Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring*

Vikram Ramanarayanan
ETS Research
90 New Montgomery St
San Francisco, CA
vramanarayanan@ets.org

Chee Wee Leong
ETS Research
600 Rosedale Road
Princeton, NJ
cleong@ets.org

Lei Chen
ETS Research
600 Rosedale Road
Princeton, NJ
lchen@ets.org

Gary Feng
ETS Research
600 Rosedale Road
Princeton, NJ
gfeng@ets.org

David Suendermann-Oeft
ETS Research
90 New Montgomery St
San Francisco, CA
suendermann-oeft@ets.org

ABSTRACT

We analyze 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 compute both time-aggregated and time-series based features from these data streams—the former being statistical functionals and other cumulative features computed over the entire time series, while the latter, dubbed histograms of cooccurrences, capture how different prototypical body posture or facial configurations co-occur within different time-lags of each other over the evolution of the multimodal, multivariate time series. We examine the relative utility of these features, along with curated speech stream features in predicting human-rated scores of multiple aspects of presentation proficiency. We find that different modalities are useful in predicting different aspects, even outperforming a naive human inter-rater agreement baseline for a subset of the aspects analyzed.

Keywords

multimodal analysis, speech recognition, emotion tracking, motion capture, face tracking, presentation assessment

1. INTRODUCTION

Accurate assessment of performance in presentational skills is gaining importance in educational institutions and within workplaces, where the effectiveness and delivery of presentation tasks are critical in securing teacher licensure, job offers, business contracts, etc. Multimodal data capture techniques based on video, audio and motion feeds provide a rich source of information for such assessment, but the complexity of these data streams brings with it a need for signal analysis tools to automatically process and make sense of this data.

Researchers have made many advances towards understanding and modeling these multimodal data streams. For example, Naim et al. [10] analyzed job interview videos of internship-seeking students and found, using machine learning techniques, that prosody, language and facial expression features were good predictors of human ratings of desirable interview traits such as excitement, friendliness or engagement. Nguyen et al. [11] proposed a computational framework to predict the hiring decision using non-verbal behavioral cues extracted from a dataset of 62 interview videos. While there is much work on automatic recognition of one or more social cues and verbal and nonverbal behavioral traits in the speech and larger multimodal analysis communities (see for example [9, 12, 14, 15]), this problem has been also been highlighted as a particularly important one, evidenced by a number of challenges at international conferences on related topics [17, 18]. Recently [2] also presented a framework for collecting multimodal data of subjects giving presentations for purposes of automated assessment. They further presented preliminary results suggesting that basic features in the speech content and delivery, and movements related to the head, body and hands significantly predicted holistic human ratings of public speaking skills. While those works focus on (i) time-aggregated, and (ii) hand-selected features that can each have an explicit interpretation, in this paper we focus on time-series features that are also motivated from an interpretability standpoint, but are relatively high-dimensional, since they have to encapsulate information about the evolution of the entire time-series. In addition, we build upon and extend this work to predict not only a holistic score of presentation proficiency, but other aspects as well, using a combination of features derived from speech, visual and Kinect data.

ICMI '15 November 09 - 13, 2015, Seattle, WA, USA

ACM 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2820765

Table 1: *Performance standards adapted from the Public Speaking Competence Rubric (PSCR) [16] that human raters were asked to score each multimodal presentation on.*

Score Dimension	Shorthand	Description of Item Competency
1	Intro	Formulate an introduction that orients the audience to the topic and speaker
2	Org	Use an effective organizational pattern
3	Conc	Develop a conclusion that reinforces the thesis and provides psychological closure
4	WC	Demonstrate a careful choice of words
5	VE	Effectively use vocal expression and paralanguage to engage the audience
6	NVB	Demonstrate nonverbal behavior that reinforces the message
7	AudAdap	Successfully adapt the presentation to the audience
8	VisAid	Skillfully make use of visual aids
9	Persuasion	Construct an effectual persuasive message with credible evidence
10	Holistic	Overall holistic performance

Despite the research advances made so far in multimodal signal analysis and presentation scoring, there is little work that explicitly models the temporal evolution of these signals and exploits this information for presentation scoring and understanding. Explicitly modeling temporal information in such data is important because a person's presentation competency need not stay constant over the course of the presentation – he/she could get fatigued over time, or be more nervous at the very beginning (resulting in repetitive, cyclic fidgeting behavior), but gradually settle into a comfort zone later. For similar reasons his/her body language and emotional state can also fluctuate over the time series. However, current feature extraction approaches that aggregate information across time are not able to explicitly model temporal cooccurrence patterns; consider for instance that a certain prototypical body posture follows a second particular posture in a definitive pattern during certain parts of the presentation. Capturing such patterns might help us (i) explicitly understand the predictive power of different features (such as the occurrence of a given emotion) in temporal context (such as how often did this emotional state occur given the previous occurrence of another emotional state), thus allowing us to (ii) obtain features that are more interpretable. It is this gap that we attempt to bridge in this study. Specifically, we propose a feature based on histograms of cooccurrences [19, 20, 13] that models how different "template" body postures or facial expressions co-occur within different time lags of each other in a particular time series. Such a feature explicitly takes into account the temporal evolution of face/body posture and facial features in different presentation contexts. This feature has been previously shown to perform well on phone classification tasks [13] as well as for unsupervised pattern discovery [19, 20]. We aim to explore how much of a benefit such time-series-based modeling can provide in assessment of presentation and interview performance.

The rest of the paper is organized as follows: Section 2 goes into the details of the multimodal data corpus, including the tasks, data collection and processing, and human scoring of different aspects of presentation proficiency. We then describe the different Kinect and speech-based features in Section 3, followed by the results of the regression experiments for presentation score prediction using these features in Section 4.

2. DATA

2.1 Assessment Tasks and Multimodal Data Collection

Five public speaking tasks were utilized for data collection. Among these tasks, the first one, task A, was an "ice-breaker", in which the speaker introduced him or herself; this task is not analyzed due to the personally identifiable information involved. Tasks B and C were modeled after prepared informational speeches, in which the speaker was given a pre-prepared slide deck and up to 10 minutes to prepare for the presentation. Task B was a business presentation, where the speaker was to present a financial report. Task C is a simulated teaching task on a topic targeting middle school students. The other two tasks were persuasive and impromptu speeches. Task D asked speakers to consider a movie they did not like but nonetheless recommend it to others. Task E asked speakers to consider a place inconvenient to live in and discuss the benefits of living there. Note that there has no visual aid for for Tasks D and E. Figures 1 and 2 show examples of the different multivariate data streams recorded.

We collected multimodal data using the following equipment and software tools: (a) Microsoft Kinect (Windows Version 1) for recording $3D$ body motions, (b) Brekel Pro Body Kinect tracking software (v1.30 64 bit version) for recording 58 body joints' motion traces in the Biovision hierarchical data format (BVH), and (c) a JVC Everio GZ-HM35BUSD digital camcorder for audio/video recording. Note that the camcorder was mounted together with Kinect on a tripod. Both Kinect and camcorder were placed 1.83m away from the front of the speaking zone that was marked on the ground. Additionally, during Task B and C, a SMART Board projector system was used to show the PowerPoint slides.

17 volunteers were recruited from within a non-profit organization, with ten male participants and seven female participants. Seven of the participants were experienced public speakers from the Toastmasters Club. The rest varied widely in their experience in public speaking. After being familiarized with the recording equipment, participants were informed that they were expected to speak for 4 to 5 minutes for Task B and C and 2 to 3 minutes for Task D and E. For Tasks B and C, which involved PowerPoint slides, they were given 10 minute to prepare for their presentation. They

Figure 1: Example of face tracking.

were not allowed to bring notes during the presentation. In Task D and E, the participants were given no preparation time. They would start speaking as soon as they were given the topic of the impromptu speech. Before each recording, the speaker was asked to clap, which served as a signal synchronizing the multimodal data. Data from 3 speakers were lost due to equipment failure. In total, we obtained 56 presentations from 14 speakers (4 per speaker) with complete multimodal recordings. After getting raw recordings from our lab sessions, the motion and video data streams were synchronized.

2.2 Human Rating

Since the ultimate goal of this study will be developing a valid assessment for measuring public speaking skills via presenters' multimodal behaviors, we chose the Public Speaking Competence Rubric (PSCR) [16] as an assessment rubric due to its favorable psychometric properties. Using the PCSR tailored to our tasks, human raters scored these presentation videos along 10 dimensions that represent various aspects of presentation proficiency on a five-point Likert scale from 0 to 4 [16]. See Table 1 for the complete list of scoring dimensions.

Five raters were recruited from within an educational testing company. Two expert raters had background in oral communication/public speaking instruction at the higher education level. The other three (non-expert raters) had extensive experience in scoring essays, but not in scoring public speaking performances. For reliability purposes, the presentations were double-scored. In the event that the scores between two raters were discrepant, the following adjudication process was used to generate final scores. If the first two raters *did not* agree with each other, a third rater (expert) was brought in to make another judgment, and the final score assigned was the average of all three scores. However, in the event that the first two raters agreed with each other, that score was used as the final score.

2.3 Head pose and eye gaze

A successful presentation entails speaker engagement with the audience, which translates to head postures and eye gazes that are necessarily directed towards the audience. Here, we extract a set of features that target these aspects of the presentation. Head postures are approximated using the rotation attribute (i.e., pitch, yaw, and roll) of the head through Visage's SDK FaceTrack[1], a robust head and face tracking engine. See Figure 1. The tracking is activated if and only if the detector has detected a face in the current frame. Additionally, gaze directions are approximated through the *gazeDirectionGlobal* attribute of the Visage tracker SDK, which tracks gaze directions taking into account both head pose and eye rotation. Note that, different from head rotation, gaze directions represent estimated "eyeball" directions regardless of head postures, and can potentially measure a speaker's level of engagement with the audience. Thus, for each presentation, we used the time evolution of basic head pose measurements (Cartesian X, Y, Z coordinates along with pitch, yaw, and roll) as well as gaze tracking information over the entire presentation to extract features.

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

2.4 Emotion tracking

Facial expressions from presenters also contribute to an effective presentation. Therefore, we utilized an off-the-shelf emotion detection toolkit, Emotient's FACET SDK[2], to analyze facial expressions. FACET outputs the intensity (ranging from 0 to 1) and confidence values for seven primary emotions (i.e., anger, contempt, disgust, joy, fear, sadness and surprise). See Figure 2. We primarily focus on the time-evolution of the estimated intensities of different emotional states for each presentation in order to compute features. Note that we faced certain technical challenges during emotion intensity tracking, such as: (a) the presenters' faces were relatively small as compared to the screen size due to the large distance from the camera, (b) considerable background brightness from using a SmartBoard during the informative task, and (c) that presenters frequently turned their heads during their presentations, which sometimes resulted in inaccurate tracking output. However, since we are considering the entire time series of emotion data as opposed to any one particular frame, we hope to minimize the effects of tracking errors.

3. METHOD

3.1 Computing time-aggregated Kinect features

For time-aggregated Kinect features, we computed statistical functionals of certain body point markers that correlated the best with the human-rated holistic scores and that captured the degree of locomotion and hand movement. A basic feature set was extracted based on the statistical functionals (such as the mean and standard deviations of the hip markers, hand movement markers and their speeds).

3.2 Computing histograms of co-occurrences (HoC) features

Recall that one of the primary goals of this work is to investigate the efficacy of time-series features computed on different multimodal time series vis-a-vis scoring of multimodal presentations. The motivation is that explicitly examining and modeling the evolution of each of these time series will result in richer features as opposed to time-aggregated features. With this in mind, we elucidate below a general methodology to compute such a feature called histograms of cooccurrences (or HoC) that can be applied to any multivariate time-series – in this paper we compute this feature for the Kinect, Face/Gaze and Emotion data streams. Note that we compute *one* high-dimensional HoC feature vector for each time-series; this is particularly useful in cases like ours, where we have one rating for an entire time-series. Further, the advantage of this feature vector over conventional time-aggregated one is that it explicitly encapsulates information regarding temporal co-occurrence patterns; so, for example, it would model how often a certain prototypical body posture (such as folded hands) follows a second prototypical posture (say, an open stance) in a definitive pattern during different parts of the presentation.

So that being said, the idea behind the histogram of cooccurrence (HoC) feature is to count the number of times different prototypical body postures co-occur with each other *at different time lags* over the course of the time series. As

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

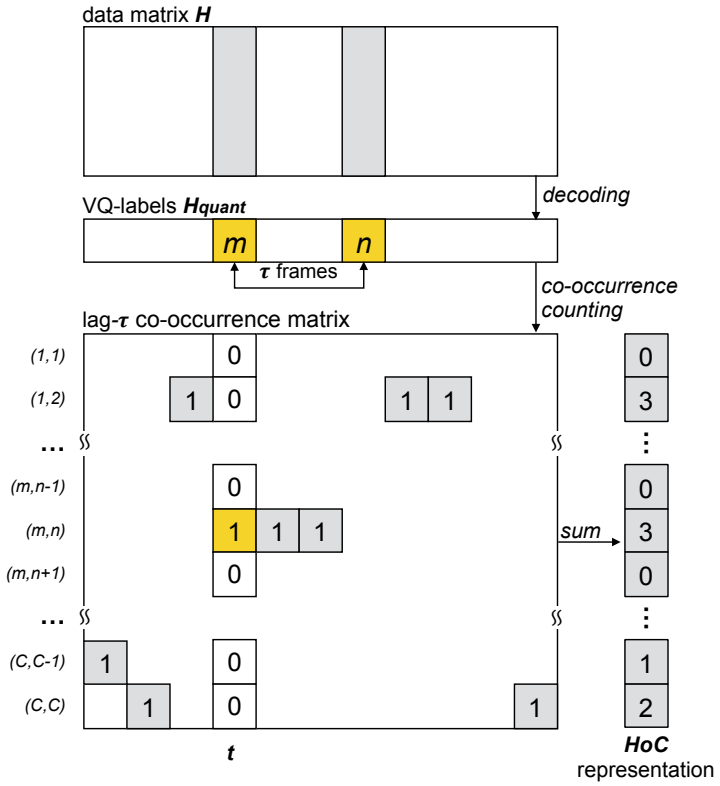

Figure 3: Schematic depiction of the computation of histograms of cooccurrences (HoC) (adapted from [13]). For a chosen lag value, τ, and a time step t, if we find labels m and n occurring τ time steps apart (marked in gold), we mark the entry of the lag-τ cooccurrence matrix corresponding to row (m, n) and the t^{th} column with a 1 (corresponding entry also marked in gold). We sum across the columns of this matrix (across time) to obtain the lag-τ HoC representation.

to what these prototypical body postures are – while this is an interesting research question in itself, for the purposes of this paper we use cluster centroids derived from simple K-means clustering on the space of body postures (in the training dataset) as prototypical body postures. We experimented with different cluster sizes $(16, 32, 64)$ and found that 32 clusters gave us the best empirical performance on the prediction task described below.

Once we perform this clustering, we can replace each frame of the input time series data matrix $\mathbf{H}$ with the best matching cluster label (corresponding to the cluster to which it belongs). This way, the data matrix is now represented by a single row vector of cluster labels, $\mathbf{H}_{quant}$. A HoC-representation of *lag* τ is then defined as a vector where each entry corresponds to the number of times all pairs of cluster labels are observed τ frames apart. In other words, we construct a vector of lag-τ co-occurrences where each entry (m, n) signifies the number of times that the input sequence of activation frames is encoded into a cluster label m at time t (in the row vector H_{quant}), while encoded into cluster label n at time $t + \tau$ [19, 20, 13]. By stacking all (m, n) combinations, each interval can be represented by a

Table 2: *Speaking proficiency features extracted by our speech rating engine, SpeechRater.*

Category	Sub-category	# 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 ($\geq$ 0.5 sec.), number of filled pauses (*uh* and *um*). See [22] for detailed descriptions of these features.
	Intonation & 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 (promincences 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) [22] as well as features based on the distribution of vowel, consonant, and syllable durations (overall percentages, standard deviation, and Pairwise Variability Index) [5].
Pronunciation	Likihood-based	8	This category includes features based on the acoustic model likelihood scores generated during forced alignment with a native speaker acoustic model [6].
	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 [7].
	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 [6].
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 [3],[4].
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 [8].

single column vector where the elements express the sum of all C^2 possible lag-τ co-occurrences (where C is the number of clusters; in our case, 32). See Figure 3 for a schematic of the HoC feature computations. We can repeat the procedure for different values of τ, and stack the results into one "supervector". Note however, that the dimensionality of the HoC feature increases by a factor of C^2 for each lag value τ that we want to consider. In our case, we empirically found that choosing four lag values of 1 to 10 frames (corresponding to 100-1000ms) gave an optimal prediction performance on regression experiments described below.

3.3 Computing speech features

Regarding measuring speech delivery skills demonstrated in public speaking, we included features widely used in automated speech scoring research area, covering diverse measurements among lexical usage, fluency, pronunciation, prosody, and so on. In particular, following the feature extraction method described in [6], we used *SpeechRater*, a speech rating system that processes speech and its associated transcription to generate a series of features on the multiple dimensions of speaking skills, e.g., speaking rate, prosodic variations, pausing profile, and pronunciation, which typically is measured by Goodness of Pronunciation (GOP) [21] or its derivatives. For more details on these features, please see Table 2.

3.4 Regression experiments

We used linear support vector machines (SVM) to perform regression experiments [1] on each of the 10 scoring dimensions with leave-one-speaker-out (or 14-fold) cross-validation. We experimented with both linear as well as radial basis function (RBF) kernels and empirically found that the former performed better on the prediction task. This could be due to the large dimensionality of the HoC feature space. We further tuned hyperparameters using a grid-search method.

4. OBSERVATIONS AND RESULTS

Table 3 lists the performance of various feature sets in predicting different human-rated dimensions of the multimodal presentation. We compare the performance of several feature sets—speech features obtained from SpeechRater, time-aggregated Kinect features, time-series HoC-based Kinect, Face and Emotion features as well as their combinations—as measured by the magnitude of Pearson correlation with the final human-adjudicated score. We also present, the Pearson correlations between the first and second human raters (denoted for $\rho_{R_1R_2}$), and finally, for benchmarking purposes, the Pearson correlation between each of the *individual* human raters' scores and the final human-adjudicated score. These last two correlation numbers can be thought of as a upper bound on the prediction performance.

Table 3: *Performance of various feature sets in predicting ten different aspects of multimodal presentation proficiency. The numbers represent Pearson correlations with the final human-adjudicated score (except for the row enclosed by dashed lines, which represents Pearson correlations between scores predicted by human raters 1 and 2, $\rho_{R_1R_2}$). The best machine score in each dimension relative to $\rho_{R_1R_2}$ are marked in* ***bold****. Also shown as a reference benchmark, is the Pearson correlations between each of the raters (1,2) and the final human-adjudicated score.*

Rater	Feature Set	Score Dimension									
		1	2	3	4	5	6	7	8	9	10
		Intro	Org	Conc	WC	VE	NVB	AudAdap	VisAid	Persuasion	Holistic
Machine	Kinect HoC	0.13	0.14	0.16	**0.23**	0.01	0.25	0.06	0.66	0.24	0.03
	Kinect Aggregated	0.12	**0.53**	0.09	0.08	0.16	0.26	0.31	0.03	0.11	0.12
	Speech	0.28	0.34	0.03	0.12	0.37	0.22	0.30	0.75	**0.48**	0.44
	Kinect Both	0.13	0.35	0.19	**0.23**	0.01	0.27	0.18	0.69	0.25	0.01
	Speech + Kinect	0.20	0.17	0.16	0.16	0.23	0.08	0.07	0.82	0.34	0.31
	Face HoC	**0.45**	0.39	0.09	0.09	0.33	0.39	**0.47**	0.16	**0.49**	**0.69**
	Emotion HoC	0.21	0.14	0.49	0.20	0.06	**0.65**	0.26	0.01	0.13	0.03
	Speech + Face HoC	0.39	0.01	**0.52**	0.05	0.25	0.05	0.13	0.03	0.27	0.03
	Speech + Emo HoC	0.36	0.04	0.47	0.15	0.03	0.01	0.07	0.03	0.32	0.02
	All	0.15	0.18	0.18	0.09	0.29	0.08	0.06	0.79	0.34	0.36
Inter-rater agreement, $\rho_{R_1R_2}$		0.24	0.33	0.48	0.11	**0.60**	0.40	0.15	**0.88**	0.02	0.39
Human	Rater 1	0.70	0.76	0.86	0.79	0.89	0.82	0.70	0.94	0.69	0.81
	Rater 2	0.80	0.83	0.83	0.61	0.86	0.83	0.73	0.97	0.63	0.82

Let us first look at the human-rater correlations. Consider the the human inter-rater agreement $\rho_{R_1R_2}$ which is the correlation between the ratings of the first and second human raters, who need not be experts in the field. We observe only two cases where $\rho_{R_1R_2}$ is greater than 0.5, and these are also the only cases where $\rho_{R_1R_2}$ outperforms machine correlations. This exemplifies the inherent subjectivity and difficulty involved in scoring various aspects of presentation proficiency – non-expert human raters tend to disagree when it comes to rating aspects of performance such as word choice, persuation and audience adaptability, for example, which are higher-level constucts that are not easily defined. However, the correlations between each of these raters' scores and the *final* rating for each scoring dimension are much higher, and close to 1 in many cases. Recall that the final score was adjudicated by an human expert with substantial prior experience in the field after considering the ratings of raters 1 and 2, and that this score need not be a simple average of those scores[3]. This suggests that although non-experts are able to score some aspects of presentation proficiency in line with how an expert would, a substantial amount of expertise is required for the scoring task. In other words, this is a non-trivial problem for not only machines to solve, but naive humans as well.

Let us now focus on the last four scoring dimensions – for instance, the 8^{th} score, representing skillful use of visual aids, is predicted with correlations coming close to the human inter-rater agreement correlation $\rho_{R_1R_2}$. Kinect HoC features and SpeechRater features are particularly useful in this regard, and a combination of these features provides the best correlation of 0.82. This suggests that features that capture temporal information about body movement and speech are very useful in predicting how well subjects use visual aids in presentations, which makes intuitive sense. Further, we see that the 7^{th}, 9^{th} and 10^{th} score dimensions, representing audience-adaptation, persuasiveness, and overall holistic performance respectively, are predicted well by face and gaze (time-series) features in particular. This, again, stands to reason – maintaining an appropriate head posture and direction of one's gaze are important for communicating effectively and persuasively to a given audience. Note also that in these cases the machine correlations are higher than the human agreement correlation $\rho_{R_1R_2}$. Notice that in some of these cases feature combinations may perform worse than the standalone features themselves – this could be due to the relatively larger dimension of the feature fusion (this is especially true in the case of the HoC features, which are sparse and have of the order of 10000 dimensions).

We see that time-series features computed over the intensities of different emotional states (estimated from facial expressions) perform much better than other features in predicting the 6^{th} set of scores (non-verbal behavior). This reconfirms earlier findings in the literature that emotional state information is an important predictor of non-verbal aspects of performance (for example, [10, 9]). While speech and face/gaze features are useful features in predicting the 5^{th} scoring dimension (vocal expression), these correlations are not as high as the human agreement correlation $\rho_{R_1R_2}$.

As far as the other four scoring dimensions are concerned, although our features perform much better than the baseline in all cases, these may not be readily interpretable – since these scores capture higher-level meta-characteristics of the presentation such as quality of introduction, conclusion, organization skill and word choice – and so it may not be clear why these scores perform well at the present time. Indeed, that we observe that speech and face/gaze features perform well on scoring dimensions #1 and #3 while combinations of our Kinect features perform well on scoring dimensions #2 and #4 might suggest that these features capture important behavioral aspects of these meta-characteristics, but understanding and interpreting the reason why is out of the scope of the current paper. Future work will focus on more tailored features in order to predict these scores in an interpretable

[3]Nonetheless, notice that the final adjudicated score is some nonlinear functional of the individual rater scores would result in a bump in correlation values.

manner. For example, these could be features that specifically look at the semantic/syntactic content of the speech for word choice, or that look at the beginning and ending portions of the time-series so as to focus on the introduction and conclusion.

5. CONCLUSIONS

We have presented a comparative analysis of three different feature sets – time-aggregated Kinect features, time-series (or histograms of cooccurrence) Kinect features and SpeechRater features (this combines information from both across and within time-series) – in predicting different human-rated scores of presentation proficiency. We found that certain scoring dimensions were better predicted by speech features, some on Kinect features, and others on combinations of all features. We further observed that these features allowed us to achieve prediction performance near human inter-rater agreement for a subset of these scores. 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 (indeed, we have only experimented with support vector regression here), these experiments provide us significant insight into understanding how to design better techniques for automated assessment and scoring of public speaking and presentation proficiency.

6. ACKNOWLEDGEMENTS

The authors would like to thank Christopher Kitchen, Jilliam Joe, and Chong Min Lee for their help in developing, organizing and supervising the data collection and rating process as well as the processing of speech data.

7. REFERENCES

[1] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[2] 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 Interaction*, pages 200–203. ACM, 2014.

[3] L. Chen, J. Tetreault, and X. Xi. 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*, Los Angeles, CA, 2010. Association for Computational Linguistics.

[4] L. Chen and S.-Y. Yoon. Application of structural events detected on ASR outputs for automated speaking assessment. In *Proceedings of Interspeech*, 2012.

[5] L. Chen and K. Zechner. Applying rhythm features to automatically assess non-native speech. In *Proceedings of Interspeech*, 2011.

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

[7] D. Higgins, X. Xi, K. Zechner, and D. M. Williamson. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25(2):282–306, 2011.

[8] J. H. Jeon and S.-Y. Yoon. Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In *Proceedings of Interspeech*, pages 1275–1278, 2012.

[9] A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.

[10] 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.

[11] L. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia*, 16(4):1018–1031, 2014.

[12] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.

[13] V. Ramanarayanan, M. Van Segbroeck, and S. Narayanan. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer Speech and Language*, in press.

[14] R. Ranganath, D. Jurafsky, and D. A. McFarland. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, 27(1):89–115, 2013.

[15] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez. Inferring mood in ubiquitous conversational video. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, page 22. ACM, 2013.

[16] L. M. Schreiber, G. D. Paul, and L. R. Shibley. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233, 2012.

[17] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski. The interspeech 2011 speaker state challenge. In *Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 3201–3204, 2011.

[18] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, et al. The interspeech 2012 speaker trait challenge. In *INTERSPEECH*, 2012.

[19] H. Van hamme. HAC-models: a novel approach to continuous speech recognition. In *Interspeech*, 2008.

[20] M. Van Segbroeck and H. Van hamme. Unsupervised learning of time–frequency patches as a noise-robust representation of speech. *Speech Communication*, 51(11):1124–1138, 2009.

[21] S. M. Witt. *Use of Speech Recognition in Computer-assisted Language Learning*. PhD thesis, University of Cambridge, 1999.

[22] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009.

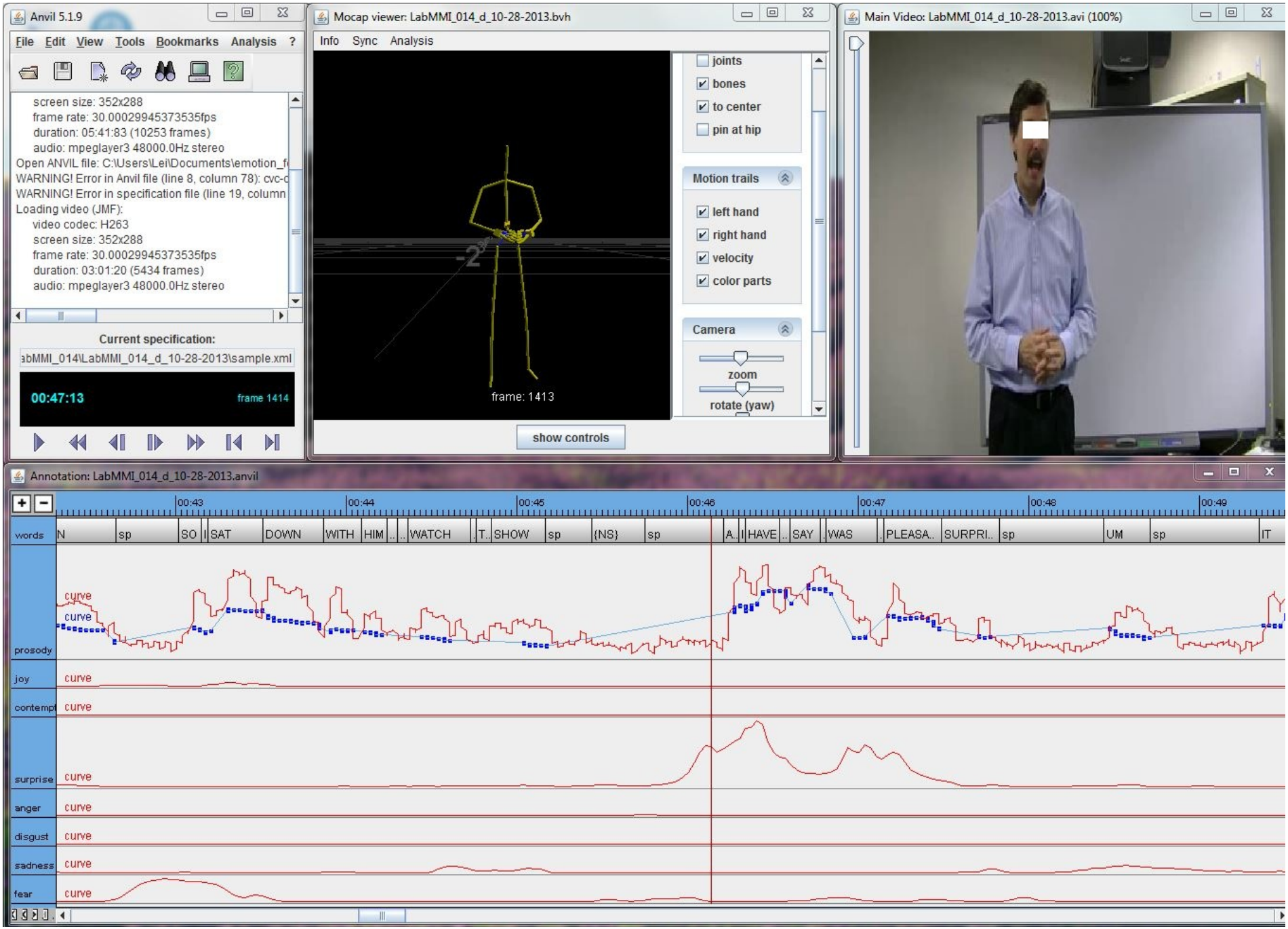

Figure 2: Example figure showing all data streams.