Measuring the Power of Learning.®

Research Report
ETS RR–17-21

A Multimodal Dialog System for Language Assessment: Current State and Future Directions

David Suendermann-Oeft
Vikram Ramanarayanan
Zhou Yu
Yao Qian
Keelan Evanini
Patrick Lange
Xinhao Wang
Klaus Zechner

May 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Multimodal Dialog System for Language Assessment: Current State and Future Directions

David Suendermann-Oeft,[1] Vikram Ramanarayanan,[1] Zhou Yu,[2] Yao Qian,[1] Keelan Evanini,[3] Patrick Lange,[1] Xinhao Wang,[1] & Klaus Zechner[3]

1 Educational Testing Service, San Francisco, CA
2 Carnegie Mellon University, Pittsburgh, PA
3 Educational Testing Service, Princeton, NJ

We present work in progress on a multimodal dialog system for English language assessment using a modular cloud-based architecture adhering to open industry standards. Among the modules being developed for the system, multiple modules heavily exploit machine learning techniques, including speech recognition, spoken language proficiency rating, speaker recognition, and the scoring of behaviors in multimodal data streams.

Keywords Multimodal analysis; dialog system; language assessment; language learning; speech recognition; speaker recognition

doi:10.1002/ets2.12149

The continued growth of English as a common language in global workforce and academic settings as well as the rise in the population of English language learners attending K – 12 classes in the United States have led to an increased need for valid assessments of English speaking proficiency. Most large-scale standardized assessments of English speaking proficiency incorporate tasks based on the prompt-response model in which the test taker is presented with certain stimulus materials (including reading passages, listening passages, images, etc.) and is then asked to provide spoken responses to one or more prompts based on these stimulus materials. These spoken responses are then recorded and assessed asynchronously by a human evaluator or by an automated assessment system.

One of the limitations of using this prompt-response model for assessment of spoken language proficiency is that it is not interactive; that is, the selection of prompts presented by the assessment system is not influenced at all by the content of a test taker's responses. In contrast to the monologic speech that is elicited by these types of assessment tasks, most spoken language produced by a language learner in an English-medium environment is dialogic and conversational. Therefore, for an assessment to be a valid measure of all of the spoken language proficiency skills that a language learner is required to master, it should also elicit dialogic speech in an interactive conversation. Some standardized assessments of English speaking proficiency, such as IELTS,[1] do elicit dialogic speech through a conversation between the test taker and a human interlocutor; however, this approach is difficult to adopt on a large scale. Furthermore, research has shown that the reliability of such oral proficiency interviews can be reduced due to variations among the human interlocutors who provide the scores (Brown, 2005).

To address this need for more interactive, scalable, and ecologically valid interactive assessments of spoken English proficiency, this report describes work in support of the development of a multimodal dialog system that engages a test taker in an interactive conversation without involving a human interlocutor. This approach has the following advantages:

- reduced costs
- feasibility to be administered at scale
- use of input/output modalities in addition to speech (such as facial expressions and gestures)
- in formative assessment and language learning scenarios, ability for the language learner to practice English at any time, even when an instructor is not available
- availability of immediate feedback about the language learner's performance on the task
- in summative as well as formative assessment scenarios, potentially increased score reliability (in terms of objectiveness, consistency, reproducibility, fairness, and validity)

Corresponding author: V. Ramanarayanan, E-mail: vramanarayanan@ets.org

- detailed knowledge of and control over the scoring features used by the system

The Multimodal Dialog System section briefly describes the architecture of the multimodal dialog system, which is based on the following design principles:

- open source (the use of state-of-the-art open-source technology allows for rapid system development and a large community of supporters and testers)
- open standards compliant (adhering to open and W3C industry standards facilitates the combination of multiple technologies that are required in a complex multimodal dialog system without the need for developing proprietary interfaces)
- cloud based (the use of cloud computing makes solutions easily scalable and provides for easily accessible web-based resources to process large amounts of multimodal data in real time while storing captured media data and logs in a central database)
- distributed and modular (because certain components require dedicated computational resources and even different operating systems — e.g., the speech recognizer or the face tracker — it is advisable to have dedicated machines for separate modules and services; furthermore, depending on the specific application, certain modules might not be necessary, for example, the voice biometrics module in a training application)
- telephony based (this includes regular, IP, or video telephony via regular terminals, soft phones, web browsers, etc. — the advantage of using telephony over audio and video captured in the client and sent asynchronously include the ability to record the full-duplex interaction on the telephony server for assessment purposes and significantly faster reaction, including incremental processing or barge-in)

While the current system adheres to traditional rule-based methods for spoken language understanding and dialog management, multiple modules have been prepared that make heavy use of data-driven methodologies, including a module for automated rating of spoken language (delivery, vocabulary, grammar, and content; see the Speech Scoring section); a module for speaker recognition to enhance test security (see the Speaker Recognition section); and a module to evaluate test takers' behaviors using multimodal cues (see the Multimodal Assessment section). Because several of these modules are still under development, the Conclusion and Outlook section outlines the next steps planned to realize a fully integrated system for multimodal assessment of English speaking proficiency and describes how machine learning can play a major role in accomplishing this goal.

Multimodal Dialog System

The multimodal dialog system architecture is based on the Help Assistant — Language-Enabled and Free (HALEF) framework (Ramanarayanan et al., 2017), which is composed of the following distributed open-source modules (see Figure 1):

- telephony servers — Asterisk (van Meggelen, Smith, & Madsen, 2009) and Freeswitch (Minessale & Schreiber, 2012) — that are compatible with the session initiation protocol (SIP), public switched telephone network (PSTN), and web real-time communications (WebRTC) standards and include support for voice and video communication
- a voice browser — JVoiceXML (Schnelle-Walka, Radomski, & Mühlhäuser, 2013) — that is compatible with VoiceXML 2.1 and can process SIP traffic and that incorporates support for multiple grammar standards, such as Java speech grammar format (JSGF), Advanced Research Projects Agency (ARPA), and weighted finited state transducer (WFST)
- a media resource control protocol (MRCP) speech server, which allows the voice browser to initiate SIP or real-time transport protocol (RTP) connections from/to the telephony server and incorporates multiple speech recognizers (Sphinx, Kaldi) and synthesizers (Mary, Festival)
- an Apache Tomcat-based web server, which can host dynamic VoiceXML pages, web services, and media libraries containing grammars and audio files
- a MySQL database server for storing call-log information
- a speech transcription, annotation, and rating (star) portal that allows one to listen to and transcribe full-call recordings, rate them on a variety of dimensions such as caller experience and latency, and perform various semantic annotation tasks required to train automatic speech recognition and spoken language understanding modules

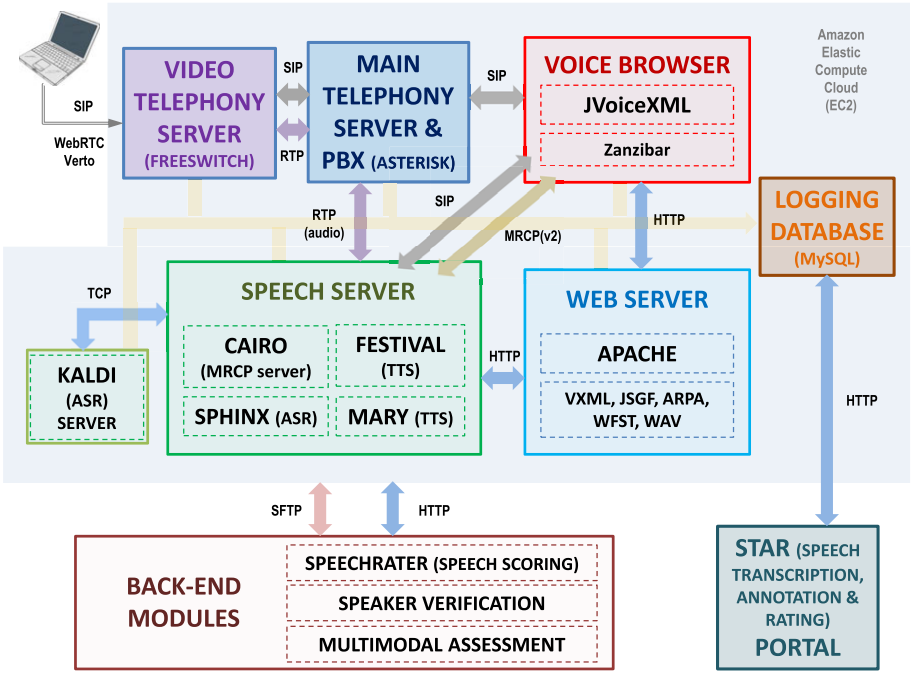

Figure 1 System architecture of the HALEF multimodal dialog system depicting the various modular open-source components.

- OpenVXML, a voice application authoring suite that generates dynamic web applications that can be housed on the web server
- multiple backend services that can be connected as needed to supply additional services to the multimodal application

We have used this architecture to create multiple dialog applications specifically tailored to conversational spoken language assessments (see the examples in Ramanarayanan, Suendermann-Oeft, Ivanov, & Evanini, 2015). While these examples are based on system initiative, for instance, in an interview application, Figure 2 shows another example dialog flow of a user-initiated task. In this task, callers are required to act as customer service representatives at a pizza restaurant and take an order from an automated customer who wants to order a pizza. In such a scenario, the automated customer waits for the user to ask a question (What is your name? What toppings would you like on your pizza? etc.) before replying with the appropriate response.

Speech Scoring

For conversational assessment of spoken language, in particular, formative assessment, it is desirable to use real-time automated speech scoring capabilities. The *SpeechRater*^{SM} Automated Scoring service (Zechner, Higgins, Xi, & Williamson, 2009) was the first system widely used for automatic nonnative spontaneous speech scoring. It consists of three main components: a speech recognizer, a feature computation module, and the scoring model. The features used by SpeechRater are designed by experts to reflect aspects of pronunciation, fluency, intonation, rhythm, vocabulary use, and grammar based on the scoring rubrics that are used by human raters. This alignment between the features used by the automated scoring system and the human scoring rubrics ensures the validity of the automated scores and prevents test takers from capitalizing on features that may be correlated with the final score but are not representative of the skills and constructs that the test is designed to assess.

For our initial work on investigating the use of automated speech rating in dialog systems, we relaxed the aforementioned constraint in favor of creating a real-time-able system. To this end, we implemented a hybrid recurrent neural network framework that comes with minimal manual effort and cost and high scoring accuracy and speed (Yu et al., 2015).

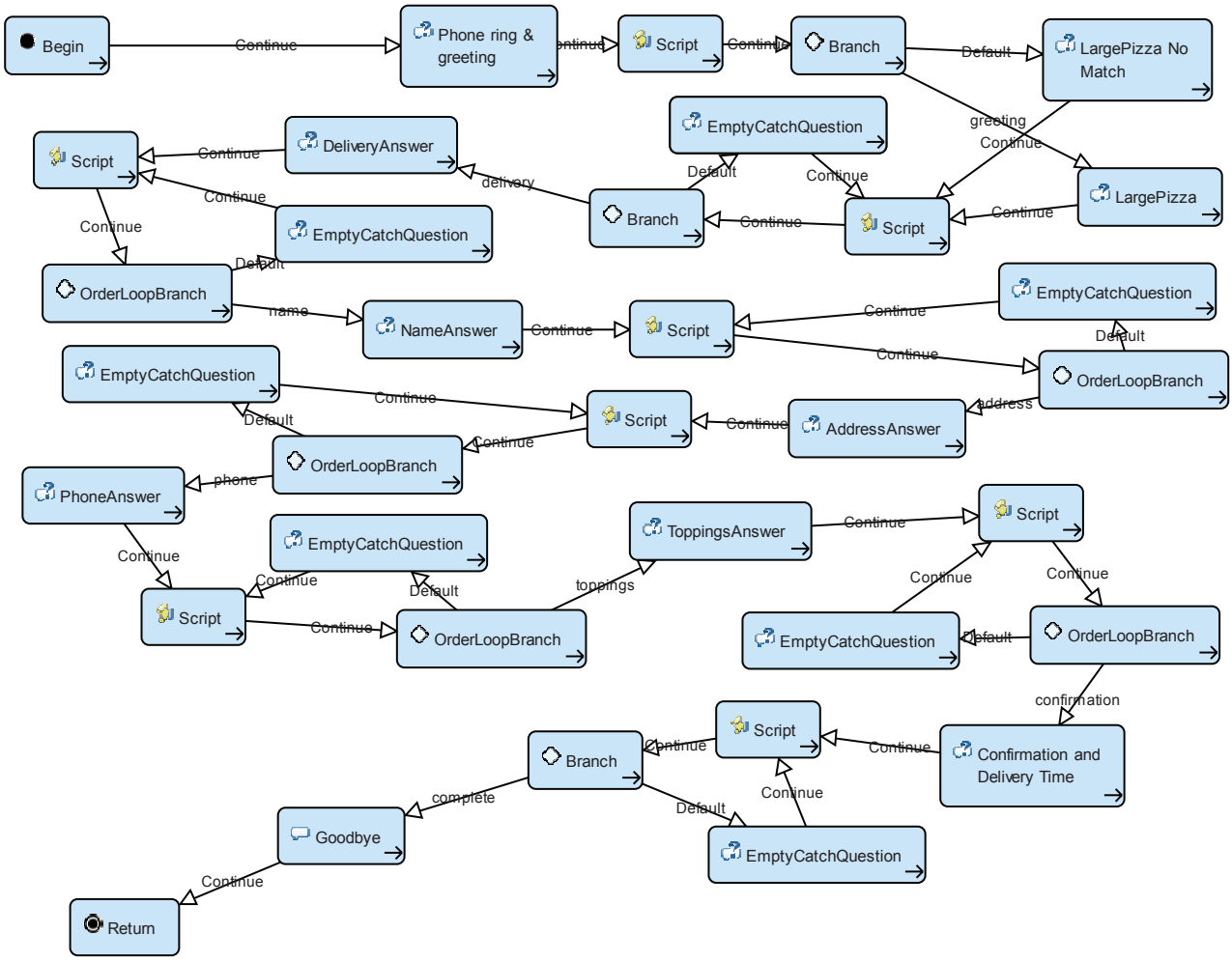

Figure 2 Example dialog flow design of a customer interaction test application.

In the proposed framework, we used generic time-sequence features extracted directly from the audio input instead of manually designed features, thus saving on human transcription effort and expert knowledge for training and optimizing the speech recognition engine for the rater. The proposed framework also jointly optimizes the module for feature engineering and the module for scoring using a hybrid model of a bidirectional long short-memory recurrent neural network (BLSTM) and a multilayer perceptron (MLP). The BLSTM automatically learns the high-level structure information from the basic prosodic features (e.g., pitch) and the mel-frequency cepstral coefficients (MFCCs). The MLP maps the features to a continuous score. The framework optimizes the BLSTM and MLP jointly using mini-batch stochastic gradient descent (SGD). The achieved results are comparable to the baseline SpeechRater model, which requires human effort in designing the features and a significant computational overhead caused by the speech recognizer.

One can use mini-batch SGD to optimize the proposed hybrid network structure in a manner that scales with respect to the amount of training data. This is useful because, as a next step, we plan to increase the amount of nonnative speech training data to explore the upper performance bound of the model (up to 600,000 hours of nonnative speech data are available at Educational Testing Service).

Speaker Recognition

The integration of a speaker recognition module into an interactive assessment of spoken language proficiency has the potential to improve the security and validity of the assessment. However, developing a speaker recognition system for

nonnative speakers of English comes with a number of challenges and opportunities. In this section, we provide a brief description of the work we have performed in this area.

In conventional i-vector-based speaker recognition systems (Dehak, Kenny, Dehak, Ouellet, & Dumouchel, 2011), speech utterances are first converted to a sequence of acoustic feature vectors, typically 20-dimensional MFCCs and their dynamic counterparts. After that, speaker- and channel-independent supervectors, which accumulate zeroth-, first-, and second-order sufficient statistics, are computed by using the posterior probabilities of classes from a pretrained Gaussian mixture model – universal background model (GMM – UBM). Next, the total variability matrix, T, is used to transform the supervectors to low-dimensional i-vectors that contain both speaker and channel variability. Then, linear discriminant analysis (LDA) is used to perform channel compensation. Finally, a score between the target and the test speaker is calculated by a scoring function such as probabilistic LDA (PLDA) for further compensation or by simply using the cosine distance.

However, conventional i-vector-based speaker recognition systems may not be able to compare the speakers at the same phonetic content due to the unsupervised training used in GMM – UBM. Recently, speaker recognition systems with phonetically aware deep neural networks (DNNs) have significantly outperformed standard i-vector-based systems (Lei, Scheffer, Ferrer, & McLaren, 2014; Richardson, Reynolds, & Dehak, 2015). Phonetically aware DNNs used for speaker recognition mainly replace the GMM components with senones and utilize the corresponding posteriors from the senones to extract Baum – Welch statistics. Thus the DNN models phonetic content (senones) in a supervised learning manner. It therefore allows comparison of different speakers at the same phonetic content.

There are many challenges in developing a unified system based on phonetically aware DNNs for speaker recognition in the context of large-scale language assessment. For instance, test takers in a global assessment of English proficiency have a wide variety of native language (L1) backgrounds, which makes it difficult to model all possible nonnative accent patterns. In addition, the speech quality of different test centers can vary substantially due to various recording settings and environments. We tested the DNN-based speaker recognition approach on a nonnative spontaneous speech corpus for the purpose of test security. Noise-aware features and multitask learning were investigated to improve the alignment of speech feature frames into the subphonemic senone space and to distill the L1 information of the test takers into bottleneck features (BNFs), which we refer to as metadata-sensitive BNFs. Experimental results show that the system with metadata-sensitive BNFs can improve speaker recognition performance by an 11.4% relative reduction in equal error rate (EER) compared to the baseline i-vector-based system (Qian et al., 2016). Measured on an enrollment and test set containing 600 nonnative speakers producing 12 utterances recorded at two different times, the best achieved EER was 2.4%. The training data for the GMM-UBM consisted of 150 hours of speech of about 2,000 test takers. We are planning to significantly increase the training data (to up to 600,000 hours available nonnative speech material) to test the influence of a more reliable GMM-UBM and make the speaker recognition system agnostic to adverse acoustic conditions and the influence of L1 variability.

Multimodal Assessment

In recent years, there has been an increasing demand for multimodal assessments of interpersonal interaction and communication skills for purposes of academic and workplace assessment and teacher licensure, among other purposes (Chen et al., 2014). Multimodal data capture techniques based on video and audio feeds as well as potential other modalities (motion capture, chat, advanced sensors) provide a rich source of information for such assessments, but the complexity of these data streams brings with it a need for signal analysis tools to automatically process and make sense of these data.

We recently analyzed (Chen, Leong, Feng, Lee, & Somasundaran, 2015; Chen et al., 2014; Ramanarayanan, Chen, Leong, Feng, & Suendermann-Oeft, 2015) how fusing features obtained from different multimodal data streams, such as speech, face, body movement, and emotion tracks, can be applied to the scoring of multimodal presentations. We analyzed multimodal data collected by Chen et al. (2015) consisting of synchronized and preprocessed recordings from 56 sessions involving speakers giving presentations on different topics. We then performed a comparative analysis of different feature sets including speaking proficiency features extracted from SpeechRater, head pose and eye gaze features, facial emotion features, and body-motion features in predicting multiple scores of presentation proficiency (such as effective formulation of the introduction and conclusion, effective verbal and nonverbal behavior, effective use of visual aids, overall persuasiveness, etc.). We found that certain scoring dimensions were better predicted by certain feature types, and others by combinations of all features. For example, we observed that time-series features computed over the intensities of

different emotional states (estimated from facial expressions) perform best in predicting nonverbal behavior scores, while another score representing the skillful use of visual aids was well predicted by a combination of time-series body-motion features and SpeechRater features. We further observed that these features allowed us to achieve a prediction performance better than the human interrater agreement (as measured by the correlation between scores provided by two nonexpert human raters) for a subset of these scores. For further details, please see Ramanarayanan, Chen, et al. (2015).

Although there is much room for improvement along the lines of better, more interpretable and predictive features as well as machine learning algorithms and methods that generalize across data sets, these experiments provide us significant insight into understanding how to design better techniques for automated assessment and scoring of speaking proficiency.

Conclusion and Outlook

To test the basic functionality of the presented multimodal dialog system, we are presently conducting a user study deploying multiple language assessment prototypes in a crowdsourcing environment. As of September 2016, more than 20,000 calls have been processed by the system (see a preliminary analysis in Ramanarayanan et al., 2016).

Multiple of the system components described in this report are already performing very well in isolation (speech and speaker recognition, speech rating, and multimodal assessment). However, to produce a system of even higher performance and robustness, there is an imminent need to fuse individual models' outputs and create a holistic multimodal understanding component using the combined knowledge of each of these sources. For example, understanding user input in a multimodal dialog can be modeled as a multicriteria optimization problem that allows sequential refinement of a pool of plausible hypotheses. This approach is based on the wide information pipeline (Varges, Riccardi, & Quarteroni, 2008) and makes use of multiple knowledge sources (e.g., acoustic and language models, shallow parsers, intonation models, content features, interpretation of facial expressions and gestures, or speaker trait detection).

Although it is desirable to increase the number and scope of data-driven methods in multimodal dialog systems to both improve performance and reduce the manual labor involved in building hand-crafted systems, this needs to be done with care. For example, the dialog flow designs of the four prototypes currently tested in the crowdsourcing environment were built by test developers bringing essential domain expertise of English language assessments to the table. Furthermore, one of our fundamental design premises is that the building of new applications be simple and hardly require consultation of scientific staff. Our ultimate goal is hence to allow domain experts to build multimodal conversational applications and enjoy the maximum benefit of data-driven methods while being fast and scalable for operational use.

Acknowledgments

The authors would like to express their gratitude for the manifold team contributions without which the presented work would not have been possible: Alan Black, Liz Bredlau, Lei Chen, Gary Feng, Chee Wee Leong, Melissa Lopez, Julie Marlier, Zydrune Mladineo, Robert Mundkowsky, Lydia Rieck, Nehal Sadek, Ayana Stevenson, Jidong Tao, Eugene Tsuprun, Phal Vaughter, and Katie Vlasov. We also express our appreciation to the larger HALEF community: Abdelrahman Abdelkawy, Christian Gaida, Jonathan Grupp, Markus Gutbrod, Youmna Heikal, Ahmed Malatawy, Tarek Mehrez, Jochen Mohrmann, Martin Mory, Michael Muck, Hadeer Nabil, Felix Neutatz, Tim von Oldenburg, Ramin Safarpour, Dennis Schmidt, Stephan Schuenemann, and Moritz Teckenbrock.

Note

1 http://www.ielts.org/test_takers_information/test_sample/speaking_sample.aspx

References

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*, Frankfurt am Main, Germany: Peter Lang.

Chen, L., Feng, G., Joe, J., Leong, C., Kitchen, C., & Lee, C. (2014). Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 200 – 203). New York, NY: ACM. https://doi.org/10.1145/2663204.2663265

Chen, L., Leong, C. W., Feng, G., Lee, C. M., & Somasundaran, S. (2015). Utilizing multimodal cues to automatically evaluate public speaking performance. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 394–400). Piscataway, NJ: IEEE. https://doi.org/10.1109/ACII.2015.7344601

Dehak, N., Kenny, P., Dehak, R., Ouellet, P., & Dumouchel, P. (2011). Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307

Lei, Y., Scheffer, M., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1695–1699). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2014.6853887

Minessale, A., & Schreiber, D. (2012). *FreeSWITCH cookbook*. Birmingham, England: Packt.

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., Ivanov, A., & Ramanarayanan, V. (2016). *Metadata sensitive bottleneck features for speaker recognition with non-native spontaneous speech*. Unpublished manuscript.

Ramanarayanan, V., Chen, L., Leong, C., Feng, G., & Suendermann-Oeft, D. (2015). Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 23–30). New York, NY: ACM. https://doi.org/0.1145/2818346.2820765

Ramanarayanan, V., Suendermann-Oeft, D., Ivanov, A., & Evanini, K. (2015). A distributed cloud-based dialog system for conversational application development. In *Proceedings of the 16th annual meeting of the Special Interest Group on Discourse and Dialogue* (pp. 432–434). Stroudsburg, PA: Association for Computational Linguistics.

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Ivanov, A., Evanini, K., Yu, Z., … Qian, Y. (2016). *Bootstrapping development of a cloud-based spoken dialog system from scratch using crowdsourced data* (Research Report No. RR-16-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12105

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Mundkowsky, R., Ivanou, A., Yu, Z., … Evanini, K. (2017). Assembling the jigsaw: How multiple W3C standards are synergistically combined in the HALEF multimodal dialog system. In D. Dahl (Ed.), *Multimodal interaction with W3C standards: Toward natural interfaces for everything* (pp. 295–310). New York, NY: Springer.

Richardson, F., Reynolds, D., & Dehak, N. (2015). A unified deep neural network for speaker and language recognition. In *Interspeech 2015: 16th annual conference of the International Speech Communication Association* (pp. 1146–1150). Baixas, France: International Speech Communication Association.

Schnelle-Walka, D., Radomski, S., & Mühlhäuser, M. (2013). JVoiceXML as a modality component in the W3C multimodal architecture. *Journal on Multimodal User Interfaces, 7*, 183–194. https://doi.org/10.1007/s12193-013-0119-y

van Meggelen, J., Smith, J., & Madsen, L. (2009). *Asterisk: The future of telephony*. Sebastopol, CA: O'Reilly.

Varges, S., Riccardi, G., & Quarteroni, S. (2008). Persistent information state in a data-centric architecture. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 68–71). Stroudsburg, PA: Association for Computational Linguistics.

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., … Ivanov, A. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 338–345). Piscataway, NJ: IEEE. https://doi.org/10.1109/ASRU.2015.7404814

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Suggested citation:

Suendermann-Oeft, D., Ramanarayanan, V., Yu, Z., Qian, Y., Evanini, K., Lange, P., … Zechner, K. (2017). *A multimodal dialog system for language assessment: Current state and future directions* (Research Report No. RR-17-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12149

Action Editor: Beata Beigman Klebanov

Reviewers: Chee Wee Leong and Lei Chen