Automatic Estimation of Presentation Skills Using Speech, Slides and Gestures

Conference Paper · August 2017

DOI: 10.1007/978-3-319-66429-3_17

4 authors, including:

Abualsoud Hanani
Birzeit University
18 PUBLICATIONS **292** CITATIONS
SEE PROFILE

Mohammad Al-Amleh
Technische Universität München
1 PUBLICATION **3** CITATIONS
SEE PROFILE

Automatic Estimation of Presentation Skills Using Speech, Slides and Gestures

Abualsoud Hanani(✉), Mohammad Al-Amleh, Waseem Bazbus, and Saleem Salameh

Birzeit University, Birzeit, Palestine
ahanani@birzeit.edu

Abstract. This paper proposes an automatic system which uses multimodal techniques for automatically estimating oral presentation skills. It is based on a set of features from three sources; audio, gesture and power-point slides. Machine learning techniques are used to classify each presentation into two classes (high vs. low) and into three classes; low, average, and high-quality presentation. Around 448 Multimodal recordings of the MLA'14 dataset were used for training and evaluating three different 2-class and 3-class classifiers. Classifiers were evaluated for each feature type independently and for all features combined together. The best accuracy of the 2-class systems is 90.1% achieved by SVM trained on audio features and 75% for 3-class systems achieved by random forest trained on slides features. Combining three feature types into one vector improves all systems accuracy by around 5%.

Keywords: Presentation skills · Audio features · Gesture · Slides features · Multi-Modality

1 Introduction

Performing a good presentation in front of a crowd is an essential skill that every successful and professional person should master. This is one of the student outcomes most undergraduate programs aim to develop in their study journey and after that in their work life. Throughout courses, people obtain such skill and nourish it. But what remains an issue is judging how well a person is performing or how better he/she has become since last time. Watching presentations is both time consuming and harder than it seems for evaluators to judge and provide feedback. Without feedback none can get better. In most cases, presenters do not receive objective feedback after their presentations since this requires tremendous amount of effort and time from the evaluators. Usually, presentation performance assessment is done by focusing on multi-modality, speech cues, gesture and slides. Speech cues include way of speaking, volume, intonation, speaking rate, etc. whereas, gesture cues include facial expressions, eye contact, head poses, hand gesture and body posture.

Most of the current rubrics for presentation performance assessment rely on both verbal and non-verbal aspects, and it is mainly done by humans. Doing this process automatically and providing instants feedback is highly desirable.

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 182–191, 2017.
DOI: 10.1007/978-3-319-66429-3_17

In this work, we are proposing an automatic system which uses multi-modality, namely speech, gesture and slides content and formatting, for presentation performance assessment. Most of the previous studies in this field used one or two modalities, and to our knowledge, this is the first work which combines cues from three modalities for assessing presentation performance.

2 Previous Work

In [1] prosodic audio features and personality assessment provided by humans were used (each alone and combined) to classify speakers as professional and non-professional (2-classes). The audio features they used are: pitch, energy, first and second formants, length of voiced and unvoiced segments and their respective statistics (minimum, maximum, mean and entropy of feature variation). In personality assessment by humans, the score for each audio clip is the average of 10 judges' assessment (the BFI-10 questionnaire). They obtained an accuracy of 87.2%, 75.5% and 90.0% when they used prosodic features, personality assessments and when prosodic features and personality assessments were combined together, respectively.

The liveliness of a voice is defined as the degree to which a voice varies in intonation, rhythm and loudness [2]. In [2], the Pitch Dynamism Quotient (PDQ) was used to analyze the liveliness of speech and it was hypothesized that monotonous speech has PDQ values around 0.10 and lively speech has PDQ values around 0.25. In [3], "high-dimensional acoustic feature extractions" approach was employed to develop a system to assess the oral presentations skills of pre-service principals. Their approach incorporates multimodal behavioral data (audio and video) to classify pre-service principal's presentations into low and high presentations.

In [4], The level of the cognitive load that a person is experiencing was classified to low, moderate and high cognitive load based on some speech features, namely, articulation rate, pause rate and pause duration.

In [5], features extracted form audio and slides were used to classify students' performance in presentations into two classes, high and low. From audio, they used some prosodic features namely, Minimum Pitch value (MINP), Maximum pitch value (MAXP), Average Pitch value (AVGP) and Pitch Standard Deviation value (STDP). They used also the speech rate, articulation rate and The Average Syllable duration. For slides, they used the total number of images, minimum and maximum font size, maximum number of different font sizes per slide, total number of words, total number of chart and total number of tables. Also, they processed each slide as a gray JPGE image to calculate its entropy and they computed the following features: maximum entropy value, minimum entropy value, average of entropy values and standard deviation of entropy values. It is worth mentioning that we will use the dataset that they used in their work and to build on what they have done.

Many researchers tried to automate the way in which presentation slides are assessed. They have chosen numbers as their reference starting with simple count of images or tables inside a presentation slide and ending with whether a footer exists at

the bottom of the slide or not. According to Seongchan Kim et al. in [6], with the huge increase in PowerPoint slides count and their hosting sites, an efficient way to estimate the quality of slides without any human intervention is required [6]. They extracted a set of useful features from slides.

Gestures and body movements can alter how people conceptualize abstract concepts [7] and even their sense of their own dominance [8]. Despite the fact that gestures are a substantial aspect in a presentation, studies do not seem to give it much attention. However, there is some research done that uses gestures to predict the emotions of the speakers. This is helpful for our work in terms of how to capture and use gestures information for presentation assessment.

Burgoon et al. [9] proposed an approach for analyzing cues from multiple body regions for the automated identification of emotions displayed in videos, focusing on hands and arms movement, facial pleasantness and head movement. This work does not have clear results. All what they conclude is that "this research has already shown great promise and is setting the stage for real-world relevance". In addition, S. Kopf et al. in [10] developed a software tool using Microsoft's Kinect and captured gestures, eye-contact, movement, speech, and the speed of slide changes to provide real time feedback for presentation skills. Speaker movement and body gestures were detected well while not all spoken words and slide changes could be recognized.

3 Dataset

3.1 MLA'14 Data Set

In all of reported experiments in this paper, we have used a dataset that was collected by international Multimodal Learning Analytics workshop and challenges (MLA 2014) which seeks to answer questions like how multimodal techniques can help the assessment of presentation skills? And how to integrate between individual performance (audio, video and posture) and the quality of the slides used, in determining how good a presentation is.

This dataset is composed of 448 multimodal recordings on 86 oral presentations of undergraduate students' groups. Each group consists of a varied number of students (1–6). It is important to note that each PowerPoint file is shared for each students' group, where, audio and gestures are recorded for each student individually.

Human coded information about the quality of the presentation was included, six aspects was taken into consideration in determining the quality of presentation. The coding process was done by four individuals, taking the average as the final rate for each criteria. The human coding was recorded with a rubric that measured: speech organization, volume and voice quality, use of language, slides presentation quality, body language and level of confidence during the presentation. Table 1 shows all evaluation criteria used to assess the quality of the presentation. The score goes from 1 (low) to 4 (high). The students of each group were evaluated individually using these metrics. The evaluation of the metrics related to the slides was the same for all group members.

Table 1. Evaluation criteria used for scoring the student oral presentations

Metric	Description
Speech organization	Structure and connection of ideas
Volume/voice	Presents relevant information with good pronunciation
	Maintains an adequate voice volume for the audience
Language	Language used in presentation according to audience
Slides presentation	Grammar
	Readability
	Impact of the visual design of the presentation
Body language	Posture and Body language
	Eye contact
Confidence during the presentation	Self-confidence and enthusiasm

As shown in Table 1, there are three classification aspects human experts used in the presentations evaluation; one is related to the voice of the presenter, one is related to the slides of the presentation and one is related to the gestures and body language of the presenter. In order to build an automatic system for presentation evaluation, this system should consider features extracted from these three aspects; voice, slides and gestures. To do so, we built three sub-systems, each uses one of these features, i.e. one system uses features extracted from voice, one uses features extracted from slides and one uses features extracted from gestures.

3.2 Two-Class Labeling

For building voice based system, the dataset was divided into two classes; high performance (average voice-related scores > 2.5) and low performance (average voice-related scores $\leq$ 2.5). By applying this criteria, 331 audio files were labeled as high performance and 117 were labeled as low performance. Similarly, the dataset was re-divided into two classes (High and Low), but this time, according to the average rate of the slides-related evaluation criteria, and one time according to the gestures based scores for the gesture based system. By Appling this criteria, 45 PowerPoint files were labeled as class 'High' and 41 files as class 'Low'. Similarly, 231 Kinect csv files were labeled as class 'High' and 217 as class 'Low'.

3.3 Three-Class Labeling

To have more details of the presentation quality, all students in the dataset were re-divided into three classes; High (rating range 1–2), Average (rating range 2–3) and Low (rating greater than 3) for our three sub-systems. Table 2 shows the number of data files of each class after applying the above criteria, for each sub-system.

Table 2. Three-class data division

Model	Class	No. of instances
Audio	High	195
	Average	191
	Low	62
Slides	High	32
	Average	26
	Low	28
Gesture	High	102
	Average	246
	Low	100

4 Feature Extraction

4.1 Audio Features

To build an automatic assessment system based on voice, audio recordings were used to extract representative features from speech signal. The following subsections describe the audio features used in our experiments.

- *Short frame energy:* Speech signal is divided into 50% overlapped 20 ms frames. Each frame is multiplied by Hamming window and then the energy, in decibel, is calculated for each frame and used as an audio feature for our system.
- *Short frame Zero-Crossing Rate (ZCR):* After subtracting average (dc) of speech signal from each sample, the number of zero-axis crossings is calculated for each short frame. These counts are then divided by the total number of zero-crossings of the whole utterance.
- *Mel Frequency Cepstral Coefficients (MFCCs):* MFCC features are the most commonly used in the speech processing applications. They represent the general shape of power spectrum for each frame with low dimensional feature vectors (typically 12). More details about MFCC technique can be found in [11]. The first 12 MFCCs of each frame are appended to the audio feature vectors of our audio-based system.
- *Short frame pitch:* Pitch refers to the fundamental frequency of the voiced speech. Pitch is an important feature that contains speaker-specific information. It is a property of vocal folds in the larynx and is independent of vocal tract. A single pitch value is determined from every windowed frame of speech. There is a number of algorithms for estimating pitch form speech signal. Among these, one of the most popular algorithms is the Robust Algorithm for Pitch Tracking (RAPT) proposed by Talkin [12]. This algorithm was used to extract pitch for use in all experiments reported in this paper.
- *Formant Frequencies:* The general shape of the vocal tract is characterized by the first few formant frequencies. *Praat* toolkit [13] was used to estimate the first three formants and their gains and appended to the acoustic feature vectors.

- *Speaking rate:* Speaking rate has been used as a feature in numerous speech processing applications. In this work, speaking rate was estimated from the number of syllables divided by the total duration in seconds of each participant presentation.
- *Articulation Rate:* The number of syllables divided by the speaking time.
- *The Average Syllable duration:* The ratio of the speaking time over the number of syllables. The number of syllables in each audio recording is found by counting the detected syllables nuclei using Praat script by Nivja de Jong [14].
- *Pauses:* There are two types of pauses; presence of silent intervals (empty pauses) and vocalizations (filled pauses) which do not have a lexical meaning. Usually, non-confident presenters need time for selecting proper words and making meaningful sentences while speaking. These times are longer than the natural pauses confident presenters usually make while they are speaking. Therefore, the length and number of occurrences of pauses may carry a useful information about the presenter skills. A simple algorithm based on the short frame energy and zero-crossing rates has been developed for estimating length and number of occurrences of pauses in each utterance. Frames with low energy and high zero-crossing are usually resemble pauses, whereas, frames with high energy and relatively low zero crossings resemble speech frames. If a number of successive pause frames exceeds a practically specified threshold, they are considered as a pause. So, if it exceeds certain duration time or if it is repeated many times while talking, this may indicate that the presenter has a low presentation skills.
- *Rhythm Patterns (RP), Statistical Spectrum Descriptor (SSD) and Rhythm Histogram (RH):* Rhythm Patterns are features sets derived from content-based analysis of audio, particularly music, and reflect the rhythmical structure in the audio recording. According to the occurrence of beats or other rhythmic variation of energy on a specific critical band, statistical measures (e.g. mean, median, variance, skewness, kurtosis, min- and max-value) are able to describe the audio content. The Rhythm Histogram features are a descriptor for general Rhythmic in an audio segment. Contrary to the RP and SSD, information is not stored per critical band. Rather, the magnitudes of each modulation frequency bin of all critical bands are summed up, to form a histogram of "rhythmic energy" per modulation frequency. 1440 RP features, 1500 RH and 168 SSD are computed using open-source Musical Information Retrieval toolkit[1].

4.2 Slides Features

Each students group has one PowerPoint presentation file. Therefore, unlike audio and gesture features, features extracted from slides are the same for each group member.

A macro was created to automatically compute a set of features from each slide of the presentation files to be used for presentation assessment, as shown in Table 3.

[1] http://www.ifs.tuwien.ac.at/mir/downloads.html.

Table 3. Set of features extracted from slides

Slide features	
Words count per slide	Unique font mean per slide
Total number of images per slide	Entropy of a slide
Font sizes per slide	Delta Entropy of two consecutive slides
Unique font sizes per slide	Word to image ratio per slide
Minimum font size per slide	Mean font per slide
Maximum font size per slide	Font difference per slide

4.3 Gesture Features

In the dataset, each student has a Kinect recording (csv format) which includes XYZ coordinates of 20 joint body positions in a rate of 120 frames per second. We extracted a set of features from Kinect motion traces for each presenter, as shown in Table 4 below.

Table 4. Gesture features extracted from Kinect csv files

Gesture feature	
Position of the 20 points of the skeleton	Contraction index
Position of the 7 points related to the head, shoulders and arms	Energy and power
Speed of movement	Overall activity
Acceleration of the movement	Shape of movement
Fluency and smoothness	

5 Experiments and Results

5.1 Experiments Setup

In all reported experiments in this paper, 10-fold cross validation technique was used for training and validation of each system. In order to investigate the usefulness of each feature type (audio, slides and gesture) for estimating presentation skills, we conducted one experiment for each feature type alone and then combined all together. As mentioned earlier, presentation skill of each participant in the dataset was classified into two classes (high vs. low), one time, and into three classes (high, average and low) another time. This classification was done based on the human ratings. Therefore, for each feature type, there are two classification experiments, one with two classes and one with three classes. In all experiments, three different classifiers implemented in Weka toolkit were used, namely, Support Vector Machines (SVM), Simple Logistic (SL) and Random Forest (RF).

The above mentioned audio features are computed for each audio file and concatenated together to form feature vectors of a dimension 3140 (energy, ZCR, 12 MFCCs, pitch, 6 formant frequencies with their gains, speaking rate, articulation rate, average Syllable duration, average pauses length, number of pauses, 1440 Rhythm Patterns, 168 Statistical Spectrum Descriptor and 1500 Rhythm Histogram features).

5.2 Results

For each experiment, accuracy, precision, recall, F-measure and ROC area were used as evaluation measures. The results of the three classifiers when trained and tested on audio features, slides features and gesture features are presented in Table 5 (two-class) and Table 6 (three-class).

Table 5. Two-class (high vs low) experiments results of the three systems

System	Classifier	Accuracy	Precision	Recall	F-Measure	ROC Area
Audio	SVM	**0.901**	**0.901**	**0.901**	**0.903**	**0.909**
	Simple logistic	0.869	0.872	0.869	0.87	0.907
	Random forest	0.809	0.804	0.809	0.802	0.894
Slides	SVM	0.701	0.69	0.701	0.69	0.646
	Simple logistic	0.694	0.68	0.694	0.677	0.739
	Random forest	**0.83**	**0.829**	**0.83**	**0.829**	**0.908**
Gesture	SVM	0.635	**0.694**	0.305	0.421	0.602
	Simple logistic	**0.714**	0.689	**0.632**	**0.618**	**0.704**
	Random forest	0.668	0.612	0.598	0.586	0.653
Combined	SVM	**0.953**	**0.952**	**0.951**	**0.952**	**0.919**
	Simple logistic	0.912	0.906	0.911	0.911	0.914
	Random forest	0.871	0.864	0.866	0.861	0.903

Table 6. Three-class (high, average, low) experiments results of the three systems

System	Classifier	Accuracy	Precision	Recall	F-Measure	ROC Area
Audio	SVM	**0.675**	**0.667**	**0.675**	**0.67**	**0.784**
	Simple logistic	0.559	0.56	0.559	0.56	0.692
	Random forest	0.492	0.506	0.492	0.495	0.653
Slides	SVM	0.465	0.457	0.465	0.449	0.587
	Simple logistic	0.475	0.468	0.475	0.467	0.647
	Random forest	**0.753**	**0.752**	**0.753**	**0.752**	**0.902**
Gesture	SVM	0.383	0.352	0.346	0.343	0.497
	Simple logistic	**0.433**	**0.428**	**0.429**	**0.432**	**0.551**
	Random forest	0.381	0.371	0.368	0.364	0.489
Combined	SVM	0.722	0.714	0.721	0.693	0.825
	Simple logistic	0.653	0.635	0.646	0.623	0.782
	Random forest	**0.812**	**0.817**	**0.812**	**0.813**	**0.911**

As expected, treating presentation skills estimation as a two-class problem gives better results than a three-class problem for all systems and all classifiers. All of the systems which use audio features outperform the systems that use slides and gesture features. As shown from the results, SVM system outperforms RF and SL when using audio features with an accuracy of 90.1% and 67.5% for 2-class and 3-class experiments, respectively. Random forest worked the best for the slides features,

with accuracies of 83.0% and 75.3% for 2-class and 3-class experiments, respectively. Simple logistic classifier worked the best for the body language features (gesture) with accuracies of 66.8% and 43.3% for 2-class and 3-class respectively.

Combining the three feature types into one feature vector, then train and evaluate the three classifiers on the resulting vectors, improves all systems accuracy by around 5%. Combining three systems by fusing their output scores is considered in the future work.

6 Conclusion

In this paper, we presented a comprehensive framework for automatically estimating presentation skills by extracting features from presenter voice, slides and body language. MLA'14 dataset was used for training and testing (10-fold cross validation) in all of reported experiments. Presentation skills prediction was treated as a 2-class and 3-class classification problems. In each case, three different classifiers were built on audio features, slides features and gesture features, independently. SVM worked the best for the audio features, whereas, Random forest and simple logistic worked better for the slides features and gesture features respectively.

The best accuracy of the 2-class systems is 90.1% achieved by SVM trained on audio features, and 75% for 3-class systems achieved by Random forest classifier. Combining three feature types into one vector improves all systems accuracy by around 5%.

References

1. Mohammadi, G., Vinciarelli, A.: Humans as feature extractors: combining prosody and personality perception for improved speaking style recognition. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, 9–12 October 2011, pp. 363–366 (2011)
2. Hincks, R..: Processing the prosody of oral presentations. In: InSTIL/ICALL Symposium (2004)
3. Shan-Wen, H., Sun, H.C., Hsieh, M.C., Tsai, M.H., Lin, H.C., Lee, C.C.: A multimodal approach for automatic assessment of school principals' oral presentation during pre-service training program. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
4. Gorovoy, K., Tung, J., Poupart, P.: Automatic speech feature extraction for cognitive load classification. In: Conference of the Canadian Medical and Biological Engineering Society (CMBEC) (2010)
5. Luzardo, G., Guaman B., Chiluiza, K., Castells, J., Ochoa, X.: Estimation of presentations skills based on slides and audio features. In: Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge, MLA 2014, pp. 37–44. ACM, New York (2014)
6. Kim, S., Jung, W., Han, K., Lee, J.-G., Yi, Mun Y.: Quality-based automatic classification for presentation slides. In: Rijke, M., Kenter, T., Vries, A.P., Zhai, C., Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 638–643. Springer, Cham (2014). doi:10.1007/978-3-319-06028-6_69
7. Jamalian, B., Tversky, T.: Gestures alter thinking about time. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci) (2012)

8. Carney, A.J.C., Dana, R., Yap, A.J.: Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance. Psychol. Sci. **21**(10), 1363–1368 (2010)
9. Burgoon, J.K., Jensen, M.L., Meservy, T.O., Kruse, J., Nunamaker, J.F.: Augmenting human identification of emotional states in video. In: Intelligence Analysis Conference, McClean, VA (2005)
10. Kopf, S., Guthier, B., Rietsche, R., Effelsberg, W., Schon, D.: A real-time feedback system for presentation skills. In: MLA 2014, Mannheim, Germany, pp. 1633–1640 (2015)
11. Fang, Z., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. J. Comput. Sci. Technol. **16**(6), 582–589 (2001)
12. Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Kleijn, W., Paliwal, K. (eds.) Speech Coding and Synthesis, pp. 495–518. Elsevier, New York (1995)
13. Paul, B., David, W.: PRAAT: doing phonetics by computer [Computer program]. Version 6.0.26. http://www.praat.org/. Accessed 2 Mar 2017
14. Jong, D., Nivja, H., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behav. Res. Methods **41**(2), 385–390 (2009)