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

Shan-Wen Hsiao[1], *Hung-Ching Sun*[1], *Ming-Chuan Hsieh*[2], *Ming-Hsueh Tsai*[2], *Hsin-Chih Lin*[2], *Chi-Chun Lee*[1]

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

10.21437/Interspeech.2015-545

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 effort 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 becoming 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 in order 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 Dim_1: content + structure + word, Dim_2: prosody, Dim_3: 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 investigates 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 shoud/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 personnels 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):

- **Dim_1: content + structure + word**

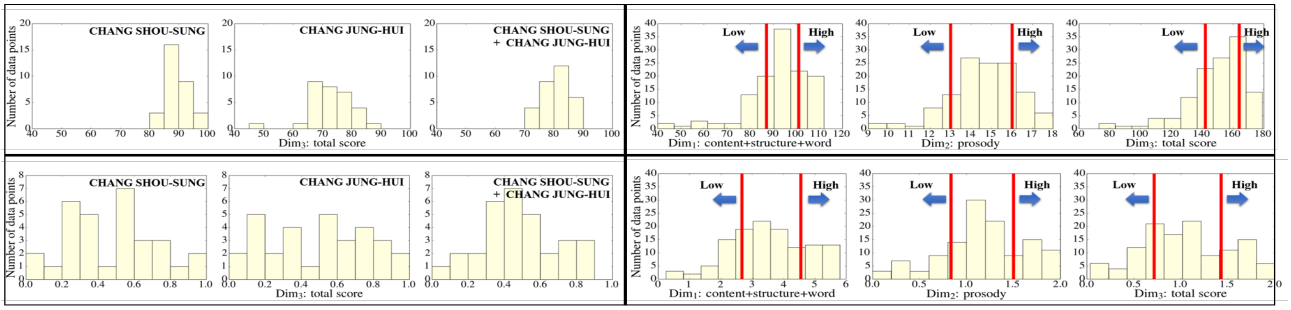

Figure 1: (Left) On the top row, it show the distribution of the original scores of Dim_3 graded by two different coaching principals on the same set of data. The bottom row shows the label distribution after the rank normalization. (Right) It shows the cut off score into *high* and *low* labels for all three dimensions of interest

- **Dim$_2$: prosody**
- **Dim$_3$: total score**

With multimodal fusion of video and audio information, our system achieves unweighted average recall of (0.63, 0.70, 0.67) or (0.67, 0.68, 0.67) depending on the choice of labeling schemes on each dimension respectively. These initial promising results indicate the feasibility of applying BSP techniques in designing computational frameworks that are domain-driven and encourage our continuing effort into developing a complete oral presentation scoring system.

In this paper, we describe the BSP pipeline from data collection, subjective labeling, to multimodal fusion framework. Section 2 describes about database collection and label preprocessing, section 3 includes experimental setups and results, and section 4 concludes with future works.

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:

- Dim_1 = content + structure + word
- Dim_2 = 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 1: *Spearman correlation between four dimensions of ratings: content, structure, word, and prosody*

	content	**structure**	**word**	**prosody**
content	1.0	0.86	0.80	0.54
structure	0.86	1.0	0.85	0.57
word	0.80	0.85	1.0	0.56
prosody	0.54	0.57	0.56	1.0

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 (Dim_1) but leave *prosody* as its own dimension (Dim_2). In this work, we also include the third dimension (Dim_3) to be modeled,

- Dim_3 = 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

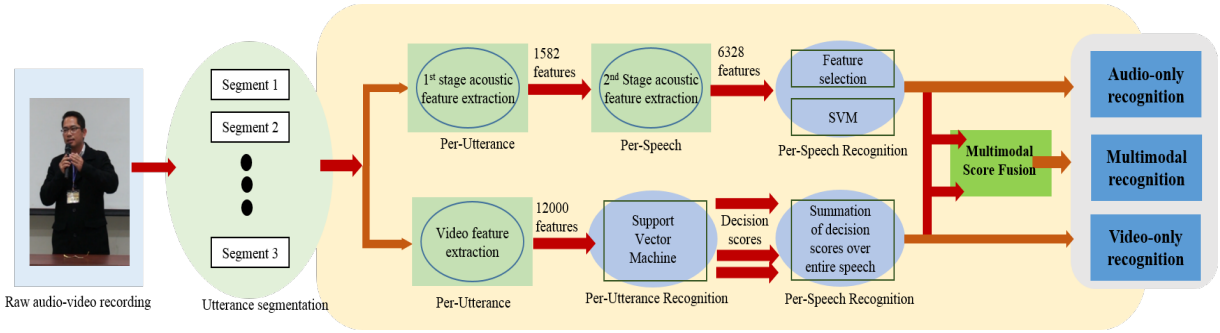

Figure 2: Our experimental setup: the raw recording is first manually-segmented into utterances and each utterance is run through audio and video feature extractor component. Video-only system is trained on individual utterances, and audio-only system is trained on entire speech by utilizing second stage 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.

coaching principal can score the speech with a different dynamic range. For example, Figure 1 (top left) shows a histogram of the original scores (Dim_3) 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: $\text{Dim}_{1\text{ori}}$, $\text{Dim}_{2\text{ori}}$, $\text{Dim}_{3\text{ori}}$
- Rank fusion scores: $\text{Dim}_{1\text{rnk}}$, $\text{Dim}_{2\text{rnk}}$, $\text{Dim}_{3\text{rnk}}$

Note that $\text{Dim}_{1\text{rnk}}$ 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.

3.1.1. Audio-only recognition

The high-dimensional acoustic feature extractions 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 have 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

Table 2: *Summary of recognition experiments' accuracies as measured by unweighted average recall*

Modality	Audio-only			Video-only			Multimodal Fusion		
	Dim$_1$	**Dim**$_2$	**Dim**$_3$	**Dim**$_1$	**Dim**$_2$	**Dim**$_3$	**Dim**$_1$	**Dim**$_2$	**Dim**$_3$
Original	0.70	0.44	0.57	0.67	0.68	0.70	0.63	**0.70**	0.70
Rank-Normalized	0.68	0.60	0.62	0.67	0.69	0.68	0.67	0.68	0.67

- **MHB**$_x$: motion boundary histogram in the x direction
- **MHB**$_y$: motion boundary histogram in the y direction

HOG mostly describes about the static appearance of the image. Since we are interested in the bodily gestural movement, we do not include HOG. Furthermore, while HOF contains motion information, it is about absolute motion, which can be sensitive to camera motion. Hence, we only compute **Traj**, **MHB**$_x$, and **MHB**$_y$: **Traj** has 30 dimensions per frame, **MHB**$_x$ and **MHB**$_y$ each has 96 dimensions per frame (frame rate is 15 Hz).

In order to construct a single video feature vector per utterance, we use the bag-of-feature approach. For each of the feature type, i.e., **Traj**, **MHB**$_x$, and **MHB**$_y$, we construct a code book of *visual words* by first randomly sample 100,000 frames of our complete database to perform *k*-means clustering ($k = 4000$). Then for each utterance, we use the histogram counts of the occurrences of each visual word (i.e., using Euclidean distances to find the closest cluster); lastly, we normalize the histogram counts, i.e., 4000 dimensional feature vector for each video feature type, using z-normalization.

Support vector machine is trained on 12000 dimensional features per utterance by assuming each utterance segment carries the same label as the entire speech (linear kernel with $C = 1$). In order to decide whether a speech belongs to the *high* or *low* class, we average the decision function, i.e., the distance to hyperplane, over the entire speech; depending on the final sign, we then assign the label to its respective class.

3.1.3. Multimodal fusion

Our multimodal fusion occurs at the *speech-level* by training a second stage logistic regression on the decision function from the audio recognition output and the length normalized decision function from the video recognition output.

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 indeeds exists useful behavior 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 **Dim**$_2$ and **Dim**$_3$.

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 in the coaching principals' grading dynamic range (as Figure 1 left demonstrates a *harsher* grading done by a coaching principal would bias the selection of *low* performing speech in our dataset), and that sensitivity could results in a degradatio of audio-based recognition. However, the same effect does not show in the video case, where the recognition rate is quite consistently high, e.g., possibly due to the high discriminatory ability and the robustness in the dense trajectories-based features.

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

	Original Score		Rank-Normalized	
	high class	*low* class	*high* class	*low* class
Dim$_1$	0.71	0.69	0.70	0.67
Dim$_2$	0.62	0.35	0.60	**0.61**
Dim$_3$	0.55	0.59	0.57	**0.67**

In general, we also see that the recognition rate is higher in video-only system, possibly due to the fact that image quality is much better and consistent as compared to audio recording. The recording microphone is situated at a far field location with the constantly changing background noise, and even the audiences would interact with the speaker during their speech. However, in the final multimodal fusion for **Dim**$_2$: prosody, by using audio and video system, the recognition rate improves to 70%.

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

[1] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[2] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[3] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[4] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[5] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2012.

[6] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[7] G. Varni, G. Volpe, and A. Camurri, "A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 576–590, 2010.

[8] A. S. Pentland, "Social signal processing [exploratory dsp]," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 108–111, 2007.

[9] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[10] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[11] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan, "thats aggravating, very aggravating: Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 87–96.

[12] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[13] B. Xiao, D. Bone, M. Van Segbroeck, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[15] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1766–1778, 2014.

[16] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 1015–1028, 2011.

[17] S. Watson, T. Miller, L. Johnston, and V. Rutledge, "Professional development school graduate performance: Perceptions of school principals," *The Teacher Educator*, vol. 42, no. 2, pp. 77–86, 2006.

[18] P. S. Keung, "Continuing professional development of principals in hong kong," *Frontiers of Education in China*, vol. 2, no. 4, pp. 605–619, 2007.

[19] P. S. Salazar, "The professional development needs of rural high school principals: A seven-state study." *Rural Educator*, vol. 28, no. 3, pp. 20–27, 2007.

[20] Y. Cheong Cheng, Y.-Q. Mao, W. Yan, and L. Catherine Ehrich, "Principal preparation and training: a look at china and its issues," *International Journal of Educational Management*, vol. 23, no. 1, pp. 51–64, 2009.

[21] D. L. Keith, "Principal desirabilitiy for professional development," *Academy of Educational Leadership Journal*, vol. 15, no. 2, p. 95, 2011.

[22] N. A. of Secondary School Principals, "Selecting and developing the 21st century school principal," http://www.nassp.org/tabid/3788/default.aspx?topic=26775, 2014.

[23] I. Muse, *Oral and Nonverbal Expression*. Routledge, 2013.

[24] D. S. Cheng, H. Salamin, P. Salvagnini, M. Cristani, A. Vinciarelli, and V. Murino, "Predicting online lecture ratings based on gesturing and vocal behavior," *Journal on Multimodal User Interfaces*, vol. 8, no. 2, pp. 151–160, 2014.

[25] P. Salvagnini, H. Salamin, M. Cristani, A. Vinciarelli, and V. Murino, "Learning how to teach from videolectures: automatic prediction of lecture ratings based on teacher's nonverbal behavior," in *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*. IEEE, 2012, pp. 415–419.

[26] H. D. Kim, C. Zhai, and J. Han, "Aggregation of multiple judgments for evaluating ordered lists," in *Advances in information retrieval*. Springer, 2010, pp. 166–178.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[30] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[31] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 702–707.

[32] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3681–3688.