A Multimodal Approach for Automatic Assessment of School Principals’ Oral Presentation during Pre-service Training Program

Shan-Wen Hsiao¹, Hung-Ching Sun¹, Ming-Chuan Hsieh², Ming-Hsueh Tsai², Hsin-Chih Lin², Chi-Chun Lee¹

¹Department of Electrical Engineering, National Tsing Hua University, Taiwan
²National Academy for Educational Research, Taiwan

Abstract

Developing automatic recognition systems of subjective rating using behavior data, collected using audio-video recording devices, has been at the forefront of many interdisciplinary research efforts between behavior science and engineering in order to provide objective decision-making tools. In the field of education, pre-service training program for school principals has become more critical due to the increasingly complex and demanding nature of the job. In this work, we collaborate with researchers from the National Academy for Educational Research to develop a system to assess pre-service principals’ oral presentation skill. Our recognition framework incorporates multimodal behavioral data, i.e., audio and video information. With proper handling of label normalization and binarization, we achieve an unweighted average recall of (0.63, 0.70, 0.67) or (0.67, 0.68, 0.67) depending on the choice of labeling schemes, i.e., original or rank-normalized, on differentiating between high versus low performing scores. The three oral presentation rating dimensions used in this work are Dim₁: content + structure + word, Dim₂: prosody, Dim₃: total score.

Index Terms: behavioral signal processing (BSP), oral presentation, multimodal signal processing, education research

1. Introduction

Modeling humans’ observable behaviors and hidden internal states computationally have gained tremendous interest in the engineering community. Recent works in various emerging fields, such as affective computing [1], social signal processing (SSP) [2], and behavioral signal processing (BSP) [3], have all made advancement in deriving novel computational algorithms to objectively quantify and automatically recognize humans’ emotion (e.g., [4, 5, 6]), social behaviors (e.g., [7, 8]), and various domain-specific behavior attributes (e.g., [9, 10, 11]) through the use of signal processing and machine learning techniques. In specific, the interdisciplinary field of BSP focuses on developing computational methods in close collaboration with domain experts such that the research outcomes could provide domain-sensitive decision-making tools for the experts. Exemplary BSP works in mental health, e.g., couple therapy [9, 12], addiction [13], and autism spectrum disorder [10, 14], in professional acting, e.g., [15], and in education, e.g., literacy assessment [16], have all demonstrated that by applying BSP techniques in modeling human behaviors, it would result not only in new development of signal processing algorithms but also in promises of novel scientific insights.

In this paper, we investigate the use of BSP techniques toward developing an automatic system to evaluate school principal candidates’ oral presentation skill. In the current climate of high expectation of education, principal of the school serves a critical role in both advancing teaching and driving the needed changes throughout the existing education system. Researchers have pointed out that since educational environment becomes increasingly complex as a result of irresistible school changes and constant rapid educational reforms, school leaders should/would benefit from pre-employment training and continuing professional development ([17, 18, 19, 20, 21]). Research around designing appropriate and effective training program for school principals has become a prevalent topic in the field of education research. Furthermore, the anticipated skills of successful principal-ship include the ability of being an effective instructional leader, i.e., to resolve complex problems and communicate effectively [22]. Especially, oral presentation skilled has been regarded as an important ability because school principals often face a variety of challenging tasks everyday and are required to communicate to different levels of personnel at school, such as teachers, students, and staff members [23].

National Academy for Educational Research (NAER) has been entrusted by Ministry of Education (MOE) in Taiwan with pre-service school principal training program. There are around 200 candidates participated in the training program at NAER every year. As part of the training program, each candidate has to perform a 3-minute long impromptu speech presentation as part of their final program evaluation. The impromptu speech aims at assessing principals immediate organization of speech planning and communication strategy. Their performances are graded by coaching principals on seven dimensions of their speech (listed in Section 2.1), and this score counts 5% toward the final grade that each participant receives at the end of the program. Due to the nature of subjectivity of oral presentation assessment, grading impromptu speech is not only time-consuming but also error-prone. This program repeats every year with fresh candidates. However, access to experienced and willing coaching principals is difficult and often results in having the same group of coaching principals every year. NAER has hence launched a research effort into automating oral presentation assessment in order to mitigate these perennial issues. Relatively few related works that we know of have worked on automatic oral presentation skill assessment in the educational setting. Cheng et al. and Salvagnini et al. have utilized multimodal information to predict ratings of online lectures, i.e., the 5-scale rating done by internet users on recordings of lectures on videolectures.net [24, 25]. These works demonstrated promising recognition accuracy. In this work, we utilize a multimodal approach in classifying the high versus low performing candidate principals along three specific dimensions of the impromptu speech (detailed in Section 2.2):

- Dim₁: content + structure + word
These four dimensions cover 70% of the total final score, and performance ratings: content, structure, word, and prosody. In this work, we start with four different dimensions of speech. The total overall score is a summation of the above seven dimensions. We used one high-definition Sony camcorder equipped with an external directional microphone. The recording environment was in a classroom with audiences where the speaker used a hand-held microphone connected to loudspeakers; the placement of camcorder was consistent in order to preserve a constant upper-body view of the speaker.

2. Research Methodology

2.1. Database Collection

We collected audio-video data at the premise of NAER as part of the 2014 pre-service principal training program. There were four different classes, including a total of 200 pre-service principals for elementary and secondary schools, in this training program. Out of 200 principals, only 128 had at least two different coaching principals (a total of eight coaching principals) rated their speech along the following seven dimensions:

1. content: content in line with the topic (0-20)
2. structure: well-formed structure (0-20)
3. word: appropriate usage of words targeted for the audience (0-20)
4. etiquette: proper etiquette (0-10)
5. enunciation: correct enunciation (0-10)
6. prosody: appropriate and fluent expressive prosody (0-10)
7. timing: proper timing control (0-10)

The total overall score is a summation of the above seven dimensions. We used one high-definition Sony camcorder equipped with an external directional microphone. The recording environment was in a classroom with audiences where the speaker used a hand-held microphone connected to loudspeakers; the placement of camcorder was consistent in order to preserve a constant upper-body view of the speaker.

2.2. Evaluation Labels of Interest

In this work, we start with four different dimensions of speech performance ratings: content, structure, word, and prosody. These four dimensions cover 70% of the total final score, and we also believe the ratings of these dimensions are more correlated with the speakers’ expressive behaviors, i.e., speech acoustic characteristics and gestural movements, than other codes, e.g., proper etiquette – depends mainly on the dress of the speaker. These four dimensions of ratings are further grouped into two dimensions:

- Dim1 = content + structure + word
- Dim2 = prosody

Table 1 depicts the correlation between these four dimensions (computed by first averaging the score between the two coaching principals, then calculate the correlation between codes).

<table>
<thead>
<tr>
<th>content</th>
<th>structure</th>
<th>word</th>
<th>prosody</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.86</td>
<td>0.80</td>
<td>0.54</td>
</tr>
<tr>
<td>0.86</td>
<td>1.0</td>
<td>0.85</td>
<td>0.57</td>
</tr>
<tr>
<td>0.80</td>
<td>0.85</td>
<td>1.0</td>
<td>0.56</td>
</tr>
<tr>
<td>0.54</td>
<td>0.57</td>
<td>0.56</td>
<td>1.0</td>
</tr>
</tbody>
</table>

From Table 1, it is evident that content, structure, and word dimensions are highly correlated with each other but not with prosody. Hence, in this work, we combine these three ratings into one dimension by summing them into a single score (Dim1) but leave prosody as its own dimension (Dim2). In this work, we also include the third dimension (Dim3) to be modeled, Dim3 = total score which is essentially the summation of the all seven dimensions.

2.2.1. Binarizing Labels and Rank Normalization

Binarizing labels is a common practice in several previously published related works in training machine learning systems to recognize subjective attributes [9, 24]. The key idea to is allow the system to learn from the ground truth labels that are more reliable and consistent (extreme behaviors are easier to be rated consistently by humans). At the same time, researchers can examine whether developing such an automatic system is feasible under this setup. Therefore, in this work, we employ the same approach. We binarize each of the three dimensions (see Section 2.2) of assessment ratings into high versus low by choosing data samples, i.e., the 3-minute long impromptu speech, that are rated in the top and bottom 20% of the entire corpus.

Furthermore, while there is instruction on how to rate the speech based on each of the seven dimensions, individual
coaching principal can score the speech with a different dynamic range. For example, Figure 1 (top left) shows a histogram of the original scores (Dim1) rated by principal Shuo-Sung Chang and Jung-Hui Chang on the same set of speech data. It is evident by simply adding the original scores before binarizing the label would bias the choice of high versus low toward a particular coaching principal with wider score spread (especially obvious at the low performing group), i.e., in this case Jung-Hui Chang.

In order to mitigate this issue, we employ a rank (fusion) label normalization method (Figure 1 (bottom left)) to combine the scores between the two coaching principals into a single representative score for each sample. The rank fusion method is a common methodology in information retrieval and label normalization [26]. The method first turns the scores of each individual evaluator into rank order, then normalize this rank by dividing with the total number of samples for this evaluator. Finally, we add this normalized rank score for the two coaching principals before binarizing them into high and low classes.

Finally, in this work we utilize the following list of labels to train and test our automatic recognition system:

- Original scores: Dim\textsubscript{ori}, Dim\textsubscript{2ori}, Dim\textsubscript{3ori}
- Rank fusion scores: Dim\textsubscript{1rank}, Dim\textsubscript{2rank}, Dim\textsubscript{3rank}

Note that Dim\textsubscript{ori} is computed by first performing rank fusion along each of the codes, i.e., content, structure, and word, before adding them. The dataset of interest is the top 20% and bottom 20% of each of the dimensions, which results in around 60 samples (30 high and 30 low). Figure 1 (right) shows the cut-off scores that we use.

3. Experimental Setup and Results

3.1. Experimental Setup

Figure 2 summarizes our system components and experimental setup. Each recorded video consists of a candidate principal’s 3-minute long impromptu speech, we perform manual segmentation first to segment the speech into meaningful utterances. The utterances lengths are around 10 seconds. Our multimodal system consists of separate recognition system (the base classifier is support vector machine [27]) using audio and video information, and the final decision of high versus low for the three dimensions of interest (see Section 2.2) is done by decision-level fusion. Train-test evaluation scheme is done via leave-one-speaker-out cross validation.

Figure 2: Our experimental setup: the raw recording is first manually-segmented into utterances and each utterance is run through audio and video feature extractor component. Video-only system is trained on individual utterances, and audio-only system is trained on entire speech by utilizing second stage statistical functional computation. Classifier of choice is support vector machine, and the multimodal fusion is done by training logistic regression on the decision scores of each modality.

3.1.1. Audio-only recognition

The high-dimensional acoustic feature extraction approach has been utilized in many recognition tasks from speech, e.g., paralinguistic recognition [4], behavior detection [9], and emotion recognition [28]. This approach of quantifying speech attributes is effective especially when dealing with challenging recognition tasks involves subtle modulating information that is multi-scale and non-linear, e.g., emotion effect on speech acoustics.

In this work, we employ the same high-dimensional acoustic feature extraction procedure on each utterance using OpenSMILE toolbox with emo2010 configuration [29], which results in 1582 features per utterance; we then z-normalize these features with respect to individual speaker. Furthermore, since the label is assigned at the speech-level, we additionally compute four different descriptive statistics, i.e., mean, variance, skewness, and kurtosis, over the entire 3-minute speech to capture aspects of dynamics on these z-normalized 1582 features. It results in $1582 \times 4 = 6328$ features per data sample.

Lastly, we perform feature selection on this large amount of features before training support vector machine on the training data within each fold of cross validation. The feature selection technique employed here is based on univariate test, i.e., one-way ANOVA F-test done on each one of the 6328 features sequentially. Empirically, we select 500 of top ranked features (i.e., the lowest 500 features as ranked by p-values) to be trained in support vector machine (linear kernel with $C = 1$).

3.1.2. Video-only recognition

In this work, we use dense trajectory method to compute video features. The framework was proposed by Wang et al. [30] for action recognition, and this approach has been successfully utilized in other recognition tasks from videos [31, 32]. The basic idea is to densely sample the video sequences to find feature points instead of the usual techniques of sampling sparse key points. The methodology tracks the dynamics of the feature point using optical flow and median filtering over time. With these densely-sampled feature points’ trajectory, i.e., so called dense trajectories, we can then derive the following local descriptors in spatio-temporal grid (details about the algorithmic procedure can be found in Wang’s paper, and all the parameters involved in computing videos descriptors are the same [30] except that pixel width is set to eight to reduce the size of data):

- Traj: trajectories’ $(x, y)$ information
- HOG: histogram of oriented gradients
- HOF: histogram of optical flow
3.2. Experimental Results and Discussions

Table 2 summarizes our experimental results. Metrics reported in table are unweighted average recall. The row of original corresponds to results on classifying high versus low class that are defined on the original raw scores, and the row of rank-normalized corresponds to results on classifying high versus low class that are defined on the scores after label rank normalization (Section 2.2.1).

There are several major points to make. First is that most entries in Table 2 exceed chance performance, i.e., 50%, by a significant margin indicating there indeed exists useful behavioral information in audio and video stream that we are capable of modeling. Second, our results also indicate that rank normalization technique improves recognition accuracies for audio-based system significantly, especially Dim2 and Dim3.

From Table 3, it shows the per-class recognition results on the rank-normalized labels, and the effect is quite evident in the audio recognition. Results in Table 3 further points to a trend that with rank normalization, recall rate on the low class improve much more significantly than the high class in the audio-based system. We hypothesize that the reason could be that if our data samples in the low class are assigned based on the original score, this process can be sensitive to the differences

<table>
<thead>
<tr>
<th>Modality</th>
<th>Audio-only</th>
<th>Video-only</th>
<th>Multimodal Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dim1</td>
<td>Dim2</td>
<td>Dim3</td>
</tr>
<tr>
<td>Original</td>
<td>0.70</td>
<td>0.44</td>
<td>0.57</td>
</tr>
<tr>
<td>Rank-Normalized</td>
<td>0.68</td>
<td>0.60</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3: Audio-only recognition: per-class recall percentage

<table>
<thead>
<tr>
<th></th>
<th>Original Score</th>
<th>Rank-Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high class</td>
<td>low class</td>
</tr>
<tr>
<td>Dim1</td>
<td>0.71</td>
<td>0.09</td>
</tr>
<tr>
<td>Dim2</td>
<td>0.62</td>
<td>0.35</td>
</tr>
<tr>
<td>Dim3</td>
<td>0.55</td>
<td>0.59</td>
</tr>
</tbody>
</table>

4. Conclusions

Oral presentation skill is regarded as one of the most important attribute for a school leader. NAER has been entrusted by MOE in Taiwan with pre-service principals’ training program. Due to the time-consuming and subjective process in terms of evaluating these candidate principals’ 3-minute long impromptu speech, a new research effort for developing an automatic assessment system based on behavioral data has been initiated to mitigate these issues and to counter the problem of limited availability of coaching principals. From Section 3.2, it is encouraging to see that our initial system is capable of classifying between high versus low (i.e., good and bad) oral presentations along the three major behaviorally-related dimensions within a multimodal framework by using proper rank-normalization dealing with coaching principals’ idiosyncrasy.

There are multiple directions of future works. One of the short time goals is to further improve this process, e.g., utilization of voice activity detector to automatically segment the speech, and completion of the system by scoring the entire database. On the behavior modeling side, the framework does not include information about the lexical content of the speech, and the computational framework is at the moment static and requires heavy computational power to extract features. Lastly, as the 2015 pre-service principal training program has started, we plan to collect more data and to mitigate some of the labeling issues mentioned by recruiting more experiences coach principals to provide extra ratings on these oral presentations.

5. Acknowledgments

Thanks to MOST, Taiwan (103-2218-E-007-012-MY3) and NAER, Taiwan (NAER-104-12-B-2-02-00-1-02) for funding.
6. References


