate separate rubrics for delivery and language use.

We use SpeechRater (Zechner et al. 2009), a state-of-the-art automated speech scoring engine designed to process non-native spontaneous speech, to extract a wide variety of features that measure different aspects of the spoken response to prompt questions. These include fluency features such as number of words per second, rate of silent pauses, repetitions and repairs, pronunciation and vocabulary features

such as number of distinct lexical types, word-level acoustic model likelihood amongst others (see (Loukina et al. 2015) for further detail).

2.2 Social Engagement

We developed models to detect two different aspects of student engagement (Fredricks et al. 2004) from video data: behavioral and affective. The behavioral engagement model detects whether students appear to be paying attention and putting forth discretionary effort, on-task behaviors, and participation. We used 3D facial feature tracking (ref) to detect students' faces and extract head yaw (side to side angle) in consecutive 10-second windows in the video stream. We considered the proportion of video frames within the windows in which the face could be detected as an indicator of engagement, based on the intuition that failure to register the face for prolonged periods might suggest disengagement from the interface.

The affective engagement model observes affective and emotional states such as enthusiasm, energy, lethargy, sadness, or distress. It utilizes facial expressions to infer engagement. The model utilizes facial expressions to infer engagement. We applied supervised machine learning methods to facial features extracted from 21 participants in the publically available SEMAINE database (McKeown et al. 2012) of dyadic human interactions. These features included head pose (yaw, pitch, and roll), head position in 3 dimensions, eye gaze direction, and 17 facial action units (AUs) (Ekman and Friesen 1978). We experimented with logistic regression, support vector machines, and multi-layer perceptron classifiers using Weka, a machine learning toolkit. Cross-validation was done by repeatedly (50 iterations) selecting a random 67% of SEMAINE participants for training data and the rest as testing data. Synthetic minority oversampling (SMOTE) was applied to the training data in order to create 400% more examples of engagement so as to bias the model toward predicting the minority class of engagement. A logistic regression classifier produced the best results, with an area under the receiver-operating characteristic curve (AUC) = .612 (chance level = .500).

2.3 Security Biometrics

To enhance security and user identity authentication for validity of the interactive communication competency assessments, we developed voice biometrics (i.e., speaker recognition) and face recognition functionality into the MAP system. Our approach for voice speaker recognition it to utilize i-vectors a compact representation of a speech utterance in a low-dimensional subspace based upon factor analysis. Speech utterances are first converted to a sequence of acoustic feature vectors, and their dynamic counterparts. These are fed into a Gaussian Mixture model to compute a match score between the target and the test speaker (or imposter). A detailed review of a preliminary voice biometrics study can be found in (Qian et al. 2016).

Our face recognition consists of four stages: face detection, face alignment, feature extraction and classification (or verification). Face detection, the first step, identifies a bounding box around the human face in the query image. This is followed by automated face alignment process that uses the face bounding box and localizes accurate positions of various facial feature points in a predefined template. For face

detection, we utilized the mature OpenCV detection algorithm based on Haar cascades [10]; for face alignment, we will employ the method of Kazemi et al. 2014 using an ensemble of regression trees. Aligned face images are used as input for feature extraction and identity verification using a Deep-Convolutional Neural Network (D-CNN) approach (Parkhi et al. 2015).

3 CONCLUSION AND DISCUSSION

In this paper we have presented MAP: Multimodal Assessments Platform a system for the assessment of interactive English communication ability from multimodal data. The system is designed to evaluate performance along a number of dimensions that underlie the construct including speaking fluency, language use, social affect and behavioral engagement. Additionally, for user authentication our system has multimodal biometrics functionality including voice-based speaker identification and facial recognition. The system is designed using a service-oriented architecture and can be used for large scale batch processing.

The MAP system is currently a prototype and has not yet been deployed with end users. One of our potential end users are university admissions offices interested in better understanding interactive English communication ability of non-native students. To that end we are working University of Rochester to capture pilot user data. “Through both our own research and testimonials of staff, students, and alumni, Rochester has long understood and prioritized face-to-face interaction in the admission process. Admission interviews make an important difference in how much we know about an applicant and how much they know about us. We continue to welcome this partnership with ETS in their evaluations." – Jon Burdick, Vice Provost and Dean of College Admission, University of Rochester.

In the future we plan to incorporate avatar-based fully autonomous dialogic interactions and real-time processing of multimodal data. A key area of focus is the fusion of features extracted from multiple modalities for improved accuracy of the assessment constructs as well as better performance on biometric identification. Finally, we believe functionalities developed in this system may also find application in areas such as workforce staffing and healthcare where interactive, conversational communication skills are highly valued.

ACKNOWLEDGEMENTS

We would like to thank Tom Florek, Keith Kiser, Robert Mundkowsky, Christopher Kurzum, Jun Xu and Janet Stumper of ETS for important contributions towards software systems development and data collection activities. We would also like to thank Scott Clyde and Jon Burdick our collaborators from University of Rochester.

REFERENCES

Pentland, A. (2008). Honest Signals: How they shape our world. Cambridge, MA: MIT Press.

Devault, D., Rizzo, A., and Morency, L.P. SimSensei: A Virtual Human Interviewer for Healthcare Decision Support. In Proceedings of the Thirteenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2014

Hoque, M.E., Courgeon, M., Mutlu, M., Martin, J.C., Picard, R.W. MACH: My Automated Conversation coacH, Proceedings of 15th International Conference on Ubiquitous Computing (Ubicomp), September 2013

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication 51, 10, 883-895.

Loukina, A., Zechner, K., Chen, L., and Heilman, M. 2015. Feature selection for automated speech scoring. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 12–19, Denver, Colorado, June 4, 2015

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing, 3(1), 5–17.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A. V., and Ramanarayanan, V. 2016. Noise and Metadata Sensitive Bottleneck Features for Improving Speaker Recognition with Non-native Speech Input. Subm. to Interspeech, San Francisco, CA

Viola, P. and Jones M.J. Robust Real-time Face Detection, International Journal of Computer Vision archive Volume 57 Issue 2, May 2004

Kazemi, Vahid and Sullivan, Josephine. One Millisecond Face Alignment with an Ensemble of Regression Trees. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)

Parkhi, O.M., Vedaldi, A., Zisserman, A. Deep Face Recognition. British Machine Vision Conference, 2015