An analysis of time-aggregated and time-series features for scoring different aspects of multimodal presentation data

Vikram Ramanarayanan[†], Lei Chen[‡], Chee Wee Leong[‡], Gary Feng[‡] and David Suendermann-Oeft[†]

Educational Testing Service R&D
[†]90 New Montgomery St, #1500, San Francisco, CA
[‡]660 Rosedale Road, Princeton, NJ
<vramanarayanan,lchen,cleong,gfeng,suendermann-oeft>@ets.org

Abstract

We present a technique for automated assessment of public speaking and presentation proficiency based on the analysis of concurrently recorded speech and motion capture data. With respect to Kinect motion capture data, we examine both time-aggregated as well as time-series based features. While the former is based on statistical functionals of body-part position and/or velocity computed over the entire series, the latter feature set, dubbed histograms of cooccurrences, captures how often different broad postural configurations co-occur within different time lags of each other over the evolution of the multimodal time series. We examine the relative utility of these features, along with curated features derived from the speech stream, in predicting human-rated scores of different aspects of public speaking and presentation proficiency. We further show that these features outperform the human inter-rater agreement baseline for a subset of the analyzed aspects.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics, multimodal computing

1. Introduction

In recent years there has been an increasing demand for assessments of interpersonal interaction and communication skills in general for purposes of business, teacher licensure, etc. Such assessments include problems of automated interviewing, interview assessment, and automated presentation scoring, among others, and although in this paper we choose to focus on the latter problem, the methods we describe can in general be applied to other applications as well.

Multimodal data capture techniques based on motion capture, video and audio feeds provide a rich source of information for such assessment, but the complexity of this data stream 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. [1] 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. [2] used non-verbal behavioral cues extracted from a dataset of 62 interview videos as key components of a computational framework to predict the hiring decision on those videos. Indeed, 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 [3, 4, 5, 6]), this problem has been also been highlighted as a particularly important one, evidenced by a number of challenges at international conferences on related topics [7, 8]. Recently Chen et al. [9] also presented a framework for collecting multimodal data of people giving presentations for purposes of automated assessment. They further presented preliminary results suggesting that basic features in the speech content, speech delivery, and hand, body, and head movements significantly predicted holistic human ratings of public speaking skills. In this work, 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 and Kinect data.

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. It is this gap that we attempt to bridge in this study. Specifically, we propose a feature based on histograms of cooccurrences [10, 11, 12] that models how different "template" body postures 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 body posture in different presentation contexts. 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

Table 1: *Performance standards adapted from the Public Speaking Competence Rubric (PSCR) [13] 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

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.

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 we do not use the video data for this study*. 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 Educational Testing Service (ETS), 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. 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.

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) [13] 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 [13]. 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 agreed with each other, the score was used as the final score. Otherwise, a third rater (expert) was brought in to make another judgment, and the final score assigned was the average of all three scores.

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. We extracted a feature set consisting of the following statistical functionals – means and standard deviations of the left and right hip markers, left and right hand markers as well as their log-transformed values.

3.2. Computing histograms of co-occurrences (HoC) features from Kinect time-series data

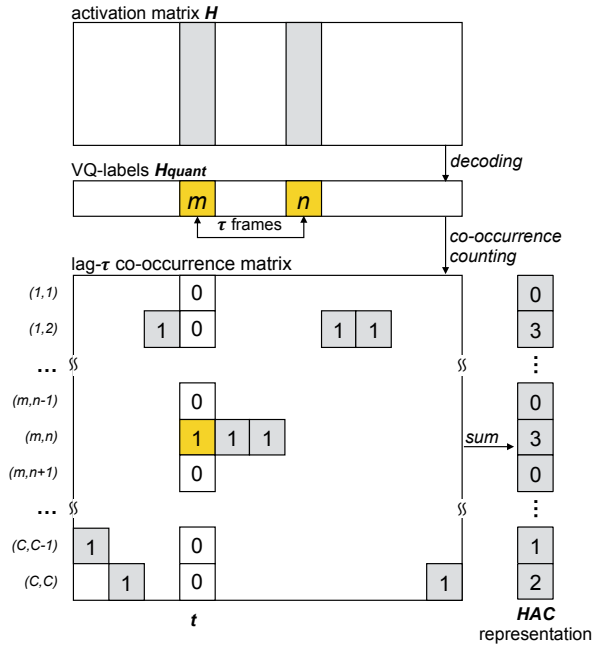

Figure 1: Schematic depiction of the computation of histograms of cooccurrences (HoC) (adapted from [12]). For a chosen lag value, τ, and a time step t, if we find labels m and n occurring τ time steps apart (marked in gold), we mark the entry of the lag-τ cooccurrence matrix corresponding to row (m, 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.

The idea behind the histogram of cooccurrence (HoC) fea-

Table 2: *Speaking Proficiency Features Extracted by 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 [14] 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) [14] as well as features based on the distribution of vowel, consonant, and syllable durations (overall percentages, standard deviation, and Pairwise Variability Index) [15].
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 [16].
	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 [17].
	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 [16].
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 [18],[19].
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 [20].

ture 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 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 Kinect 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 $t+\tau$ [10, 11, 12]. By stacking all (m, n) combinations, each interval can be represented by a 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 1 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 [21], we used SpeechRater, an ETS-internal 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) [22] or its derivatives. For more details on the SpeechRater features, please see Table 2. While prosody and grammar features are directly applicable to assessment of presenters' speech patterns, the rationale for including pronunciation and audio quality features in the feature set was to capture aspects of speakers' pronunciation and intelligibility, which could be useful for score prediction.

3.4. Regression experiments

We used linear support vector machines (SVM) to perform regression experiments [23] 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.

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 Intro	2 Org	3 Conc	4 WC	5 VE	6 NVB	7 AudAdap	8 VisAid	9 Persuasion	10 Holistic
Machine	Kinect HoC	**0.35**	0.03	0.29	0.11	0.11	0.27	0.05	0.62	0.42	0.06
	Kinect Aggregated	0.12	**0.53**	0.09	0.08	0.16	0.26	0.31	0.03	0.11	0.12
	SpeechRater	0.28	0.34	0.03	0.12	0.37	0.22	**0.30**	0.75	**0.48**	**0.44**
	Kinect Both	0.27	0.22	0.36	0.12	0.10	0.16	0.07	0.61	0.37	0.08
	All	0.21	0.14	0.20	**0.16**	0.25	0.12	0.14	0.79	0.17	0.11
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

4. Observations and results

Table 3 lists the performance of various feature sets in predicting different human-rated scores of the multimodal presentation. We compare the performance of time-aggregated Kinect features, time-series HoC Kinect features and SpeechRater 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 of sorts on the prediction performance.

Let us first 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 all features provides the best correlation of 0.79. This suggests that features that capture temporal information about body movement 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} score dimensions, respresenting audience-adaptation, persuasiveness, and overall holistic performance respectively, are predicted well by speech features in particular, although body language captured by both time-series Kinect features (for score #9) and time-aggregated features (for score #7) perform well also. Note though 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 standalong 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 Kinect features (both HoC and time-aggregated) perform well in predicting the 6^{th} set of scores (non-verbal behavior), while speech features are likewise useful for the 5^{th} dimension (vocal expression). Even though these correlations are not as high as the human agreement correlation $\rho_{R_1R_2}$, their higher correlation values relative to other features in each case agrees with our intuitive understanding that non-verbal behavior can be better captured by looking at temporal and time-aggregated statistics of body posture data while vocal expression can be captured by appropriate speech features.

As far as the other four scoring dimensions are concerned, although our features perform much better than the baseline in three out of four 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 combinations of our Kinect features perform well on scoring dimensions #1, #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 interpreting the relevance of different features to predicting the various aspects of the contruct as well as more tailored features (such as features that specifically look at the beginning and ending portions of the time-series so as to focus on the introduction and conclusion) in order to predict these scores in an interpretable manner.

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. We also thank Keelan Evanini for help with the SpeechRater feature writeup.

7. References

[1] 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."

[2] 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*, vol. 16, no. 4, pp. 1018–1031, 2014.

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

[4] 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*. ACM, 2008, pp. 53–60.

[5] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech & Language*, vol. 27, no. 1, pp. 89–115, 2013.

[6] 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*. ACM, 2013, p. 22.

[7] 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*, 2011, pp. 3201–3204.

[8] 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.

[9] 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*. ACM, 2014, pp. 200–203.

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

[11] M. Van Segbroeck and H. Van hamme, "Unsupervised learning of time–frequency patches as a noise-robust representation of speech," *Speech Communication*, vol. 51, no. 11, pp. 1124–1138, 2009.

[12] 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.

[13] L. M. Schreiber, G. D. Paul, and L. R. Shibley, "The development and test of the public speaking competence rubric," *Communication Education*, vol. 61, no. 3, pp. 205–233, 2012.

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

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

[16] 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.

[17] 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*, vol. 25, no. 2, pp. 282–306, 2011.

[18] 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: Association for Computational Linguistics, 2010.

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

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

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

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

[23] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.