PRESENTATION QUALITY ASSESSMENT USING ACOUSTIC INFORMATION AND HAND MOVEMENTS

Fasih Haider, Loredana Cerrato, Nick Campbell[*]

SCSS, Trinity College Dublin, Ireland
{haiderf,cerratol,nick}@tcd.ie

Saturnino Luz

IPHSI, University of Edinburgh, UK
S.Luz@ed.ac.uk

ABSTRACT

This study focuses on prosodic and gestural features that contribute to the positive judgement of public oral presentations. The general hypothesis is that certain prosodic characteristics, such as high pitch variation and perceived loudness, together with the production of natural hand gestures, influence the audience's perception of the speaker as a good presenter. Being able to identify features that can give an indication of a good presenter is useful for applications in the field of skills training, where automatic feedback could be provided to trainees at the end of their presentation about the extent to which they have been able to use their voices and gestures to keep the audience engaged. For this reason, we also propose a method, based on prosodic and visual features, able to categorise presentation quality with high accuracy.

Index Terms— Multimodal Signal Processing, Prosodic Analysis, Gestures Analysis

1. INTRODUCTION

Prosody is believed to be of fundamental importance in contributing to the success of a public speech. Several manuals on public speaking advise the presenter to speak with a lively voice, where by lively voice is meant a voice that varies in intonation, rhythm and loudness [10, 7]. Liveliness has also been associated with enthusiasm [17] and this is what we also assume in this study. Previous studies have formulated and tested the hypothesis that the higher the variability (or standard deviation) of fundamental frequency (F0), the more a spoken utterance is perceived as lively [20, 8]. However F0 deviation alone might not always be an optimal feature discriminating lively speech from monotonic speech (typical of depressive states) [19].

Another aspect that seems to contribute to the success of public speech is speaking rate. This has shown to be more strongly correlated than pitch variation with perceptions of liveliness [20] and has been considered, together with voice level and intensity, as an indicator of self confidence. Fast rate of speech, lower voice level and high speech intensity are listed among the characteristics of self confident voices in several studies [7, 10].

Other characteristics believed to contribute to the success of a presentation include the speaker's ability to establish contact with their listeners (e.g. eye contact) and be aware of their body language. Specific postures that supposedly denote self confidence, such as standing straight with feet aligned under the shoulders, and both feet flat on the ground, are recommended by public speaking guides. Other postures that denote the lack of self confidence, such as fidgeting, crossing the legs, gesturing widely without purpose are considered inappropriate [3]. However, to the best of our knowledge, these recommendations have not been validated against large sets of presentation data. In this paper we investigate the hypothesis that hand movements produced in the upper part of the body (close to the shoulders) can contribute to good ratings for the speaker's body language in the analysed dataset.

In addition to testing hypotheses that relate presentation quality to prosodic and gestural features, we propose a method for automatic inference of presentation quality using a large array of low-level audio and video descriptors. Automatic detection of presentation quality is a challenging task. Few studies have been conducted in this field. In one study on self confidence detection [9] the authors compared multiple classifiers using a set of prosodic and spectral features, on a very limited dataset consisting of fourteen females speakers giving regular lectures ranked by 5 experts judging self confidence. Their classifiers are able to detect two classes (low self confidence and high self confidence) with a maximum accuracy of 87.7 % and 75.2 % for speaker-dependent and speaker-independent settings respectively. There are other studies conducted on the MLA dataset [14]. Luzardo et al. [12] employ features extracted from presentation slides to predict overall presentation quality (2-class problem), obtaining up to 65% accuracy. When audio features are used, pitch and filled-pause related features improve accuracy to 69%. Chen et al. [2] propose a different approach, performing a clustering of presentation ratings to derive two principal components (roughly corresponding to delivery skills and slide

[*] This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at School of Computer Science and Statistics, Trinity College Dublin and FP7-METALOGUE project under Grant No. 611073.

quality) which they use as the target functions of a regression task. Finally, Echeverria et al. [4] employed machine learning models to classify presentations according to performance (good vs. poor), achieving accuracy scores of 68% and 63%. Our results compare favourably to these other studies.

2. THE DATASET

A sub-set of the presentations contained in a sub-corpus of the Multimodal Learning Analytics (MLA) dataset [14] is used as dataset to run the experiment: 416 oral presentations given by Spanish-speaking students presenting projects about entrepreneurship ideas, literature reviews, research designs, software design etc. The dataset contains: speech, facial expressions and physical movements in video, skeletal data gathered from Kinect[1] for each individual, and slides of presentations, making up a total of 19 hours of multimodal data. In addition, individual ratings for each presentation, and group ratings related to the quality of the slides used when doing each presentation are available. Each presentation has a rating based on the following performance factors: a) structure and connection of ideas, b) presentation of relevant information with good pronunciation, c) maintenance of adequate voice volume for the audience, d) usage of language according to audience, e) grammar of the slides, f) readability of the slides, g) impact of the visual design of the slides, h) posture and body language, i) eye contact, and j) self-confidence and enthusiasm.

3. HYPOTHESES

In the MLA dataset, each student is judged by the audience on a scale ranging between good (4) and poor (1). In this analysis, we assume that a presentation factor (such as self confidence) is considered good if the rating assigned to it is $>= 2.5$, otherwise the presentation factor is considered poor. Based on the assumptions and results found in the literature we formulate the following hypotheses for the investigation of prosodic features [10, 7, 17, 20, 8, 10, 19]:

1. Standard deviation (sd) of F0 is an indication of liveliness and enthusiasm. So a higher value is an indication of a lively and enthusiastic voice.
2. The harmonic-to-noise ratio (HNR) may indicate abnormality in the voice, so that speakers with lack of self confidence tend to exhibit high values of HNR [21].
3. High values of perceived loudness reflect a loud voice, which is considered an indication of a good presenter.
4. A fast speech rate is an indication of a fluent speaker.

Regarding last hypothesis, speech rate is usually measured in number of words spoken per minute. In the available dataset, however, we measured vocalisation to pause ratio as an alternative measure to indicate the fluency of speech. Pauses and vocalisation lengths are known to play an important role in structuring both discourse and interactive speech [15, 11], so we expected this feature to provide a reasonable index of fluency in presentations.

For the analysis of visual features we formulate the following hypothesis: production of hand gestures in the upper part of body is assumed to be an indication of fluent gestures produced by good presenters. This hypothesis is based on our observations of the behaviour of the top ten speakers who obtained good ratings for their body language and the top ten speakers who received poor ratings for their body language. The two groups follow a clear trend: the good speakers produce fluent arm and hand gestures concentrated in the upper part of the body. Their gestures have the following main functions: a) provide discourse with continuity and coherence [13], b) mark stress and rhythm of utterances, c) point at the slides and d) describe something.

The speakers who received the lowest ratings for body gestures seem, in general, to produce fewer gestures in the upper part of the body. They tend to keep their arms down, parallel to their body, or keep their hand in hands at the level of their belly. When they produce gestures they produce particular types of hand movements which are not connected to the co-occurring discourse (neither semantically nor structurally), as described by Ekman [5]. These gestures seem to be produced by the speaker to manage particular emotional states, such as tension or anxiety and are not concentrated in the upper part of the body. Analysing the skeletal data, we attempted to find a measure that could detect the position of the arm gestures in relation to the shoulder centre used as a reference. We chose the mean value of the Euclidean distance between the hands and shoulder as the basic measure. Despite its relative simplicity, this measure provides a good indication of hand and arm movements concentrated in the upper part of the body, and thus can be used in testing our hypothesis.

4. PRESENTATION QUALITY ASSESSMENT

We start by analysing correlations among the different categories of ratings in order to estimate how visual and prosodic features might contribute to the prediction of overall presentation quality.

Figure 1 depicts the correlations of all ratings (correlation matrix) as a corrgram [6], where blue indicates positive correlation and red indicates negative correlation, with darker hues indicating stronger correlations. We note that ratings that appear to be motivated by voice feature (e.g. self-confidence and enthusiasm) are sometimes highly correlated to visually-motivated ratings (e.g. body language and pose) are highly correlated. This might imply that either visual or voice features alone might suffice to distinguish some aspects of presentation quality. Alternatively, it is possible that combined feature sets might be more effective overall.

In order to investigate these issues in more detail, in the

[1] https://www.microsoft.com/en-us/kinectforwindows/

following sections, we analyse how the various prosodic and visual features vary according to two broad performance categories (poor vs. good presentation), in line with the hypotheses formulated in section 3, and then describe a method for automatic categorisation of presentation quality level based on a rich set of such features, presenting its results on the MLA dataset.

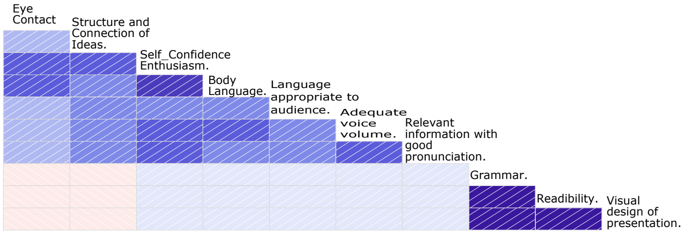

Fig. 1. Correlation matrix for rating categories

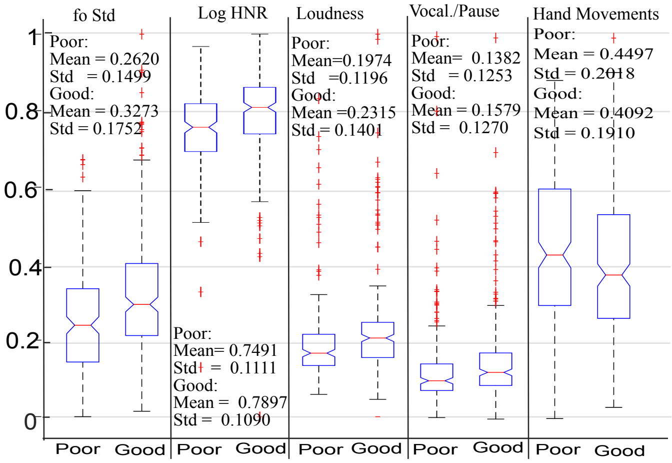

Fig. 2. ANOVA Test Results.

4.1. Feature Extraction

In total we use 6376 audio features for the classification tasks: the complete audio set of the ComParE challenge [16] (6,373 features) with the addition of perceived loudness (sd and mean) and V/P (vocalisation to pause ratio).

To extract the features related to the speakers' hand movements we calculate the Euclidean distance (ED) between wrist joint and shoulder centre joint (tracked by Kinect) in each frame of the video (presentation of student). Finally, we calculate the mean, standard deviation, maximum, minimum, median, maximum ratio and minimum ratio of the ED, its first (velocity) and second (acceleration) order derivative for each video/speaker. In total we extract 42 features for both hands. The maximum ratio for a speaker is measured by counting the number of frames which have higher ED compared to their preceding and following frames and then averaged over the total number of frames in that video. Similarly, the minimum ratio of a speaker is measured by counting the number of frames which have lower ED compared to their preceding and following frames and then averaged over the total number of frames in that video.

4.2. Hypothesis testing

In order to validate the hypotheses formulated for prosodic and gestural features, we performed analysis of variance (ANOVA) on the values of several features with respect to presentations rated as poor, as compared to presentations rated as good. The results show that a significant difference exists between the poor and good groups of speakers for the different measures considered: a significant difference is shown between fundamental frequency standard deviation values ($p = 0$) for good and poor presenters. The box plots shown in Figure 2 depict the quartiles for the respective distributions of values. The higher the values of the standard deviation the higher the pitch variation of the student during the presentation. This is in line with the results of previous studies [20, 8] that show that the higher the standard deviation of fundamental frequency, the more a speech sample is perceived as lively. Since we assume that liveliness is associated with enthusiasm this result is an indication that speakers rated as good presenters are very likely to have lively and enthusiastic voices. The log HNR shown in Figure 2 has higher values for good speakers ($p = 0.0002$) which might be due to the fact that the presenters perceived as good speakers do not present obvious abnormalities in their voice quality (e.g. roughness of sound [18]). The results summarised in Figure 2 show higher values of perceived loudness for speakers judged as good ones ($p = 0.009$). This is because loudness plays an important role in expressing self confidence and enthusiasm and speaking loud is generally considered a characteristic of good presenters, consistently with our hypothesis and the literature on presentation quality.

As for the results of the vocalisation over pause ratio we do not get significant differences ($p = 0.1148$) between speakers judged as good versus poor. This might depend on the fact that pauses can also be used for rhetorical purposes and in our calculation of vocalisation to pause ratio (as explained in section 4.1) we cannot take into account the filled pauses and hesitations since they were not annotated in the audio files. As for visual gestures our hypothesis is that hand gesture produced in the upper part of the body are an important factor that characterises a good presenter. To perform hand gestures in the upper part of the body, speakers need to move their hands in that region for a relatively long period of time. A good presenter can also move his/her hands in the lower body region (to relax) or over head (to point out the slides) but these gestures should not be maintained for long periods of time. So, we decided to choose the mean value of ED as a measure of these gestures. The results show that the good presenters have statistically significantly lower ED values ($p = 0.036$) as shown in figure 2. Although it is true that

Table 1. 2-Class Experiment Results (F measure)

Feature	Rank	Poor	Good
Audio	Conf.	92.64%	94.19 %
Audio	Body	94.32%	94.61%
Visual	Conf.	60.06%	73.35%
Visual	Body	64.18%	66.51%
Fusion	Conf.	94.02%	95.26%
Fusion	Body	95.80%	96.02%

the visual features are, in some sense, 'optimised' (i.e. designed by us rather than discovered automatically from data), they come not only from simply watching the videos, but are also informed by the literature on gestures and presentations [13, 5].

4.3. Classification Method

We first z-score normalise our feature set and then scale it in the range of [0 1]. To reduce the high dimensionality of features we employ PCA (Principle Component Analysis) over the feature set to reduce the number of dimensions to number of instances. After that we map our data to the reduced dimensions. From the statistical significance (***p***) of the transformed feature set with the rating (poor or good), we select the transformed features with $p < 0.5$. The classification method was implemented in MATLAB [2] and employed discriminant analysis in 10-fold cross validation experiments.

4.4. Results

Prosodic and visual features are analysed to predict presentation quality. The correlation test results (figure 1) show that the presentation quality factors under consideration are highly correlated with each other. Therefore, in principle, it should be possible to detect the body language rating ('Body') with prosodic features, and the self confidence rating ('Conf.') with skeletal features.

The motivation for this automatic inference task is to be able to distinguish those students who present really poorly (and therefore might need expert attention and extra tutoring), from those who present really good and might be selected as examples of how to present from the average presentations. The very good speakers do not necessarily need extra attention from the tutor, while the poor presenters might benefit from advice. Therefore, we performed two experiments. In the first (2-Class) experiment we have 3 types of feature vectors, 2 types of ratings and 2 types of groups of speakers (poor and good). The results are shown in table 1. In the second experiment (3-Class) we have the same settings, but the students are divided into three groups: poor (rating range is $1-2$), average (rating range is $2-3$) and good (rating $>= 3$). The results (harmonic mean) are shown in table 2.

Table 2. 3-Class Experiment Results (F measure)

Feature	Rank	Poor	Average	Good
Audio	Conf.	83.76%	84.54%	85.42%
Audio	Body	84.32%	80.85%	84.24%
Visual	Conf.	60.00%	18.46%	66.19%
Visual	Body	62.35%	07.02%	56.98%
Fusion	Conf.	82.49%	82.00%	83.07%
Fusion	Body	84.62%	76.60%	81.03%

5. DISCUSSION AND CONCLUSION

Our study uses an extended dataset including both male and female students, in contrast to the limited dataset used in a similar study [9]. Our approach is tested in speaker-independent settings and the student presentations are ranked by an audience. It yields maximum F scores of 95.26% (good) and 94.02% (poor) in detecting self confidence. Moreover, in the two-class problem, the F measure of prosodic features indicates that the prosodic features are not only able to predict the rating of *self confidence and enthusiasm* but also the rating of *body language and pose*. This may be due to the impact of good posture on speaking style. The visual features show the same behaviour, but with less accuracy. However, the fusion of prosody and visual features does in fact improve overall categorisation performance.

We also obtained promising results for three-class rating detection, with F scores as high as 83.76% (poor) 84.54% (average) and 85.42% (good) in detection of *self confidence and enthusiasm*. In the three-class problem, the features show the same behaviour as for the two-class problem, except for the fusion which causes a slight decrease in performance. At the same time they cause a slight increase in poor (body language rating) class detection, while visual features alone are almost unable to detect average class (07.02 % and 18.46%). In the three-class problem, the features show the same behaviour as for the two-class problem, except that feature fusion does not seem to improve performance in this case.

Being able to automatically analyse non-verbal components to predict public speaking performance is a significant contribution to this field. In this paper, we presented a method for exploiting visual and prosodic features for presentation quality detection, shedding light on how prosodic and visual features are related to the quality of a presentation. The proposed approach may be useful in the field of multimodal learning analytics which seeks to analyse different aspects of public presentations in order to understand the learning process and provide feedback to the trainee presenter. These techniques have been implemented as a component of a multimodal dialogue system intended to monitor the presentation performance of a public speaker, in the EU METALOGUE project [1]. Regarding future work, we are considering incorporating text analysis, and adding more gestures and body poses. It may be also relevant to examine co-occurrence of speech and gesture, in the spirit of multimodal analysis, and to attempt finer-grained class detection for quality ranking.

[2] http://uk.mathworks.com/products/matlab/

6. REFERENCES

[1] J. Alexandersson, M. Aretoulaki, N. Campbell, M. Gardner, A. Girenko, D. Klakow, D. Koryzis, V. Petukhova, M. Specht, D. Spiliotopoulos, et al. Metalogue: A multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management. In *International Conference on Intelligent Environments (IE), 2014*, pages 365–368. IEEE, 2014.

[2] L. Chen, C. W. Leong, G. Feng, and C. M. Lee. Using multimodal cues to analyze MLA'14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, MLA '14, pages 45–52, New York, NY, USA, 2014. ACM.

[3] M. A. DeCoske and S. J. White. Public speaking revisited: delivery, structure, and style. *American journal of health-system pharmacy*, 67(15):1225–1227, August 2010.

[4] V. Echeverría, A. Avendaño, K. Chiluiza, A. Vásquez, and X. Ochoa. Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, MLA '14, pages 53–60, New York, NY, USA, 2014. ACM.

[5] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, pages 57–106, 1981.

[6] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56:316–324, 2002.

[7] D. Grandstaff. *Speaking as a Professional: Enhance Your Therapy Or Coaching Practice Through Presentations, Workshops, and Seminars*. A Norton Professional Book. W.W. Norton & Company, 2004.

[8] R. Hincks. Measuring liveliness in presentation speech. In *INTERSPEECH*, pages 765–768, 2005.

[9] J. Krajewski, A. Batliner, and S. Kessel. Comparing multiple classifiers for speech-based detection of self-confidence-a pilot study. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3716–3719. IEEE, 2010.

[10] J. Lamerton. *Public Speaking. Everything you need to know*. Harpercollins Publishers Ltd, 2001.

[11] S. Luz. The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Transactions on Information Systems*, 30(3):17:1–17:24, 2012.

[12] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa. 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 '14, pages 37–44, New York, NY, USA, 2014. ACM.

[13] D. McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.

[14] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz. Mla'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 531–532, New York, NY, USA, 2014. ACM.

[15] M. Oliveira. *The role of pause occurrence and pause duration in the signaling of narrative structure*, volume 2389 of *LNAI*, pages 43–51. Springer, 2002.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013.

[17] J. Sinclair. *The Collins COBUILD English Language Dictionary*. HarperCollins, London, 1995.

[18] R. Sousa and A. Ferreira. Evaluation of existing harmonic-to-noise ratio methods for voice assessment. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA), 2008*, pages 73–78, Sept 2008.

[19] H. Stassen et al. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, 27(3):289–307, 1993.

[20] H. Traunmüller and A. Eriksson. The perceptual evaluation of f0 excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America*, 97(3):1905–1915, 1995.

[21] E. Yumoto, W. J. Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, 71(6):1544–1550, 1982.