Measuring the Power of Learning.®

Research Report
ETS RR–17-20

Using Vision and Speech Features for Automated Prediction of Performance Metrics in Multimodal Dialogs

Vikram Ramanarayanan

Patrick Lange

Keelan Evanini

Hillary Molloy

Eugene Tsuprun

Yao Qian

David Suendermann-Oeft

April 2017

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

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

Using Vision and Speech Features for Automated Prediction of Performance Metrics in Multimodal Dialogs

Vikram Ramanarayanan,[1] Patrick Lange,[1] Keelan Evanini,[2] Hillary Molloy,[1] Eugene Tsuprun,[2] Yao Qian,[1] & David Suendermann-Oeft[1]

1 Educational Testing Service, San Francisco, CA
2 Educational Testing Service, Princeton, NJ

Predicting and analyzing multimodal dialog user experience (UX) metrics, such as overall call experience, caller engagement, and latency, among other metrics, in an ongoing manner is important for evaluating such systems. We investigate automated prediction of multiple such metrics collected from crowdsourced interactions with an open-source, cloud-based multimodal dialog system in the educational domain. We extract features from both the audio and video signals and examine the efficacy of multiple machine learning algorithms in predicting these performance metrics. The best performing audio features consist of multiple low-level audio descriptors—intensity, loudness, cepstra, pitch, and so on—and their functionals, extracted using the OpenSMILE toolkit, while the video features are bags of visual words that use 3D Scale-Invariant Feature Transform descriptors. We find that our proposed methods outperform the majority vote classification baseline in predicting various UX metrics rated by both the user and experts. Our results suggest that such automated prediction of performance metrics can not only inform the qualitative and quantitative analysis of dialogs but also be potentially incorporated into dialog management routines for positively impacting UX and other metrics during the course of the interaction.

Keywords Multimodal dialog; user experience; call rating; computer vision; machine learning; dialog management

doi:10.1002/ets2.12146

When building and deploying any spoken dialog system (SDS), it is imperative to understand how well the system is performing to ensure an optimal user experience (UX). While such an endeavor is crucial and relevant during the process of bootstrapping a dialog system for a new domain or application, it is equally important to measure UX and system performance metrics for an SDS that is more mature to ensure a high quality of service. Furthermore, the ability to rapidly prototype, evaluate, and monitor SDSs is particularly important for applications in the educational domain, because language learning and assessment applications require systems that deal gracefully with nonnative speech and varying cultural contexts.

Although much research has been conducted into the metrics one can use to quantify the performance and UX of an SDS (see, e.g., Danieli & Gerbino, 1995; Jiang et al., 2015; Möller, 2004; Pietquin & Hastie, 2013; Walker, Wright, & Langkilde, 2000; Yang, Levow, & Meng, 2012), much of this research has focused on the evaluation of dialog management strategies. The PARADISE approach in particular has been a popular approach that attempts a mapping between objective SDS metrics and subjective user ratings (Walker, Litman, Kamm, & Abella, 1997). Relatively less work has been done on predicting multiple such metrics *automatically* from the data alone. For instance, Schmitt, Schatz, and Minker (2011) showed that the use of 52 automatically extracted features based on speech recognition, language understanding, and dialog management features performed well on an evaluation conducted on the Carnegie Mellon University Let's Go Corpus. They suggested, however, that incorporating user emotional state information does not improve performance significantly. Evanini et al. (2008) presented a method to automatically predict caller experience in interactive voice response systems using primarily log information and a decision tree-based classification approach. Forbes-Riley and Litman (2004, 2011) investigated the automated prediction of metrics for the educational domain, examining the utility of multiple types of turn-level and contextual linguistic features for automatically predicting student emotions in human–human spoken tutoring dialogs.

While most studies in this area have focused on features extracted from speech, text, or log file data, there is little to no work, to our knowledge, on automatically predicting performance metrics from *video*-based features. We aim to bridge

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

this gap by using features extracted from the video data collected from each user's interaction with the dialog system to predict various performance metrics. In other words, this report attempts to further the state of the art in this field by (a) automating the prediction (b) of multiple UX and system performance SDS metrics by (c) investigating multiple video- and audio-based feature sets and machine learning algorithms as well as (d) the effect of factors like number of dialog states and task type. The remainder of the report is organized as follows: The next section describes the collection of data and ratings, while the following section outlines the features we automatically extract from the speech signal. We then describe our machine learning experiments in the subsequent section, followed by an in-depth analysis of the observed ratings and prediction results in the concluding section.

Data

Crowdsourcing Data Collection

We used Amazon Mechanical Turk for our crowdsourcing data collection experiments. Crowdsourcing, and particularly Amazon's Mechanical Turk, has been used in the past for assessing SDSs and for collecting interactions with SDSs (Jurcicek et al., 2011; McGraw, Lee, Hetherington, Seneff, & Glass, 2010; Rayner, Frank, Chua, Tsourakis, & Bouillon, 2011). We leveraged the open-source HALEF dialog system[1] to develop conversational applications within this crowdsourcing framework. The HALEF architecture and components have been described in detail in other publications (Ramanarayanan et al., 2017; Suendermann-Oeft, Ramanarayanan, Teckenbrock, Neutatz, & Schmidt, 2015; Yu et al., 2016). In addition to reading instructions and calling into the system, users were requested to fill out a 2- to 3-minute survey regarding the interaction. Approximately 88% of all participants self-reported as native speakers of English from all over the continental United States; 78% of participants were male. In all, we collected 1,133 conversations with approximately 41 hours of dialog data.

Spoken Dialog Tasks

We deployed four goal-oriented conversational tasks from common workplace communicative scenarios for the purposes of this experiment: responding to an offer of food, scheduling a meeting, interviewing for a job, and taking a customer's order.

The first spoken dialog task is a short conversation in which the system offers some food to the participant and the participant is expected to accept or decline the offer in a pragmatically appropriate manner. The second task provides the participant with a sample résumé stimulus, and the participant is instructed to act as a job candidate in an interview with an automated interviewer. In each case, the participant connects to the system and then proceeds to answer the sequence of questions posed by the automated coworker/interviewer. Depending on the semantic class of the participant's answer to each question (as determined by the output of the speech recognizer and the natural language understanding module), he or she is redirected to the appropriate branch of the dialog tree, and the conversation continues until all questions are answered.

Whereas the two aforementioned tasks are system-initiated dialog scenarios, the other two involve user-driven dialog. In the third task, the participant is 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 (e.g., What is your name? What toppings would you like on your pizza?) before replying with the appropriate response. Therefore this task might be more difficult than the other three, imposing more cognitive load on the user. In the fourth task, the participant is to arrange a time for a meeting with a coworker. Table 1 lists details of the various spoken dialog tasks used for the data collection.

Ratings

To better understand how the system performs when actual test takers call in, we asked all Turkers to rate various aspects of their interactions with the system on a scale from 1 to 5, 1 being least satisfactory and 5 being most satisfactory. Furthermore, we had six expert reviewers listen to between 30 and 45 full-call recordings each from the pizza item (from a subset of 162 calls in total[2]), examine the call logs, and rate each call on a range of dimensions (Suendermann, Liscombe,

Table 1 Four Conversational Tasks Deployed

Item	No. DS	No. DDS	No. of calls	Mean turn time (s)	Handling time (s)	
					Mean	SD
Food offer	1	1	303	4.0	45.8	18.7
Job interview	8	3	225	9.1	299.2	111.9
Pizza service	7	7	313	4.1	139.9	65.2
Meeting request	6	0	292	5.1	77.8	31.4

Note. Along with the number of dialog states for each task (No. DS), we also list the number of dialog states that required a speech recognition and subsequent language understanding hypothesis to go to the next dialog state (No. DDS) (as opposed to an inconsequential state that just moves to the next state after end of speech has been detected).

Pieraccini, & Evanini, 2010). We included expert ratings (who are speech technology researchers) because callers can potentially conflate system performance with their own performance on the test or react in certain ways due to lack of experience with the technology or the task, and having expert ratings allows us to analyze such potential biases. However, we first requested all experts to rate a smaller subset of calls and compare notes before doing a second pass to ensure interrater reliability. Table 2 provides descriptions of these different ratings in addition to information regarding whether the rating was performed by naive callers, dialog system experts, or both.

Table 2 The Various Rating Rubrics

Rating	Description	Caller	Expert
Caller experience	A qualitative measure of the caller's experience using the automated agent, with 1 for a very bad experience and 5 for a very good experience.	✓	✓
Caller engagement	A qualitative measure of caller's engagement with the task or the system, ranging from highly disengaged to highly engaged.	✓	✓
Intelligibility of system responses	This metric measures, on a scale from 1 to 5, how clear the automated agent is. A poor audio quality rating would be marked by frequent dropping in and out of the automated agent's voice or by muffled or garbled audio.	✓	
Audio quality of caller responses	This metric measures, on a scale from 1 to 5, how clear the caller audio is. A poor audio quality rating would be marked by user responses dropping in and out of the call or being muffled, garbled, echoing, or inaudible.		✓
Video quality of caller responses	This metric measures, on a scale from 1 to 5, the video quality of the call. A poor quality rating here would involve issues with lighting, other problems with the video (such as pixellation, blocking artifacts, nonconstant background), and if the user's head is not located in the center of the image as instructed in the caller guidelines.		✓
Qualitative latency score	Measures perceived system response time. How debilitating is the average delay between the automated agent's response from the time the user finishes speaking to the conversation?	✓	✓
Caller cooperation	A qualitative measure of caller's cooperation, or the caller's willingness to interact with the automated agent, with 1 for no cooperation and 5 for fully cooperative.		✓
System performance	A qualitative measure of how the system performed as per caller expectations and whether the system responses were appropriate.	✓	
System understanding degree	A qualitative measure of how well the system "understood" the caller.	✓	

Experiments

Visual Features

There is much work on computing video-based features in the computer vision literature (Forsyth & Ponce, 2011; Vedaldi & Fulkerson, 2010; Weinland, Ronfard, & Boyer, 2011). However, a large proportion of these features are computed on an image-by-image basis, not necessarily taking into account the spatiotemporal relationships between pixels and pixel regions in the sequence of images. We wanted to use a feature that explicitly captures spatiotemporal relationships in the image sequence for the subsequent classification task. Therefore we use 3D Scale-Invariant Feature Transform (SIFT) descriptors (Scovanner, Ali, & Shah, 2007) to represent videos in a bag-of-visual-words approach (Csurka, Dance, Fan, Willamowski, & Bray, 2004), which can be summarized as follows:

1. For each video in the data set, use ffmpeg[3] (or similar software) to extract image frames at a desired frame rate (we used one frame/s in our case, because we wanted to capture macro-level behavioral patterns over the entire video). Convert this into a 3D video matrix by concatenating all image frames.
2. Remove outlier frames, that is, any frame that lies more than 3 standard deviations away from the mean image.
3. Select N interest points at random.[4]
4. Extract N 3D SIFT features for each video in the data set using the procedure described in Scovanner et al. (2007).
5. Use a held-out portion of the data set to quantize the 3D SIFT descriptors into K clusters using K-means clustering.
6. Assign cluster labels to all SIFT descriptors computed for other videos in the data set using K-nearest-neighbor (KNN) clustering.
7. Finally, for each video, compute the histogram of cluster labels (also called a "signature") and use this as a K-dimensional feature descriptor for the video. Using such a histogram of cluster labels is more robust than using the raw 3D SIFT features and also allows us to build a more discriminative representation of a video, because some spatiotemporal patterns can occur in some videos more than others.

After some empirical experimentation, we chose free parameter values of $N = 50$ descriptors and $K = 64$ clusters for subsequent machine learning experiments.

Speech Features

We used OpenSMILE (Eyben, Weninger, Gross, & Schuller, 2013) to extract features from the audio signal, specifically, the standard openEAR emobase and emobase2010 feature sets containing 988 and 1,582 features, respectively, which are tuned for recognition of paralinguistic information in speech. These consist of multiple low-level descriptors — intensity, loudness, mel-frequency cepstral coefficients (MFCCs), pitch, voicing probability, F0 envelope, line spectral frequencies, and zero crossing rate, among others — as well as their functionals (such as standard moments). These feature sets have been shown to be comprehensive and effective for capturing paralinguistic information in various standard tasks (Eyben, Woellmer, & Schuller, 2010).

We also examined features that are currently used in automated speech scoring research, covering diverse measurements among lexical usage, fluency, pronunciation, prosody, and so on. In particular, following the feature extraction method described in Chen, Zechner, and Xi (2009), we used the *SpeechRater*^{SM} Automated Scoring service, a speech rating system that processes speech and its associated transcription to generate a series of features on the multiple dimensions of speaking skills, for example, speaking rate, prosodic variations, pausing profile, and pronunciation, which is typically measured by goodness of pronunciation (Witt, 1999) or its derivatives. For more details on these features, please see Table 3.

Machine Learning Experiments

We used SKLL,[5] an open-source Python package that wraps around the scikit-learn package (Pedregosa et al., 2011), to perform machine learning experiments. We experimented with a variety of learners to predict the various performance metric scores (as detailed in the Ratings section), including support vector classifiers (SVC), tree-based classifiers, and boosting-based classifiers, using prediction accuracy as an objective function for optimizing classifier performance.[6] We ran stratified 10-fold cross-validation experiments, where folds were generated to preserve the percentage of samples in

Table 3 Speaking Proficiency Features Extracted by SpeechRater

Category	Subcategory	No. of features	Example features
Prosody	Fluency	24	This category includes features based on the number of words per second, number of words per chunk, number of silences, average duration of silences, frequency of long pauses (≥0.5 s), and number of filled pauses (*uh* and *um*). See Zechner, Higgins, Xi, and Williamson (2009) for detailed descriptions of these features.
	Intonation and stress	11	This category includes basic descriptive statistics (mean, minimum, maximum, range, standard deviation) for the pitch and power measurements for the utterance.
	Rhythm	26	This category includes features based on the distribution of prosodic events (prominences and boundary tones) in an utterance as detected by a statistical classifier (overall percentages of prosodic events, mean distance between events, mean deviation of distance between events; Zechner et al., 2009) as well as features based on the distribution of vowel, consonant, and syllable durations (overall percentages, standard deviation, and Pairwise Variability Index; Chen & Zechner, 2011).
Pronunciation	Likelihood based	8	This category includes features based on the acoustic model likelihood scores generated during forced alignment with a native speaker acoustic model (Chen et al., 2009).
	Confidence based	2	This category includes two features based on the ASR confidence score: the average word-level confidence score and the time-weighted average word-level confidence score (Higgins, Xi, Zechner, & Williamson, 2011).
	Duration	1	This category includes a feature that measures the average difference between the vowel durations in the utterance and vowel-specific means based on a corpus of native speech (Chen et al., 2009).
Grammar	Location of disfluencies	6	This category includes features based on the frequency of between-clause silences and edit disfluencies compared to within-clause silences and edit disfluencies (Chen, Tetreault, & Xi, 2010; Chen & Yoon, 2012).
Audio quality	–	2	This category includes two scores based on MFCC features that assess the probability that the audio file has audio quality problems or does not contain speech input (Jeon & Yoon, 2012).

each class. We performed two sets of experiments. The first examined audio files at the *dialog turn* level, as opposed to the full-call level, as we want to be able to automatically predict scores given only audio information from a single turn.[7] Such a functionality could then eventually be integrated with dialog management routines to choose an appropriate next action based on the current caller experience or caller engagement rating, for example. The second set of experiments looked at both audio and video files at the level of the full call. Note that we only examined data from calls that were assigned ratings between 1 and 5 (eliminating NULL or spurious ratings). Furthermore, we did not examine automated prediction of the latency and system understanding degree ratings, because they would be better measured by system log information and spoken language understanding accuracy, respectively.

Observations and Results

Qualitative and Quantitative Performance Analysis

Figure 1 shows histograms of various call ratings as provided by callers (top) and experts (bottom). Although we obtained caller ratings from all calls collected, we only collected expert ratings from the pizza task. We observed that callers tended

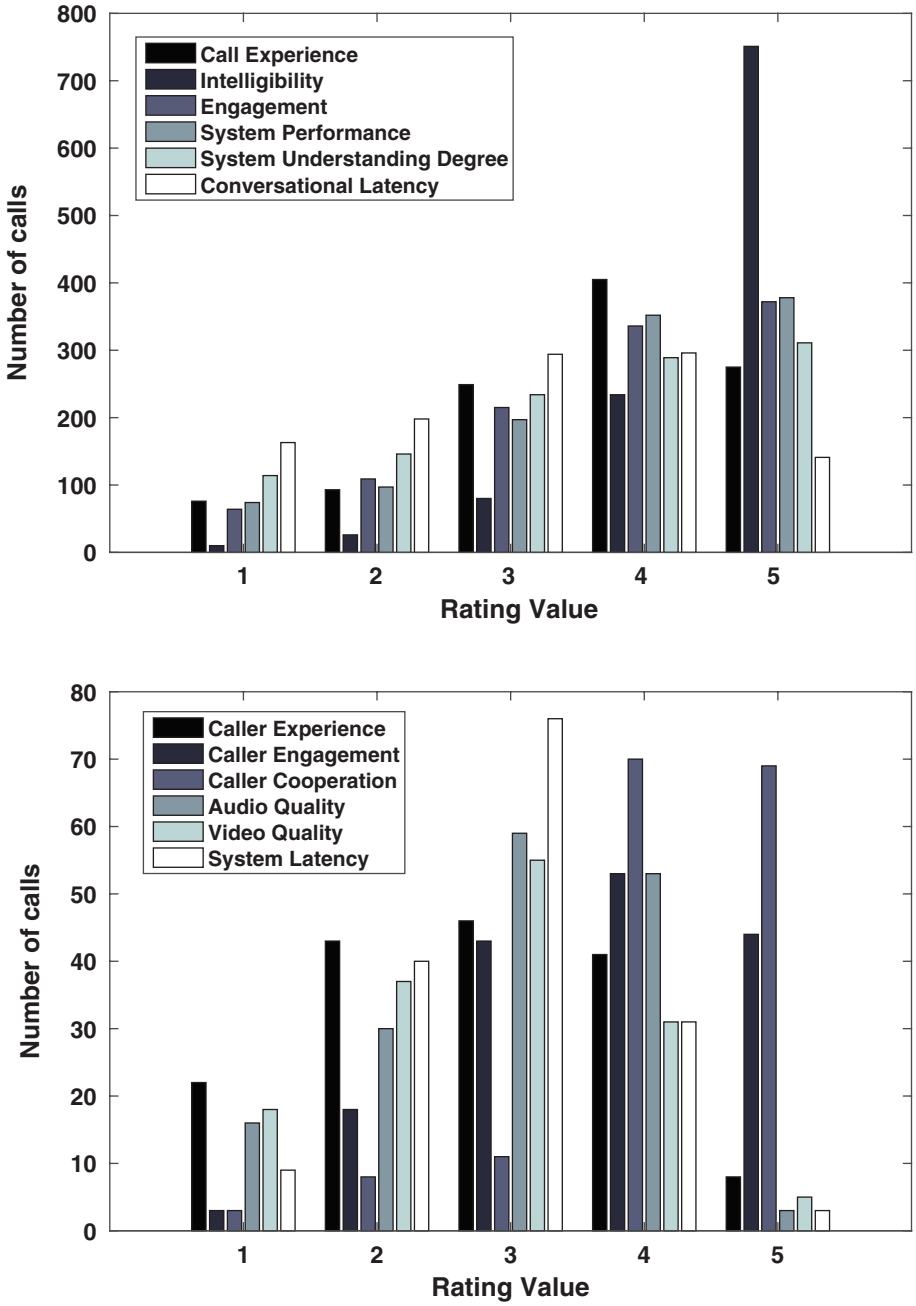

Figure 1 User ratings (top). Expert ratings (bottom).

to rate their call experiences much more highly ($M = 3.6$) than experts who listened to the calls later ($M = 2.6$), though they did come closer to experts when asked about other dimensions, such as conversation latency. Callers also tended to rate the system intelligibility and performance and understanding degree highly and self-reported a higher engagement rating. Experts tended to give above neutral ratings to caller engagement ($M = 3.7$) and cooperation ($M = 4.1$), which is in general agreement with the caller ratings and suggests that callers were invested in the task for the most part.

Automated Prediction Results

Let us first examine the results of prediction experiments performed at the dialog turn level, summarized in Tables 4 and 5. Table 4 shows 10-fold classification accuracies obtained by running stratified cross-validation experiments[8]

Table 4 Prediction Results on Audio-Based Features Computed From Each Dialog Turn of the Call

		emobase2010						emobase						SpeechRater					
Rating	Majority vote baseline	SVC	KNN	DT	GB	AB	RF	SVC	KNN	DT	GB	AB	RF	SVC	KNN	DT	GB	AB	RF
Caller																			
Call experience	35.5	31.8	30.1	31.3	38.2	34.2	**40.5**	30.5	28.4	31.8	37.3	33.4	40.0	28.0	27.8	28.0	34.0	33.4	36.4
Intelligibility	65.2	42.2	57.2	51.7	64.8	47.2	**66.8**	46.4	56.0	50.6	64.6	46.0	66.7	38.5	56.5	49.2	63.4	48.5	65.3
Engagement	34.2	33.0	28.5	29.9	38.1	33.1	**40.9**	26.3	27.3	30.0	37.9	34.1	40.7	28.4	28.1	28.3	32.8	31.2	36.0
System performance	32.4	28.3	29.1	30.9	37.6	32.7	**40.2**	24.4	27.5	30.6	37.3	34.3	40.1	25.9	27.8	27.8	34.1	32.8	35.5
Expert (pizza)																			
Caller experience	30.1	25.7	27.1	27.7	34.5	30.0	**35.2**	28.6	27.5	27.3	34.2	31.6	**35.6**	25.6	27.3	26.8	30.5	26.8	**35.4**
Caller engagement	33.9	28.2	27.9	29.3	31.1	30.0	33.3	25.6	27.7	29.8	31.3	28.7	32.0	28.8	28.0	29.1	31.3	29.1	**36.4**
Caller cooperation	45.3	41.6	42.3	42.1	42.1	41.0	45.9	37.9	39.6	40.2	42.6	39.6	**46.5**	34.8	40.1	36.4	42.4	27.7	45.5
Audio quality	34.6	24.6	32.7	34.3	39.8	33.3	**41.9**	30.3	28.7	32.2	39.0	37.1	39.5	29.9	30.3	29.2	34.8	28.1	37.4

Note. AB = AdaBoost; DT = decision trees; GB = gradient boosting; KNN = *K* nearest neighbor; RF = random forests; SVC = linear support vector classifier machines. The best-performing systems are highlighted in bold.

Table 5 Caller Rating Prediction Results for the Best Performing Learner (RF) and Feature Set (emobase2010) in Table 4 Broken Down by Task

	Classifier accuracy				
Rating	Food offer	Interview	Pizza	Meeting	Overall
Call experience	39.6	40.6	38.1	45.5	40.5
Intelligibility	68.9	66.0	61.5	78.1	66.8
Engagement	35.1	46.0	35.4	46.1	40.9
System performance	39.0	42.0	36.5	45.3	40.2

using six different classifiers—linear SVC machines, KNN, decision trees, gradient boosting, AdaBoost, and random forests (RF)—on each feature set extracted from the audio corresponding to each dialog turn.[9] Recall that although performance metrics are rated at the level of the full-call recording, we assign the same rating to the audio associated with each dialog turn of that full-call recording for the purposes of this experiment. We see that the RF classifier generally performs best in most cases, while the best performance is obtained using the emobase2010 feature set. Also note that while emobase and SpeechRater perform only marginally worse, they are increasingly lower dimensional as compared to the emobase2010 feature set and therefore might find utility in some applications. Additionally, the best performing system for each rating significantly outperforms the majority vote baseline. We also experimented with feature scaling but do not report the results here as the results trended similarly with those shown in Table 4.

Table 5 provides insight into how different tasks performed on the caller ratings prediction task for the emobase2010 feature set and a RF classifier. We observe that the accuracies were higher than average for the meeting and interview tasks as compared to the pizza and food offer tasks. This trend can be explained by the longer duration of utterances in the interview and meeting scheduling tasks, soliciting more elaborate user input. Table 1 shows the average duration of speech utterances per task.

Now, let us consider the results of experiments performed at the level of the full-call recording; these are summarized in Table 6. Note that for this level of analysis, we only considered the best performing audio feature—the emobase2010 feature set extracted using OpenSMILE—as opposed to all three speech feature sets examined in Table 4. Furthermore, we only tested audio-only features to predict audio quality ratings and video-only features to predict video-only ratings. We generally observe that (a) the best performing feature sets outperform the majority vote baseline in all rating categories, while (b) RF classifiers still perform well for this experiment, and other classifiers, such as the KNN, DT, and GB, also perform competently in predicting certain ratings; moreover, (c) the fusion of emobase2010 audio- and video-based 3D SIFT bag-of-visual-words features performs better than audio or video features alone. An exception to the latter

Table 6 Prediction Results for Video and Audio Features Computed Over the Entire Call

		Audio (*emobase2010*) features only						Video features only						Fused video and audio features					
Rating	Majority vote baseline	SVC	KNN	DT	GB	AB	RF	SVC	KNN	DT	GB	AB	RF	SVC	KNN	DT	GB	AB	RF
Caller																			
Call experience	36.9	28.2	27.7	32.2	37.9	31.7	40.9	31.8	29.5	26.5	32.0	28.7	36.5	29.6	27.8	33.0	37.5	30.6	**41.9**
Intelligibility	68.0	64.2	59.9	56.3	67.2	60.1	**69.9**	66.2	60.4	53.0	62.1	63.9	68.6	61.0	59.9	56.4	66.3	64.2	**69.6**
Engagement	34.3	29.5	31.4	28.9	41.0	35.4	**42.8**	33.0	28.5	29.9	38.1	33.1	40.9	32.2	31.4	31.0	41.2	36.6	41.8
System performance	34.6	27.4	28.8	29.9	39.4	34.0	41.8	28.3	29.1	30.9	37.6	32.7	40.2	27.1	28.8	29.4	37.5	32.7	**42.3**
Expert (pizza only)																			
Caller experience	29.1	33.6	33.3	30.2	34.3	32.3	37.7	24.8	24.1	27.9	20.9	29.6	33.1	36.0	33.3	24.6	**40.7**	32.3	37.9
Caller engagement	32.7	25.3	27.4	**42.5**	37.0	32.2	31.5	22.9	24.0	21.9	24.0	27.4	29.5	21.2	27.4	40.4	37.0	33.6	37.7
Caller cooperation	43.4	38.5	**46.6**	39.0	40.4	38.4	44.6	35.6	41.1	38.3	37.7	45.1	45.2	38.3	**46.6**	**46.6**	36.3	39.1	41.8
Audio quality	37.1	39.8	26.4	35.7	39.6	30.2	**40.2**	–	–	–	–	–	–	–	–	–	–	–	–
Video quality	37.7	–	–	–	–	–	–	25.9	31.2	32.1	39.3	21.7	**39.8**	–	–	–	–	–	–

Note. AB = AdaBoost; DT = decision trees; GB = gradient boosting; KNN = *K* nearest neighbor; RF = random forests; SVC = linear support vector classifier machines. The best-performing systems are highlighted in bold.

point is in the case of caller engagement (for both experts and callers), where the audio-only features perform better. This suggests that our video features are not capturing enough significant information regarding the callers' facial expressions and gestures, which have been shown to be important markers in characterizing engagement. However, this poor performance is not surprising, given that the video feature extraction procedure involves the computation of space-time interest points at random, which does not guarantee that salient regions on the face and body of the caller are analyzed. While improving this area of the procedure to select more relevant and meaningful space-time interest points is definitely a priority for future research, it is nonetheless interesting to note that even though the current procedure for 3D SIFT feature extraction selects interest points at random, it performs competently in predicting video quality ratings (above the baseline) as well as other ratings (when fused with audio features), which suggests that these features are already capturing meaningful discriminative information and can only perform better with more careful interest point selection.

Summary and Outlook

We have examined how features extracted from just the audio signal can be used to automatically predict different spoken dialog performance metrics, such as call experience, engagement, intelligibility, and system performance. We have further analyzed callers' self-ratings vis-à-vis experts' ratings and found that callers tend to generally rate their experiences higher than experts do.

Many important avenues for future research remain. First, we would like to conduct a deeper investigation of video-based features, including the choice of more meaningful space-time interest points (using methods such as difference of Gaussian filtering or face/body/pose trackers to find more relevant points of interest) as well as other useful robust image descriptors, such as histograms of oriented gradients or Fisher vectors. In addition, we plan to explore more meaningful higher level face- and emotion-based features obtained using face-tracking algorithms. Second, we intend to look into better feature fusion and machine learning methods to improve prediction accuracy. Finally, we envision incorporating such prediction modules into real-time dialog management routines in multimodal dialog settings to improve UX and system performance on the fly during interactions.

Acknowledgments

We would like to thank Lydia Rieck, Elizabeth Bredlau, Katie Vlasov, Juliet Marlier, Phallis Vaughter, Nehal Sadek, and Veronika Laughlin for their help in designing the conversational tasks and Robert Mundkowsky and Chong Min Lee for their engineering support.

Notes

1 http://halef.org
2 We only had expert raters rate a small subset of the calls owing to time and availability constraints.
3 https://ffmpeg.org/
4 While the computer vision literature contains many methods for interest-point detection, we chose to select these at random in the interest of speeding up processing time.
5 https://github.com/EducationalTestingService/skll
6 We also ran initial experiments using the quadratic weighted kappa metric (which takes into account the ordered nature of the categorical labels) as the objective function but found that it performed similarly.
7 We only examined audio features at the turn level in this study owing to time and resource constraints required to segment the video into user and system turns. However, we plan to examine the utility of video features computed at the turn level in future research.
8 Owing to the possibility that data from the same call might be present in both train and test sets of each fold, we also ran 10-fold cross-validation with call ordering to avoid this. However, the results we obtained had similar trends to what we present in Table 4 and are therefore omitted for brevity.
9 Owing to poor audio quality of some of the recordings, we were only able to extract SpeechRater features for 84% of the data set and therefore report results on that subset.

References

Chen, L., Tetreault, J., & Xi, X. (2010). Towards using structural events to assess non-native speech. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT* (pp. 74–79.). Stroudsburg, PA: Association for Computational Linguistics.

Chen, L., & Yoon, S.-Y. (2012). Application of structural events detected on ASR outputs for automated speaking assessment. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association* (pp. 767–770). Baixas, France: International Speech Communication Association.

Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association* (pp. 1861–1864). Baixas, France: International Speech Communication Association.

Chen, L., Zechner, K., & Xi, X. (2009). Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (pp. 442–449). Stroudsburg, PA: Association for Computational Linguistics.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). *Visual categorization with bags of keypoints*. Paper presented at the Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV), Prague, Czech Republic.

Danieli, M., & Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In *Empirical methods in discourse interpretation and generation. Papers from the 1995 AAAI spring symposium, US-Stanford CA* (pp. 34–39). Menlo Park, CA: AAAI Press.

Evanini, K., Hunter, P., Liscombe, J., Suendermann, D., Dayanidhi, K., & Pieraccini, R. (2008). Caller experience: A method for evaluating dialog systems and its automatic prediction. In *2008 IEEE Spoken Language Technology Workshop (SLT 2008) Proceedings* (pp. 129–132). Piscataway, NJ: IEEE. https://doi.org/.1109/SLT.2008.4777857

Eyben, F., Weninger, F., Gross, F. & Schuller, B. (2013). Recent developments in opensmile, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international Conference on Multimedia* (pp. 835–838). New York, NY: ACM.

Eyben, F., Woellmer, M., & Schuller, B. (2010). *OpenSMILE: The Munich open speech and music interpretation by large space extraction toolkit* (Verion 1.0.1). Retrieved from http://web.stanford.edu/class/cs224s/hw/openSMILEmanual.pdf

Forbes-Riley, K., & Litman, D. J. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Human language technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 201–208). Stroudsburg, PA: Association for Computational Linguistics.

Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, *53*, 1115–1136.

Forsyth, D., & Ponce, J. (2011). *Computer vision: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. M. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*, 282–306. https://doi.org/10.1016/j.csl.2010.06.001

Jeon, J. H., & Yoon, S.-Y. (2012). Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association* (pp. 1275–1278). Baixas, France: International Speech Communication Association.

Jiang, J., Awadallah, A. H., Jones, R., Ozertem, U., Zitouni, I., Kulkarni, R. G., & Khan, O. Z. (2015). Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on the World Wide Web* (pp. 506–516). Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2736277.2741669

Jurcicek, F., Keizer, S., Gašic, M., Mairesse, F., Thomson, B., Yu, K., & Young, S. (2011). Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association* (pp. 3061–3064). Baixas, France: International Speech Communication Association.

McGraw, I., Lee, C.-Y., Hetherington, I. L., Seneff, S., & Glass, J. (2010). Collecting voices from the cloud. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *LREC 2010: Seventh International Conference on Language Resources and Evaluation* (pp. 1576–1583). Paris, France: European Language Resources Association.

Möller, S. (2004). *Quality of telephone-based spoken dialogue systems*. New York, NY: Springer Science & Business Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pietquin, O., & Hastie, H. (2013). A survey on metrics for the evaluation of user simulations. *Knowledge Engineering Review*, *28*(1), 59–73. https://doi.org/10.1017/S0269888912000343

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: Towards natural user interfaces to everything* (pp. 295–310). New York, NY: Springer.

Rayner, E., Frank, I., Chua, C., Tsourakis, N., & Bouillon, P. (2011). For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language call application. In *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011)* (pp. 117–120). Baixas, France: International Speech Communication Association.

Schmitt, A., Schatz, B., & Minker, W. (2011). Modeling and predicting quality in spoken human–computer interaction. In *Proceedings of the SIGDIAL 2011 Conference* (pp. 173–184). Stroudsburg, PA: Association for Computational Linguistics.

Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia* (pp. 357–360). New York, NY: ACM. https://doi.org/10.1145/1291233.1291311

Suendermann, D., Liscombe, J., Pieraccini, R., & Evanini, K. (2010). "How am I doing?": A new framework to effectively measure the performance of automated customer care contact centers. In A. Neustein (Ed.), *Advances in speech recognition: Mobile environments, call centers and clinics* (pp. 155–179). New York, NY: Springer.

Suendermann-Oeft, D., Ramanarayanan, V., Teckenbrock, M., Neutatz, F., & Schmidt, D. (2015). HALEF: An open-source standard-compliant telephony-based modular spoken dialog system: A review and an outlook. In G. Geunbae Lee, H. Kook Kim, M. Jeong, & J.-H. Kim (Eds.), *Natural language dialog systems and intelligent assistants* (pp. 53–61). New York, NY: Springer.

Vedaldi, A., & Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1469–1472). New York, NY: ACM. https://doi.org/10.1145/1873951.1874249

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 271–280). Stroudsburg, PA: Association for Computational Linguistics.

Walker, M., Wright, J., & Langkilde, I. (2000). Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system. In P. Langley (Ed.), *Proceedings of the 17th International Conference on Machine Learning* (pp. 1111–1118). San Francisco, CA: Morgan Kaufmann.

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, *115*, 224–241. https://doi.org/10.1016/j.cviu.2010.10.002

Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning* (Unpublished PhD thesis). University of Cambridge, England.

Yang, Z., Levow, G.-A., & Meng, H.-Y. (2012). Predicting user satisfaction in spoken dialog system evaluation with collaborative filtering. *Journal of Selected Topics in Signal Processing*, *6*, 971–981. https://doi.org/10.1109/JSTSP.2012.2229965

Yu, Z., Ramanarayanan, V., Mundkowsky, R., Lange, P., Ivanov, A., Black, A. W., & Suendermann-Oeft, D. (2016, January). *Multimodal HALEF: An open-source modular web-based multimodal dialog framework*. Paper presented at the International Workshop on Spoken Dialog Systems (IWSDS 2016), Saariselka, Finland.

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

Suggested citation:

Ramanarayanan, V., Lange, P., Evanini, K., Molloy, H., Tsuprun, E., Qian, Y., & Suendermann-Oeft, D. (2017). *Using vision and speech features for automated prediction of performance metrics in multimodal dialogs* (Research Report No. RR-17-20). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12146

Action Editor: Beata Beigman Klebenov

Reviewers: Saad Khan and Chee Wee Leong

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/