Toward Automating Oral Presentation Scoring During Principal Certification Program Using Audio-Video Low-Level Behavior Profiles

Shan-Wen Hsiao, *Student Member, IEEE*, Hung-Ching Sun, *Student Member, IEEE*, Ming-Chuan Hsieh, Ming-Hsueh Tsai, Yu Tsao, *Member, IEEE*, and Chi-Chun Lee, *Member, IEEE*

Abstract—Effective leadership bears strong relationship to attributes of emotion contagion, positive mood, and social intelligence. In fact, leadership quality has been shown to be manifested in the exhibited communicative behaviors, especially in settings of public speaking. While studies on the theories of leadership has received much attention, little has progressed in terms of the computational development in its measurements. In this work, we present a behavioral signal processing (BSP) research to assess the qualities of oral presentations in the domain of education, in specific, we propose a multimodal framework toward automating the scoring process of pre-service school principals' oral presentations given at the yearly certification program. We utilize a dense *unit-level* audio-video feature extraction approach with session-level behavior profile representation techniques based on bag-of-word and Fisher-vector encoding. Furthermore, we design a scoring framework, inspired by the psychological evidences of human's decision-making mechanism, to use confidence measures outputted from support vector machine classifier trained on the *distinctive* set of data samples as the regressed scores. Our proposed approach achieves an absolute improvement of 0.049 (9.8 percent relative) on average over support vector regression. We further demonstrate that the framework is reliable and consistent compared to human experts.

Index Terms—Behavioral signal processing (BSP), oral presentation, multimodal signal processing, educational research

1 Introduction

Charismatic and effective leadership has been shown to be related to the phenomenon of emotion contagion [1], [2], positive mood [3], and attribute of social intelligence [4]. A recent meta study summarizing twenty years of research in studying the relationship between leadership, affect, and emotions in social science shows that there exists a wealth of literature in formulating these theoretical concepts; in fact, the conceptualization in the complex interplay between these constructs have been quite established. However, the methodologies in deriving valid measurements of these attributes, i.e., largely based on different forms of self reports or human observers perception, have progressed very little over the years [5]. This continues to be a critical hurdle needed to be overcome computationally in further substantiating the concept of leadership in the perception of charisma and moving the theories forward [5], [6].

Furthermore, researchers have also pointed out evidences that this affect-based perception of charisma/leadership is, in fact, reflected in an individual's public speaking skill [7] or broadly in his/her communicative styles and strategies [8]. Studies have revealed that more than the word usage, the communicative strategy and expressivity (i.e., verbal characteristics and non-verbal behaviors) play defining roles in the perception of leadership and charisma [9]-it is, in fact, a result from intertwining effect between *content* and *delivery* when giving a speech [10], [11], [12]. Various domain experts have been striving not only to advance the theoretical underpinning of these fundamental attributes of an effective leader but also to derive appropriate methods in quantitatively assessing the quality of these attributes as manifested in these individuals. This is particularly relevant in fields such as political science [13], [14], business (organizational) management [15], and also education. While there is abundant research in understanding the *content*, the research into studying the *delivery* is much more limited, potentially due to a lack of adequate methodology. In this work, we present a multimodal behavior profiles framework to quantify communicative behaviors for tasks of assessing public speaking in the domain of education.

In fact, the notion that human internal states are manifested in different communicative channels has already sparked a tremendous computational effort. Engineers have developed algorithms in order to recognize humans socio- and emotional-attributes from observable behavioral cues through the use of signal processing algorithms and machine learning techniques. These effort have led to the emergence of

- *S.-W. Hsiao, H.-C. Sun, and C.-C. Lee are with Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, China. E-mail: qoopoo30@gmail.com, felicityphoebe@hotmail.com, cclee@ee.nthu.edu.tw.*
- *M.-C. Hsieh and M.-C. Tsai are with National Academy for Educational Research, New Taipei City 23703, Taiwan, China. E-mail: {mhsieh, mhtsai}@mail.naer.edu.tw.*
- *Y. Tsao is with Research Center for Information Technology, Academia Sinica, Taipei, Taiwan 11529, China. E-mail: yu.tsao@citi.sinica.edu.tw.*

Manuscript received 25 Nov. 2016; revised 9 Aug. 2017; accepted 24 Aug. 2017. Date of publication 7 Sept. 2017; date of current version 25 Nov. 2019. (Corresponding author: Chi-Chun Lee.)
Recommended for acceptance by S. Scherer.
Digital Object Identifier no. 10.1109/TAFFC.2017.2749569

1949-3045 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

several cross-cutting interdisciplinary fields, e.g., affective computing [16], social signal processing [17], and behavioral signal processing [18]. For example, progresses in affective computing have resulted in a large body of works on algorithmic designs for automatic emotion recognition. Researchers have utilized a wide range of behavioral descriptors across different modalities, such as speech [19], [20], body gestures [21], [22], facial expressions [23], and even multimodal behaviors [24], [25], to obtain robust recognition of human's emotional states. Aside from emotions, other subjective attributes, such as paralinguistic and social aspects of human behaviors, have also received much attention recently in the development toward automatic recognition [26]; some notable examples includes intoxication detection [27], dialog backchannel recognition [28], interest level recognition [29], etc.

While the general human behavior modeling tasks has progressed significantly, an interdisciplinary field, behavioral signal processing (BSP), emerges by building upon this wealth of research and focusing on modeling human behaviors in a tightly-integrative and contextualized manner. Instead of working on the general problem domain of human behaviors, it stresses close integration with the domain experts to enhace the scientific-rigor in the research, i.e., from multimodal data acquisition in ecologically-valid *real* setting, domain-sensitive algorithmic development, to proper experimental validation, in order provide meaningful analytics to enhance the experts decision-making process. Exemplary BSP research already exists in tasks of recognizing high-level and subjective attributes for domains of mental health, e.g., couple therapy [30], [31], drug addiction therapy [32], [33], and autism spectrum disorder [34], [35], of professional acting [36], and of education, e.g., second language literacy assessment [37]. Each of these BSP effort have demonstrated that it would result not only in novel signal processing algorithms that can measure domain-relevant constructs in real world problems, but also in promises of advancing the current scientific understanding of human.

In this work, we present a thorough BSP research toward quantitative modeling of leaders' communicative abilities in the domain of education. In specifics, we collaborate with educational researchers to contextualize such an analytical development in assessing the school candidate principals' impromptu presentation at an yearly certification program-an extension from our previous published work [38].

1.1 Background

In the current climate of society's high expectation for better education, irresistible school changes and constant rapid educational reforms have made the educational environment become increasingly more complex. Designing effective pre-employment certification and continuing training program to assess and improve desired leadership qualities has become a prevalent research topic in the field of education (e.g., [39], [40], [41]). In fact, the National Academy for Educational Research (NAER) has been entrusted by the Ministry of Education (MOE) in Taiwan with the design and implementation of pre-service school principal certification program. Every year, each candidate principals has to attend a 2-month program to become a certified *principal-to-be*.

The aim of the program is to rigorously evaluate multiple aspects of each candidate on their potential in being a school leader. Similar to leadership research in other fields, a key anticipated ability to be assessed is their communicative skill, which is important not only to resolve complex problems but also to steer and lead the direction of the school development [42]. Each candidate is required to perform a 3-minute long impromptu speech as part of their final examination to demonstrate their immediate speech planning and communicative strategy as a leader. Throughout the years, the scores have been graded by two senior coaching principals, which count 5 percent toward the final grade that each participant receives at the end of the program. Due to the nature of subjectivity in the oral presentation assessment, grading impromptu speech is not only time-consuming but also error-prone. Further, this program repeats every year with fresh candidates; however, access to experienced and eligible coaching principals is becoming more difficult every year. The NAER has, hence, launched a collaborative research effort into automating this subjective oral presentation scoring in order to mitigate these perennial issues

1.2 Related Works

1.2.1 Educational Research

In the field of education, there exists many different types of commonly-used automated systems for various performance ratings, including passage summary (written or spoken presentation and response), written product (essay, email, response to problem-solving scenario), spoken form (read aloud, retell or reconstruct sentences or phrases), short answer questions (written or spoken presentation and response), and oral reading fluency (accuracy and expression) [43]. In fact, commercial companies have developed automatic scoring systems for constructed language in assessment tasks, and these systems have been applied widely in large scale assessment [44]. For example Streeter et al. reports that Pearson's Knowledge Technology has applied automatic system to score more than 20 million spoken and written responses for different kinds of language-related assessments [43]. Aside from commercial applications, Balogh et al. evaluates spoken English tests for adults and achieves a machine-human correlation of 0.95 [45]; similarly, Berstein et al. evaluates a spoken Arabic test and obtains a score correlation of 0.97 [46]. Validity of these test scores can be further strengthened by comparing with scores obtained from other concurrent administrations of existing tests. For example, Bernstein et al. analyzes the validity of test scores from oral exams of four types of languages to quantify a person's effectiveness in spoken communication [47]. Their study shows that scores from the automated tests are strongly correlated (r=0.77 to 0.92) with the scores from oral proficiency interviews; similar conclusions are obtained by Bernstein & Cheng [48].

While there is a wealth of research works on automated scoring systems in the field of education, most of these systems rely on *well-controlled* tasks that are operational often in *limited* contexts, e.g., a pre-defined set of short utterances or spoken words or short written lexical contents. The assessment is also carried out only based on *single* surface form of human communication, e.g., written texts or spoken words. However, since the impromptu speech in this context reflects a higher-level attributes of a candidate principal, not only is the talk much *less-constrained* but also the assessment is, at the same time, much *subjective* in nature. This is an important gap needed to be filled with computational methods in the field of education assessment.

1.2.2 Engineering Research

Recently, there are also several related research works in the engineering domain targeted for automatic assessment using multimodal behavior cues in education setting [49]. For example, Ochoa et al. provide a nice summary showing the feasibility and the recent trends in developing educational learning analytics by using multimodal sensor recordings and automated computational methods in assessing student's learning performances in classes [50], [51]. Furthermore, Haider et al. show promising accuracies in rating students' presentation skill while giving powerpoint-aided presentations with a suite of multimodal cues on a large Spanish corpus [52]. In the area of developing algorithms in automatic assessment of public speaking, i.e., the closest application domain to this work, Batrinca et al. presents a platform of using virtual agents as audiences and develops a multimodal automatic system for assessing public speaking ability [53], and later a similar setup is carried out in a work done by Wörtwein et al. [54]. Chen et al. and Ramanarayanan et al. also present automatic frameworks of using a rich set of multimodal behavior features (i.e., data collected from audio and kinect sensors) on ratings of oral presentations [55], [56], [57]. While this body of works only recently emerge, this effort of learning analytics development already points toward the promises of utilizing automatic method in assessing various types of oral presentation skills in less constrained scenarios.

1.3 Our Contributions

The major theme of the work is on quantitative modeling of candidate principals' multimodal behavior profiles toward automated assessment of their communicative skills. We present a novel BSP human behavior research integrating the following three major contributions:

1) *Research Settings*: conducting the research in a *real* and *contextualized* examination scenarios
2) *Behavior Representations*: using data-driven *holistic* low-level multimodal behavior profiles to obtain both a *reliable* and a *consistent* assessment
3) *Automatic Scoring*: handling challenges in modeling high-level attributes by using confidence score from binary classifier (instead of conventional regression)

First, our work represents a collaborative BSP research. The spontaneous audio-video data is collected *directly* during the real candidate principals examinations; the behaviors that they exhibit are naturally ecologically-valid. Further, many past works show the efficacy in utilizing advanced sensor technologies, e.g., Kinect senors or depth camera, in a highly-instrumented recording space. In our context, in order to avoid unnecessary alteration to the existing implementation of the impromptu speech examination and to further ensure the algorithm's wide-applicability in the real-world NAER certification programs, we maintain the existing data recording setup using high-definition video camcorder that can also be easily scaled up. Also, the grading sheet that the algorithm is built for is the one that is currently in use by the educational experts. The outcome of the automated analytics can be easily integrated into the decision-making pipeline of the experts. This realistic nature of our work ensures the potential of our framework to achieve the most direct impact and provides ecologically-valid samples for human behavior studies at the same time.

Second, we derive low-level behavior profiles as general feature representations for the assessment algorithm. The idea emerges as the goal of the research is to optimize accuracies for automatic scoring instead to seek manually predefined discrete behaviors (e.g., looking at the audiences, proper eye contacts, adequate forearm gestures, etc) that is often useful in context of educational training. The holistic profile based approach could potentially include these discrete behaviors (in)-directly and possibly beyond. In fact, many past works have demonstrated that due to the complex characteristics of human behaviors, many state-of-art recognition framework benefit from a *holistic* representation of low-level descriptors (LLDs) describing a multitude aspects of speech and video signal's spectral-temporal characteristics directly. Various examples can be found across fields, e.g., event recognitions [58], action recognitions [59], [60], emotion recognitions [61], and detection of high-level couples' behaviors during therapy [30]. In this work, we propose to derive behavior profiles from both video and audio LLDs to assess the quality of impromptu speech.

Lastly, the subjectivity in the annotated labels often creates an issue in the algorithmic developmental process. Handling of such subjectivity in the past is largely carried out conventionally at the label preprocessing stage, e.g., averaging the raters to generate ground truth. In this work, we draw our inspiration from literature in psychology stating that if the underlying true decision-to-make (e.g., in this case: assessment of a candidate principal speech) is a (series) of *binary* choice(s) by nature rather than a *continuous* scale, by imposing human to make a continuous judgment would result in a *loss-of-information* [62], [63]. Hence, aside from preprocessing labels, we also propose an technical approach to extend the binary classification framework to directly regress a real-valued score using 'sample-to-boundary' distance as the final assessment score. It is an appealing methodology, especially in highly-professional domain, where expert ratings are often limited in number and simply out-sourcing the annotation can raise concern on the validity. In fact, the effect of subjectivity handling in our technical framework is demonstrated not only in the improved recognition accuracy, but also could be more consistent and less susceptible to unwanted variation compared to human experts shown in our consistency analysis.

To the best of our knowledge, there is few works that have systematically, i.e., from research setting, computational handling of subjectivity, to analyses, modeled human behaviors computationally in this context; the analytic generated provides a necessary methodological building block to further advancing the research in understanding leadership in education setting. Additionally, the algorithmic approaches and the recognition rates achieved can also be presented as benchmark results on the database, where the effort is undergoing to be released to the community. The rest of the paper is organized as follows: Section 2 describes about our multimodal database, collection methodology,

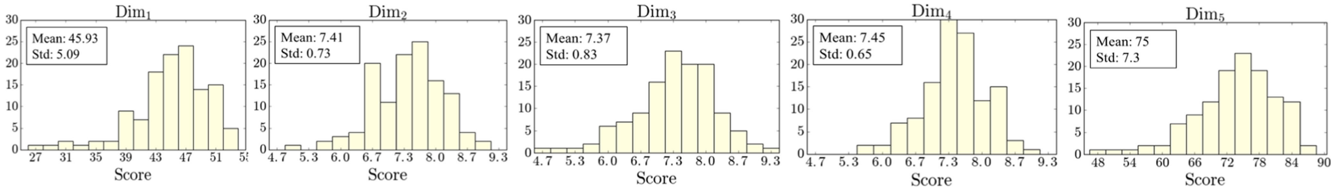

Fig. 1. It shows the distribution of each of the five dimensions of rating used in this work along with their means and standard deviations.

and annotation labels of interest. Section 3 describes about research methodology, including multimodal behavior representations, and our proposed automatic scoring method. Section 4 shows experimental setups, results, analyses, and discussions. Finally, Section 5 concludes with future works.

2 Database

2.1 The NAER Principals' Oral Presentation Database

The audio-video data was collected at the end of the 2014 NAER pre-service principals' certification program by the NAER researchers as part of the candidate principals' final examination. The program originally included a total of 200 candidate principals' impromptu speech given in Mandarin Chinese recorded using a single high-definition Sony camcorder with an externally-connected directional microphone. The final examination scores of these candidate principals served as a criteria for their future dispatchment to different schools. The camcorder was stationed on a tripod with the placement relatively consistent in order to preserve a constant and clear upper-body view of the speaker (Fig. 2 shows an actual snapshot of our raw data). The grading of each speech was done based on a score sheet that had been in use for the certification program over the years. The score sheet included the following seven items (the number in parenthesis indicates the range of possible scoring points of each item):

1) *content*: content in line with the topic chosen (0-20)
2) *structure*: well-formed speech organization (0-20)
3) *word*: appropriate word usage for the audience (0-20)
4) *etiquette*: proper etiquette and manner (0-10)
5) *enunciation*: correct enunciation (0-10)
6) *prosody*: appropriate, fluent, expressive prosody (0-10)
7) *timing*: proper timing control (0-10)

Each candidate principal was graded by multiple *coaching principals*. Qualified coaching principals were recruited due to their heavy involvement in the program over the years since the MOE launched this mandatory procedure. Finally, the certification program used the average of the summed total score over seven items computed from the rating of these recruited coaching principals as the final assessment score of each candidate principal's impromptu speech.

Fig. 2. It shows a snapshot of the raw data from two candidate principals.

There were four classes in the 2014 program with 12 coaching principals split in groups of three. Out of the 200 candidate principals, only 128 of them had their speech rated by at least three coaching principals; each triplet of coaching principals rated approximately 30 - 40 non-overlapping set of speech. This set of oral presentations constitutes the complete NAER principals' oral presentation dataset for this work. The mean duration of the speech in this corpus is 3 minutes 8.62 seconds and the standard deviation is 24.22 seconds. We additionally perform automatic speech segmentation of each audio file using an energy-based voice activity detector [64] generating a sequence of *pseudo-sentences* as a pre-processing step. Each *pseudo-sentence* is approximately 10 seconds long.

2.2 Assessment Dimensions of Interest

Table 1 summarizes the spearman correlation computed between pair-wise grading attributes. We observe that *content*, *structure*, and *word* are highly correlated (> 0.80) with each other, *prosody* and *enunciation* correlates with each other at 0.78, *etiquette* correlates with most of the ratings at least at the 0.65 level, and *timing* is the one that is least correlated with all the other rating items. One thing to note that, except for *timing* attribute, the rest correlates with each other fairly strongly. This results indicate that while these are seemingly disparate descriptions of ratings, they are, in effect, influencing each other in the evaluation process, i.e., an assessment of good presentation is holistic in nature including intertwining effect of *how* and *what* is given in the speech. In fact, principal component analysis (PCA) shows that with four hidden dimensions out of the seven categories, they reach over 95 percent of the variances in our assessment ratings.

In our previous work, we concentrated only on the rating categories of *content*, *structure*, *word*, and *prosody*, and *total* score [38]. In this work, we expand upon our past work in order to offer complete results by learning to score the following five dimensions of ratings. The value of each

TABLE 1
Average Spearman Correlation Between the 4-Pairs of Three Coaching Principals Computed for the Six Attributes Used on the Grading Sheet for Impromptu Speech Assessment (All of the Results Obtain a p-Value $< 1e^{-03}$)

Grading attributes' inter-correlation						
	structure	word	etiquette	enunciation	prosody	timing
content	0.90	0.81	0.66	0.62	0.64	0.44
structure		0.85	0.68	0.64	0.67	0.51
word			0.67	0.65	0.65	0.47
etiquette				0.69	0.72	0.44
enunciation					0.78	0.33
prosody						0.41

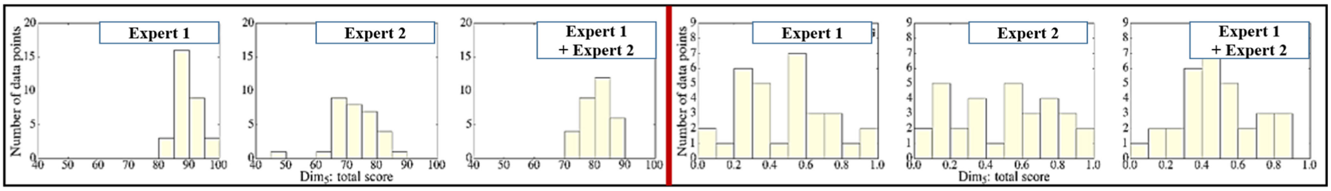

Fig. 3. (Left) it shows an example of the original scores', Dim_5, distribution that is graded by two different coaching principals, Expert 1 and Expert 2, on the same set of data; the dynamic ranges are clearly different. (Right) It shows an example of the label distribution after the rank normalization is done on each coaching principal.

category is computed by averaging the ratings given by the three coaching principals:

- **$\mathbf{Dim_1}$ = content + structure + word**
- **$\mathbf{Dim_2}$ = prosody**
- **$\mathbf{Dim_3}$ = etiquette**
- **$\mathbf{Dim_4}$ = enunciation**
- **$\mathbf{Dim_5}$ = total final score** (summation of all 7 categories)

Through our PCA analysis, the first principal axis (i.e., accounts for 81.6 percent of the total variances) weights are $[-0.54, -0.56, -0.53, -0.18, -0.13, -0.15, -0.20]$ for the attributes of *content*, *structure*, *word*, *etiquette*, *enunciation*, *prosody*, and *timing* respectively; this further demonstrates the correlative structure between these attributes. Hence, we decide to combine the three ratings, i.e., *content*, *structure*, and *word*, into a single dimension. Further note that *timing* is not included in this work as it is merely a tracking on how long does each speaker speaks. Fig. 1 shows the distribution, means, and standard deviations for each of the five dimensions of ratings used in this work.

2.2.1 *Pre-Processing Assessment Dimensions*

Furthermore, in order to mitigate the issue that each individual coaching principal may score a speech with different dynamic ranges, a common practice is to perform label normalization. In this work, we use rank label normalization technique [65] to achieve dynamic range normalization. The method transforms the original scores of each individual evaluator of each category into a rank order, and we then normalize this rank by dividing with the total number of samples for this evaluator (creating a number bounded between zero and one). Fig. 3 demonstrates an example that two different coaching principals (Expert 1 and Expert 2) having distinct scoring ranges although they both rated the same set of candidate principals' oral presentations. In this work, we finally include the following ten labels to train our assessment system and present all of their complete results in Section 4.2.

- Original: $\mathbf{Dim_{1O}}$, $\mathbf{Dim_{2o}}$, $\mathbf{Dim_{3o}}$, $\mathbf{Dim_{4o}}$, $\mathbf{Dim_{5o}}$
- Rank-normalized: $\mathbf{Dim_{1r}}$, $\mathbf{Dim_{2r}}$, $\mathbf{Dim_{3r}}$, $\mathbf{Dim_{4r}}$, $\mathbf{Dim_{5r}}$

A summary of the inter-evaluator agreement for the ten dimensions-of-interest (five original, five ranked) is listed in Table 2. We quantify inter-evaluator agreement by averaging the Cronbach's alpha computed between coaching principals' scores. It is interesting to see that these expert coaching principals achieve the highest inter-evaluator agreement in Dim_5 (the total score). This seems to implicate that experts judgment tend to agree more on higher-level rating (e.g., Dim_5 can be thought as the final evaluation of *how good is a talk overall*) more than the seemingly lower-level rating (e.g., Dim_2 is a rating on *how well is the prosody manifested*). This corroborates past research showing that the quality of the speech is often holistically integrative between content and delivery - resulting in an overall *felt-sense*. In fact, it also justifies the NAER researchers' use of Dim_5 as the assessment score marked on the final grade sheet of these candidates for the certification program. Moreover, we also see that by rank-normalizing labels, it improves the inter-evaluator agreement for all dimensions except Dim_2. One observation that we see with Dim_2 is not only that the range of scoring is small but also coaching principals tend to give a same score to multiple presentations, which may lead to a negative effect when performing rank-normalization.

2.3 Comparison to Similar Databases

The public speaking presentation databases collected for similar automated assessment task used in the past engineering works were often *simulated-in lab*, i.e., through recruiting volunteers to perform presentations in a well-instrumented lab space. For example, the setting used in work [54] was done in lab with availability of a platform of displaying virtual audiences. Their database included 2 recordings each for 45 participants using head mounted microphone, web camera, and a Microsoft Kinect as the participants engaged in interactive scene with virtual humans. Since their goal was mainly on training the subjects to improve their presentation behaviors not for large-scale assessment, the rating was designed by the authors including a pre-set list of discrete behaviors. The database used in works [55], [56], [57] included 56 presentations from 14 recruited speakers total. The multimodal behavior data was collected in lab using Kinect sensors with audio recordings. This database was collected with a similar goal as ours, i.e., to perform automatic assessment; hence, the assessment rating was done using a psychometric instrument called Public Speaking Competence Rubric (PSCR) on 10 dimensions. Each sample was annotated by *two* experts only, and the third expert was brought in only on the samples when there

TABLE 2
The Table Summarizes the Average Inter-Evaluator Agreement, i.e., Cronbach Alpha, Computed Between Coaching Principals on the Ten Dimensions of Labels Used in This Work

	Dim_{1O}	Dim_{2O}	Dim_{3O}	Dim_{4O}	Dim_{5O}
Database	0.51	0.48	0.40	0.53	0.58
	Dim_{1r}	Dim_{2r}	Dim_{3r}	Dim_{4r}	Dim_{5r}
Database	0.55	0.40	0.43	0.58	0.63

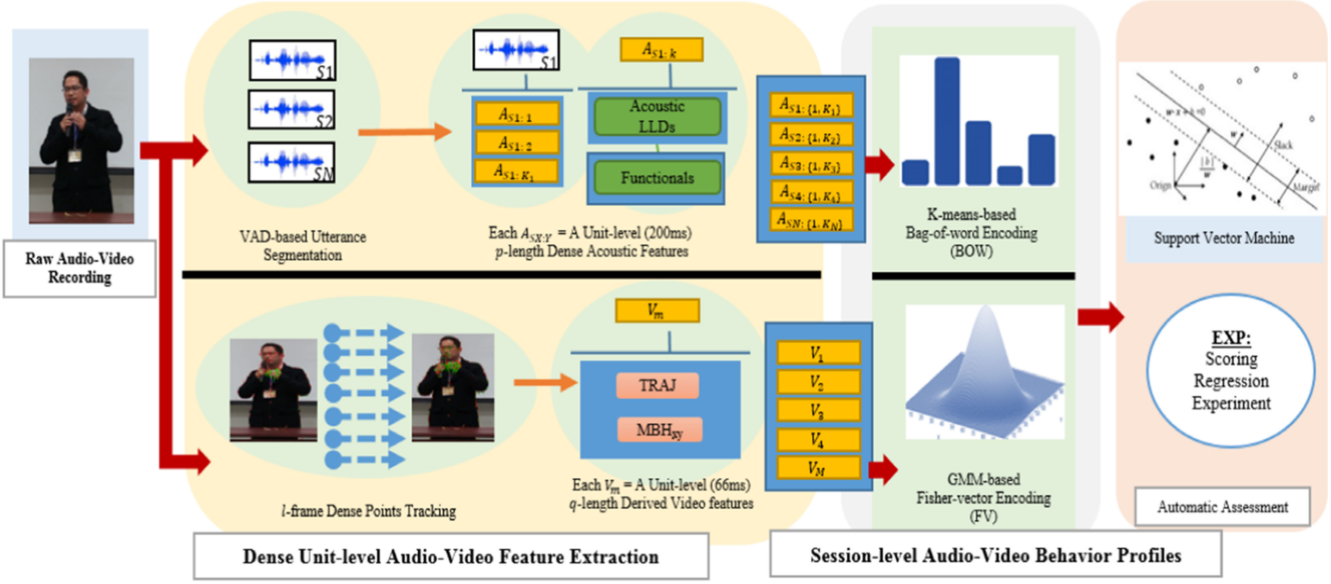
Raw Audio-Video Recording
S1
S2
SN
VAD-based Utterance Segmentation
S1
$A_{S1:k}$
$A_{S1:1}$
$A_{S1:2}$
$A_{S1:K_1}$
Acoustic LLDs
Functionals
Each $A_{SX:Y}$ = A Unit-level (200ms) p-length Dense Acoustic Features
$A_{S1:\{1,K_1\}}$
$A_{S2:\{1,K_2\}}$
$A_{S3:\{1,K_3\}}$
$A_{S4:\{1,K_4\}}$
$A_{SN:\{1,K_N\}}$
K-means-based Bag-of-word Encoding (BOW)
Support Vector Machine
Origin
Slack
Margin
l-frame Dense Points Tracking
V_m
TRAJ
MBH_{xy}
Each V_m = A Unit-level (66ms) q-length Derived Video features
V_1
V_2
V_3
V_4
V_M
GMM-based Fisher-vector Encoding (FV)
EXP: Scoring Regression Experiment
Automatic Assessment
Dense Unit-level Audio-Video Feature Extraction
Session-level Audio-Video Behavior Profiles

Fig. 4. It shows a complete diagram of our computational framework and experimentation in this work: dense *unit-level* audio-video feature extractions is performed on the raw audio-video recordings, and the k-means bag-of-word (BOW) and Fisher-vector (FV) encoding methods map the varying-length sequences of audio-video feature vectors to a single-fixed length vector as the *behavior profile* at the speech-level. Then, we carry out automatic scoring of the 10 ratings, i.e., 2 (original & rank-normalized) × 5 dimensions of interest.

was a disparate scoring between the original two experts. The inter-evaluator agreement on the 10 dimensions ranged from 0.15 to 0.88, where the overall *holistic* assessment achieved a modest agreement at 0.39.

While our setting is not directly comparable, we see that our research setting is more *contextualized-in the wild*, i.e., the NAER database utilized in this work constitutes *real* subjects in examination, *larger* sample size (128 presentations from 128 subjects), and *established* expert ratings (one that has already been used in the program over the years). Our data collection protocol is easily scalable to large-scale assessment task, especially important in domain of education.

3 Research Methodology

Our complete system is shown in Fig. 4. It consists of three major components: 1) audio-video low-level descriptors, 2) session-level behavior profiles, and 3) scoring using the support vector machine (SVM) classifier. Instead of first generating a list of manually pre-defined behaviors to look for, then to assess the overall quality of the presentation, we propose to compute *holistic* behavior profiles directly from multimodal low-level behavior descriptors; in fact, similar low-level encoding based approach has recently shown to be effective in complex tasks such as emotion and paralinguistic recognition (e.g., [66], [67]). The combination of multimodal behavior profiles with the scoring regression system using binary SVM classifier achieve desired recognition performances in this work.

In the following section, we describe the details about the multimodal *dense unit-level* audio-video feature extractions, the two different *session-level* feature encodings, i.e., k-means based bag-of-word model and Fisher-vector encoding to generate behavior profiles, and lastly, our proposed scoring regression system using binary SVM classifier trained on *distinctive* groups of samples.

3.1 Dense Unit-Level Acoustic Feature Extraction

We adopt the use of high-dimensional feature extraction approach in the acoustic modality as many past works have demonstrated this comprehensive data analysis approach is capable of modeling human acoustic in complex recognition tasks. We generate high-dimensional acoustic features at a *granular*-level using a sliding window approach, i.e., a window of 200 ms (roughly correspond to a syllable duration) with 50 percent overlap. This sliding window approach attempts to capture more *detailed* dynamics while maintaining adequate window length to properly maintaining the temporal characteristics of acoustic LLDs. We term each of this window as an *unit* at which we compute a high-dimensional acoustic feature vector using exhaustive functions. The exact extraction approach is carried out by first calculating various acoustic LLDs and then applying statistical functions on these LLDs. A complete list of acoustic LLDs and the statistical functions used is listed in Table 3.

The total number of acoustic features computed per *unit* is 9,063 (57 LLDs × 53 functions × 3 $\Delta\&\Delta\Delta$). We further perform z-score normalization on these features per speaker and discard features with zero variance; this results in a final dimension of 8,861 per *unit* - characterizing *extensive* aspects of acoustic-related properties within a *unit* window. In summary, this acoustic feature extraction approach works as follows: 1) applying VAD to automatically segment an oral presentation into N utterances, 2) generating a sequence of 8,861-dimensional feature vector per utterance using the sliding unit approach, and 3) depending on the number of *units* per utterance and the total number of N utterances, putting these features together to form a varying number of sequences of 8,861-dimensional acoustic features per speech, which is then be further converted to a fixed-length acoustic behavior profile using methods described in Section 3.3. The term *dense* essentially refers to both a more *granular* temporal scale and an *extensive* use of low-level

TABLE 3
The Table Provides a List of Statistics Applied to Various Acoustic LLDs to Form a High-Dimensional Dense Unit-Level Acoustic Feature Vector

Acoustic LLDs	
Low-level Descriptors (LLDs)	Type
zero-crossing rate, log energy, probability of voicing, F_0	prosodic
Mel-frequency Cepstral Coefficients (MFCCs) 0-12, spectral flux, spectral centroid, max, min, spectral bands 0-4 (0-9KHz), spectral roll-off (0.25, 0.5, 0.75, 0.9)	spectral
Functions applied to LLDs/ΔLLDs/ΔΔLLDs	
position of min/max, range, max − arithmetic mean, arithmetic mean − min	extremes
linear regression slope, offset, error, centroid, quadratic error, quadratic regression a, b offset, linear error, quadratic error (contour & quadratic regression)	regression
percentile range (25%, 50%, 75%), 3 inter-quartile ranges (25% - 50%, 50%-75%, 25%-75%)	percentiles
mean value of peaks, distance between peaks, mean value of peaks − arithmetic mean	peaks
arithmetic means, absolute value of arithmetic mean (original, non-zero values), quadratic mean (original, non-zero values), geometric mean (absolute values of non-zero values), number of non-zero values	means
relative duration LLD above 25%, 50%, 75%, 95% range, relative duration LLD is rising/falling, relative duration LLD has left/right curvature	temporal

descriptors. The approach mentioned above is carried out completely using opensmile toolbox [68].

3.2 Dense Unit-Level Video Feature Extraction

We use dense trajectory-based method to compute video features. The raw video resolution is 1920x1020 with a framerate of 30 Hz; we downsample the resolution to 640x480 with a framerate of 15 Hz before carrying out video descriptor extraction. The *unit* here simply refers to a video frame (15 Hz ≈ 66 ms). The framework is originally proposed by Wang et al. [59] and is effective in tasks of humans' action recognition (e.g., [58], [60]). A version of this approach has also recently been utilized in emotion recognition with body expressions [69]. In the following, we will briefly describe this video feature extraction approach. The core idea is to *densely* sample each video frame instead of trying to find key *feature* points. In essence, the algorithm first densely samples points within each frame, and then the algorithm prunes out *unnecessary* points that are either 'non-trackable' over time based on method of autocorrelation (e.g., those could be related to absence of any movement) or 'too much displacement' (most likely due to error in point-tracking). The methodology captures the movement dynamics of these densely-sampled points using an optical flow and a median filtering technique over time. After pruning, it would form a varying number of trajectories per frame; the tracking and sampling are reinitialized for every 15 frames.

Hence, with these densely-sampled points' trajectories, i.e., so called *dense trajectories*, we then derive the following descriptors in their respected spatio-temporal grid.

- **MBH**$_x$: motion boundary histogram in the x direction (the relative motion in the x direction)
- **MBH**$_y$: motion boundary histogram in the y direction (the relative motion in the y direction)
- **TRAJ**: dense trajectories' (x, y) normalized position displacement information

A description of the two derived video features is below:

TRAJ descriptors: Assume $P_t = (x_t, y_t)$ is a feature point at time t, we can track this point using the equation,

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{\hat{x}_t, \hat{y}_t}, \quad (1)$$

Where M is the median filter kernel, ω is the dense optical flow field, and $\hat{x}_t, \hat{y}_t$ is the rounded position of (x_t, y_t). Now, we can form a trajectory of a feature point as $(P_t, P_{t+1}, P_{t+2}, \ldots)$, and we constrain the length of each trajectory to be 15 to mitigate issues of drifting. Now, with the trajectory of a feature point, we can further form another sequence: $S = (\Delta P_t, \ldots, \Delta P_{t+L-1})$, where $\Delta P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. Now, we can normalize S to obtain S', i.e., the *TRAJ* descriptors

$$S' = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||}. \quad (2)$$

MBH descriptors: The algorithm would first define a space-time volume for each trajectory. This volume is further split into cell-sized spatio-temporal grid, where the parameter of the grid size is $(n_\sigma, n_\sigma, n_t) = (2, 2, 3)$. The descriptors are then computed on this cell-sized grid. Motion boundary histogram descriptors are based on derivatives of optical flow, I_ω, in order to quantify relative pixel-wise motions; the method is more robust to movement associated with camera motion. The optical flow field is first separated into each direction (I_x, I_y), where we compute spatial derivative in each direction of x, y in this cell, and the orientation information is quantized into histograms (8 bins) and then finally normalized by L^2-norm to generate the two descriptors **MBH**$_x$ and **MBH**$_y$. In this work, we further adopt the improved estimation of camera motion, i.e., based on speed up robust (SURF) features and random sample consensus (RANSAC) method. The process further removes trajectories that are caused by camera motions before the computing the descriptors [70].

In summary, the *dense unit-level* video feature extractions consists of two types of features: *TRAJ* (30 dimensions, i.e., 15 frame of a trajectory's x, y's normalized displacement information), **MBH**$_x$ and **MBH**$_y$ (each 96 dimensions, i.e., cell size $2 \times 2 \times 3 \times 8$ bins of histogram, describing the relative motion of a trajectory in the x and y direction, respectively). Video features are *dense* in terms of both in their spatial sampling and in their temporal granularity (single frame shift). The list of parameters choice is fairly standard and has been utilized in many past computer vision works (e.g., [59], [71], [72]), except that the pixel-window width is set to eight when looking for points-to-sample as compared to three to reduce the data size. This approach measures candidate principals bodily movement characteristics during impromptu speech and represent them as a high-dimensional (207 dimensions) feature vector per frame.

3.3 Behavior Profile: Session-Level Encoding

The dense unit-level audio-video feature extraction generates a p-length acoustic feature for 200 ms unit window at 100 ms step ($p = 8861$), and l-length video features at every 66 ms ($l = 207$). The labels of interest occur at the speech session-level ($\approx$ 3-minute long). Depending on the actual length of each speech, this would result in a varying numbers of feature sequences per presentation. In this work, we employ two different methods to encode these unit-level acoustic-video descriptors to generate session-level vector representation as behavior profiles: k-means bag-of-word encoding and Fisher-vector encoding. These encoding methodologies have been quite useful in processing video information in recognition tasks involving spatio-temporal movement of humans through video sequences [73]. In this work, we utilize this *holistic* behavior profile representations as inputs to the machine learning algorithm.

3.3.1 k-Means Bag-of-Word Encoding

The first encoding approach is based on k-means clustering, a.k.a., bag-of-word model. The idea is to first randomly sample feature vector sequences from the entire database to train a 'codebook' using k-means clustering approach. Once a codebook is trained, we assign each *unit* frame of the feature vectors to the closest 'code' using Euclidean distance to the mean of each cluster. Then, for a particular oral presentation, we can form a histogram of k bins with counts from the cluster assignments for that speech. After performing z-normalization on the histogram, we obtain the final session-level behavior profile, i.e., a single vector of length k, from the original sequences of unit-level acoustic-video features. This approach is different from the vector of locally aggregated descriptors (VLAD) encoding [74]. VLAD uses the summation of weighted overall distance of each unit frame to the cluster centroid to achieve session-level encoding, while BOW encoding is a counts on the number of each cluster occurrence within a session to achieve encoding.

3.3.2 Fisher-Vector Encoding

The use of Fisher-vector encoding has been shown to obtain recognition results surpassing the use of BOW in a several computer vision tasks [75]. Hence, we further employ this encoding approach on video descriptors. We briefly describe FV encoding below.

FV encoding can be derived as a special case of Fisher kernel (FK). Fisher kernel, i.e., $\mathbf{K}(\mathbf{X}, \mathbf{Y})$, is used to measure the similarity between the two sets of data samples $(\mathbf{X}, \mathbf{Y})$,

$$(\mathbf{X} = \bar{x}_t, t = 1T_1, \mathbf{Y} = \bar{y}_t, t = 1T_2),$$

where T_1, T_2 can be different. We define a scoring function,

$$\mathbf{G}_\lambda^\mathbf{X} = \nabla_\lambda \log \mathbf{u}_\lambda(\mathbf{X}), \tag{3}$$

where $\mathbf{u}_\lambda(\mathbf{X})$ denotes the likelihood of $\mathbf{X}$ given the probability distribution function (PDF), $\mathbf{u}_\lambda$. Here the choice of PDF is Gaussian Mixture Model (GMM), and λ represents the parameters of GMM, i.e., $(\bar{w}, \bar{\mu}, \Sigma)$. $\mathbf{G}_\lambda^\mathbf{X}$ is the direction where λ has to move to provide a better fit between $\mathbf{u}_\lambda$ and $\mathbf{X}$. With the use of Equation 3, we have effectively changed a varying length $\mathbf{X}$ into a fixed-length vector, i.e., a dimension equals to the total number of parameters in λ.

Since $\mathbf{u}_\lambda(\mathbf{x})$ is GMM with K mixtures expressed as,

$$\mathbf{u}_\lambda(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathbf{u_k}(\mathbf{x}),$$

with $\lambda = \{w_k, \mu_k, \Sigma_k, k = 1, \ldots, K\}$ correspond to mixture weight, mean, and covariance matrix for each mixture of Gaussian. These parameters are of the following form

$$\sum_{k=1}^{K} w_k = 1$$

$$\mathbf{u_k}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)\right)},$$

covariance matrices are set to be diagonal.

We first define a probability $\gamma_t(k)$ as

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)},$$

From this, the gradient with respect to μ_k, σ_k of a data point x_t can be derived,

$$\nabla_{\mu_k} \log \mathbf{u}_\lambda(x_t) = \gamma_t(k) \left(\frac{x_t - \mu_k}{\sigma_k^2} \right)$$

$$\nabla_{\sigma_k} \log \mathbf{u}_\lambda(x_t) = \gamma_t(k) \left(\frac{(x_t - \mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k} \right).$$

Then, by Fisher Information Matrix approximation, we can derive the Fisher encoding for the first and second order statistics below,

$$\mathbf{g}_{\mu_\mathbf{k}}^\mathbf{X} = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right) \tag{4}$$

$$\mathbf{g}_{\sigma_\mathbf{k}}^\mathbf{X} = \frac{1}{T\sqrt{2w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right). \tag{5}$$

This results in a fixed dimension vector at the session-level by concatenating the output from Equation 4 & 5, i.e.,

$$FV = \left[\mathbf{g}_{\mu_\mathbf{1}}^\mathbf{X}, \mathbf{g}_{\sigma_\mathbf{1}}^\mathbf{X}, \ldots, \mathbf{g}_{\mu_\mathbf{k}}^\mathbf{X}, \mathbf{g}_{\sigma_\mathbf{k}}^\mathbf{X}, \ldots, \mathbf{g}_{\mu_\mathbf{K}}^\mathbf{X}, \mathbf{g}_{\sigma_\mathbf{K}}^\mathbf{X} \right]. \tag{6}$$

In this work, we only compute the gradient with respect to mean and standard deviation because empirically weight possess little useful information [76] and the inclusion of weight parameter would make the dimension of FV become too large. We perform random sampling of unit-level feature vectors to train the GMM, and carry out L^2 normalization.

In summary, for audio modality profile, the encoding is carried out on just the *speaking portions* within the presentation using BOW; FV is not carried out for the audio modality due to the large acoustic LLD dimension (less suitable for GMM training). For video modality profile, both kinds of encodings are carried out on the entire presentation.

3.4 Regression Using SVM Binary Classifier

To further handle the subjectivity in developing automated assessment system for high-level attributes, we employ a novel idea to assign a regressed score to each oral presentation. The idea is inspired from the past psychological evidence and our hypothesis that when experts assess these oral presentations, they may internally have templates of *good* and *bad* presentations and by judging each sample's *closeness* to each set of the templates, they then assign a score respectively. With this idea in mind, we assess each oral presentation by utilizing *sample-to-decision boundary distance* outputted using a SVM binary classifier trained for the purpose of recognizing *good* versus *bad* performing speech along each rating dimensions of interest. Our idea intuitively corresponds to the underlying mechanism of SVM for classification. We can imagine the distance to decision boundary encodes information about how *far* or *close* is a particular speech to the support vectors, i.e., the representative set of *good* and *bad* examples that maximize the between-class margin, and we can treat this *closeness* essentially as a proxy to the subjective process of generating assessment score. The boundary of choosing *good* and *bad* speech, i.e., top and bottom rated speech, for SVM training is termed as the *distinctive cut-off boundary*. A similar concept has also been explored in utterance-level emotion classification by Mower et al. [77] and facial action unit categorization [78] though the use in regression tasks remains to be limited.

Hence, for the regression experiment, we first train a separate acoustic-only and video-only SVM classifier on the *distinctive* classes of data using its respective behavior profile. For each data, z, we then compute the distance to the decision hyperplane using the trained SVM as follows:

$$dist = \sum_{i=1}^{n} y_i \alpha_i x_i^T z + \rho, \tag{7}$$

where $y_i \in \{1, -1\}$ corresponds to the class label of each support vector, α_i is the weight parameter for each support vector x_i, and ρ is the bias term. After generating $dist$ for each modality $(dist_A, dist_V)$, we normalize the score by linearly transforming each of them to a range of $[1, 10]$ such that both modalities scores are comparable. Finally, we assign a final score to each of the rating dimensions mentioned in Section 2.2 by summation of these two distance-based measures. Fig. 5 shows our proposed framework.

4 Impromptu Speech Scoring Experiment

In this section, we present our experimental results on automatically scoring the ten dimensions-of-interest using the proposed method mentioned in Section 3.4. The evaluation is done via leave-one-speaker-out cross validation, and the evaluation metric is spearman correlation.

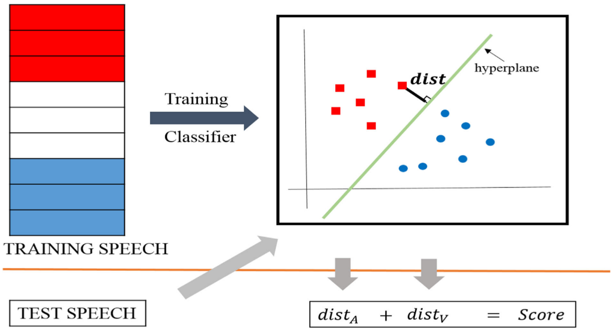

Fig. 5. It shows a schematic of our proposed regression using SVM binary classifier. At training, only the *distinctive* top-bottom scored presentations are used to train a SVM binary classifier (i.e., samples marked in red and blue), and at testing, a speech is scored by computing the distance to the learned decision hyperplane. The multimodal fusion is a simple averaging between the two distances to generate a final score.

4.1 Experimental Setup

Our proposed method, i.e., utilizing distance-to-decision hyperplane outputted from binary SVM classifier trained on the *distinctive* set of samples as the final score, is denoted as $\text{Binary}_{\text{SVM}}$. We compare our method to two other baseline scoring frameworks. One of them is the support vector regression SVR (denoted as $\text{Baseline}_{\text{SVR}}$). Another one is also based on the same idea as our proposed method, but instead of training a SVM classifier, we directly use SVR regression (denoted as $\text{Binary}_{\text{SVR}}$) to regress the score. The SVR model for both baseline models is trained using $\epsilon = 0.2$.

Moreover, the parameter K determines the dimensions of profile using BOW, and M indicates the number of mixtures used in the GMM for FV encoding. These two parameters (K, M) dictate the dimensionality of the final feature inputed to the classifiers. For BOW encoding, we have tested $K \in \{1000, 2000, 3000\}$; for FV encoding, we have tested $M \in \{128, 256, 512\}$. We present the results of $K = 2000$ for BOW and $M = 256$ for FV encodings respectively in this work.

4.2 Experimental Results and Analyses

Table 4 summarizes our experimental results using different feature sets. The following is the description for each feature set, A_* and V_* used:

A_1: the same feature set used in our previous paper [38], i.e., computing four statistical functions (mean, variance, kurtosis, and skewness) over utterance-level functional features to form a speech-level feature vector; the utterance-level features are derived from the opensmile 2010 configuration

A_2: BOW encoding on the dense unit-level acoustic low-level descriptors (Section 3.1)

V_3: BOW encoding on motion boundary descriptors (MBH_{xy})

V_4: BOW encoding on the two descriptors (Traj and MBH_{xy})

V_6: FV encoding on motion boundary descriptors (MBH_{xy})

V_7: FV encoding on the two descriptors (Traj and MBH_{xy})

4.2.1 Accuracy Validation of Automatic Scoring

First thing to note that, in general, our proposed method, i.e., BN_{SVM^*}, achieve the best average spearman correlation computed across all ten dimensions of interest. In specifics, comparing to the best baseline SVR model, our best model,

TABLE 4
Exp Results: The Metric Is Spearman Correlation

		Dim_{1o}	Dim_{2o}	Dim_{3o}	Dim_{4o}	Dim_{5o}	Dim_{1r}	Dim_{2r}	Dim_{3r}	Dim_{4r}	Dim_{5r}	AVG
Baseline$_{SVR}$	$BL_{SVR0} : A_1$	0.13	0.02	0.03	0.27	0.11	0.01	0.13	0.15	0.18	0.05	0.11
	$BL_{SVR1} : A_2$	0.36	0.27	0.31	0.52	0.39	0.37	0.25	0.35	0.48	0.52	0.38
	$BL_{SVR2} : V_3$	0.44	0.47	0.43	0.49	0.49	0.40	0.32	0.30	0.40	0.45	0.42
	$BL_{SVR3} : V_4$	0.42	0.42	0.42	0.49	0.48	0.39	0.31	0.34	0.40	0.44	0.41
	$BL_{SVR4} : V_6$	0.41	0.44	0.47	0.48	0.39	0.42	0.35	0.35	0.44	0.47	0.42
	$BL_{SVR5} : V_7$	0.50	0.43	0.47	0.48	0.52	0.38	0.31	0.34	0.42	0.45	0.43
	$BL_{SVR6} : A_2 + V_3$	0.48	$\underline{\mathbf{0.50}}$	0.47	0.59	0.54	0.46	0.36	0.41	**0.54**	0.59	0.49
	$BL_{SVR7} : A_2 + V_4$	0.49	0.45	0.47	$\underline{\mathbf{0.60}}$	0.55	0.46	0.34	0.44	**0.54**	0.59	0.49
	$BL_{SVR8} : A_2 + V_6$	0.46	0.46	$\underline{\mathbf{0.51}}$	0.57	0.48	**0.48**	**0.36**	**0.45**	**0.54**	0.60	0.49
	$BL_{SVR9} : A_2 + V_7$	**0.53**	0.45	$\underline{\mathbf{0.51}}$	0.59	**0.57**	0.46	0.32	0.44	**0.54**	**0.61**	**0.50**
Binary$_{SVR}$	$BN_{SVR0} : A_1$	0.19	0.07	0.19	0.31	0.24	0.01	0.15	0.16	0.18	0.18	0.17
	$BN_{SVR1} : A_2$	0.46	0.31	0.31	0.52	0.44	0.40	0.25	0.36	0.48	0.54	0.41
	$BN_{SVR2} : V_3$	0.45	0.47	0.43	0.49	0.53	0.41	0.32	0.33	0.40	0.48	0.43
	$BN_{SVR3} : V_4$	0.45	0.44	0.44	0.49	0.53	0.38	0.31	0.39	0.40	0.44	0.43
	$BN_{SVR4} : V_6$	0.36	0.43	0.47	0.48	0.39	0.40	0.35	0.35	0.44	0.47	0.41
	$BN_{SVR5} : V_7$	0.48	0.41	0.47	0.48	0.49	0.37	0.30	0.35	0.42	0.43	0.42
	$BN_{SVR6} : A_2 + V_3$	0.55	0.50	0.47	0.59	0.59	0.48	**0.36**	0.45	**0.54**	**0.62**	**0.51**
	$BN_{SVR7} : A_2 + V_4$	0.54	0.45	0.48	$\underline{\mathbf{0.60}}$	**0.59**	**0.49**	0.34	**0.48**	**0.54**	0.61	**0.51**
	$BN_{SVR8} : A_2 + V_6$	0.49	**0.47**	$\underline{\mathbf{0.51}}$	0.57	0.48	0.48	**0.36**	0.46	**0.54**	0.60	0.50
	$BN_{SVR9} : A_2 + V_7$	**0.57**	0.46	$\underline{\mathbf{0.51}}$	0.59	0.57	0.48	0.33	0.46	**0.54**	0.61	**0.51**
Binary$_{SVM}$	$BN_{SVM0} : A_1$	0.20	0.19	0.21	0.27	0.26	0.07	0.14	0.23	0.23	0.16	0.19
	$BN_{SVM1} : A_2$	0.39	0.29	0.33	0.50	0.42	0.47	0.33	0.39	0.56	0.60	0.43
	$BN_{SVM2} : V_3$	0.51	0.46	0.40	0.43	0.53	0.40	0.38	0.43	0.38	0.44	0.44
	$BN_{SVM3} : V_4$	0.54	0.43	0.45	0.46	0.56	0.38	0.37	0.46	0.38	0.44	0.45
	$BN_{SVM4} : V_6$	0.44	0.42	0.42	0.45	0.46	0.39	0.40	0.43	0.45	0.46	0.43
	$BN_{SVM5} : V_7$	0.48	0.41	0.45	0.46	0.52	0.39	0.36	0.41	0.39	0.41	0.43
	$BN_{SVM6} : A_2 + V_3$	0.58	**0.47**	0.44	0.56	0.59	0.57	$\underline{\mathbf{0.47}}$	0.50	0.57	0.65	0.54
	$BN_{SVM7} : A_2 + V_4$	$\underline{\mathbf{0.61}}$	0.46	**0.48**	**0.58**	$\underline{\mathbf{0.62}}$	$\underline{\mathbf{0.58}}$	0.46	$\underline{\mathbf{0.52}}$	0.56	$\underline{\mathbf{0.66}}$	$\underline{\mathbf{0.55}}$
	$BN_{SVM8} : A_2 + V_6$	0.53	0.44	0.47	0.54	0.54	0.55	0.46	0.51	$\underline{\mathbf{0.58}}$	0.65	0.53
	$BN_{SVM9} : A_2 + V_7$	0.56	0.46	**0.48**	0.57	0.60	0.56	0.46	0.51	0.57	0.65	0.54

A_1 denotes the baseline method, i.e., functionals of functionals. A_2 denotes dense unit-level acoustic feature with BOW encoding, $V_{3,4}$ denote BOW encoding on MBH_{xy} and All descriptors respectively, and $V_{6,7}$ denote FV encoding on MBH_{xy} and All descriptors respectively. Binary$_{SVR}$ indicates the method using SVR on best-distinctive set of samples, and Binary$_{SVM}$ is our proposed method that uses sample-to-hyperplane distance. The best accuracy obtained of each framework on each rating dimension is in bold, and the final best accuracy achieved across frameworks is additionally marked with underline (all of the results obtain a p-value $< 1e^{-03}$).

BN_{SVM7}, improves the average correlation by 0.049 absolute (9.8 percent relative). The result shows that both $Binary_{SVR}$ and espcially $Binary_{SVM}$ outperform straightforward $Baseline_{SVR}$ demonstrating that by learning from less-ambiguous data could provide a boost in performance. Additionally, we see that methods based on $Binary_{SVM}$ achieve a better overall accuracy as compared to methods based on $Binary_{SVR}$. Since $Binary_{SVM}$ is essentially a classification approach without optimizing directly to regress on the values of the scores, it is quite interesting to observe that simply by utilizing distance value as the predicted scores would achieve a better regression performance (number in bold). This may reinforces our hypothesis (Section 3.4) about the possible mechanism in the expert's subjective assessment in this context, i.e., not as a process of assigning a real-value score, instead, comparing to good/bad templates.

Moreover, the improvement in using dense unit-level profiles on low-level acoustic descriptors is evident by comparing accuracies obtained using A2 versus A1 feature sets. The differences in the use of BOW versus FV encodings on video descriptors, however, are less significant in this context. The reason we believe could have been that both profile encoding methodologies are equally powerful to in capturing discriminative characteristics of the speaker's bodily movement during impromptu speech. Further, the multimodal fusion of audio and video information improves the recognition for all ten dimensions. In specific, the multimodal BN_{SVM6} improves the correlation of 0.123 absolute (28.8 percent relative) and 0.105 (23.6 percent relative) over audio-only BN_{SVM2} and video-only BN_{SVM4}.

Lastly, we observe a very interesting and important result. The rating that, for most part, achieves the best accuracies is the Dim_{5^*} (the total score), i.e., 0.62 and 0.66 for original and rank-normalized total scores, respectively. First of all, within the application context of automatic oral presentation assessment for the candidate principals certification program, this result is quite promising and in fact useful. Dim_5 is the final score entered into the grading system for educational decision-maker in order to assess the qualification of these candidate principals. Furthermore, this result underscores the *holistic* modeling nature of our proposed multimodal low-level behavior profile framework; Dim_5 is a higher-level rating compared to all other dimensions of ratings, which are focused more on a specific aspect, used in this assessment. Further, while the items on the scoring sheet seem to be disjoint with each focusing on disparate dimensions from their written descriptions, in Section 2.2, we see that the correlation among dimensions are actually quite strong implicating that the assessment of individual rating are strongly affected by each other.

TABLE 5
Summary Information About the Best-Distinctive Split of Data Samples for the Ten Dimensions: Agreement$_{bb}$ Is the Coaching Principals' Agreement Level Computed for This Set of Samples, Agreement$_{ori}$ Is the Coaching Principals' Agreement Level for the Entire Database, and Usage of Data Indicates the Percentage of Samples Included in the Training of the Binary$_{SVM}$

	Agreement$_{\mathrm{bb}}$	Agreement$_{\mathrm{ori}}$	Usage of Data (%)
Dim$_{1o}$	0.53.	0.51	85.4
Dim$_{2o}$	0.61	0.48	61.8
Dim$_{3o}$	0.51	0.40	65.0
Dim$_{4o}$	0.53	0.53	100.0
Dim$_{5o}$	0.62	0.58	81.3
Dim$_{1r}$	0.68	0.55	73.2
Dim$_{2r}$	0.46	0.40	82.9
Dim$_{3r}$	0.72	0.43	38.2
Dim$_{4r}$	0.43	0.58	64.2
Dim$_{5r}$	0.75	0.63	56.9
AVG	0.59	0.41	66.0

Our computational framework can be thought as a quantitative *holistic* modeling on the multimodal delivery form of the presentation, which is integrative and influential in the overall experts *judgment's* of how well a candidate principal has carried out the speech.

4.2.2 Analysis of Best-Distinctive Cut-Off Boundary

Choosing the best distinctive boundary, i.e., top X% and bottom Y% rated speech, to train the binary SVM classifier is a major component of our proposed algorithm. Table 5 depicts the final total usage of top and bottom percentage of scoring samples used for each dimension in our framework with their associated inter-evaluator agreement computed for those particular samples. The percentage presented is chosen based on a greedy search.

The thing to note is that the best cut-off boundary seems to come at a trade-off between the amount of data used and the amount of ambiguity to be included. The more data samples included in the training of SVM classifier does not necessary correspond to an increase in the accuracy as the inter-evaluator agreement also tends to decrease; at the same time, too few data samples included cause the model to be not well-trained. Depending on the dimension of interest, the best total percentage of data used ranges from approximately 40 to 85 percent. We observe that the best-distinctive set of data samples used for predicting rank-normalized dimensions tend to be less when compared to the amount of data required to achieve the best accuracy for original labels. This phenomenon also reflects in the agreement level as the portion of data used for rank-normalized dimensions tend to have a higher inter-evaluator agreement ($\mu = 0.61$) when compared to the portion of data used for original labels ($\mu = 0.56$).

4.2.3 Comparison with Other Techniques

Aside from using support vector machine as the binary classifier in order to derive confidence score for the regression tasks, we further compare it with three other classifier approaches, i.e., logistic regression (BN$_{\mathrm{Logistics}}$), random forest (BN$_{\mathrm{RandForest}}$, and gradient boosted decision tree (BN$_{\mathrm{GBDT}}$), using the best combination of audio and video descriptors from Table 4. The confidence score for each of the three classifiers are derived as probabilities, and further, the best boundary chosen in Section 4.2.2 is also used for these classifiers. Our comparison results are summarized in Table 6. Our results show that using SVM works better than the other three methods, most likely, is due to the fact of its maximum-margin learning in discriminating between the *good* versus *bad* samples. In this work, we do not explicitly compare to time series model. On one hand, the usage of static classifier with high-dimensional encoding of low-level descriptors have obtained state-of-art recognition of events with temporal structure in many audio and video recognition tasks; at the same time, most conventional time-series models, e.g., Hidden Markov Model and its variants, are also less suitable to handle modeling tasks with high-dimensional input feature space.

TABLE 6
Comparison with Three Other Classification Methodologies: BN$_{\mathrm{Logistics}}$, BN$_{\mathrm{RandForest}}$ and BN$_{\mathrm{GBDT}}$ Indicate the Use of Logistic Regression, Random Forest, and Gradient Boosted Decision Tree, Respectively

	BN$_{\mathrm{SVM}}$	BN$_{\mathrm{Logistics}}$	BN$_{\mathrm{RandForest}}$	BN$_{\mathrm{GBDT}}$
Dim$_{1o}$	**0.61**	0.54	0.32	0.23
Dim$_{2o}$	**0.46**	0.39	0.27	0.36
Dim$_{3o}$	**0.48**	0.42	0.22	0.25
Dim$_{4o}$	**0.58**	0.57	0.39	0.36
Dim$_{5o}$	**0.62**	0.58	0.34	0.26
Dim$_{1r}$	**0.58**	0.54	0.31	0.28
Dim$_{2r}$	**0.46**	0.42	0.26	0.34
Dim$_{3r}$	**0.52**	0.46	0.28	0.23
Dim$_{4r}$	**0.56**	0.54	0.41	0.52
Dim$_{5r}$	**0.66**	0.58	0.31	0.29
AVG	**0.55**	0.51	0.31	0.31

4.2.4 Consistency Validation of Automatic Scoring

Lastly, we carry out a novel validation analysis in this work. Aside from comparing correlation to the *ground truth*, i.e., the average of the coaching principals' ratings, as conventionally done to evaluate the accuracy numbers (e.g., Section 4.2.1), we perform additional validity analysis. We reach out to a pair of coaching principals, who were the coaching principals during the 2014 certification program, to rate 10 oral presentations post-hoc again without letting them know that they have already seen/graded those speech back during the 2014 certification program. We then compute pair-wise spearman correlations between 'original scores: Ori.' (the original scores collected at the certification program), 'predicted scores: Pred.' (scores derived from BN$_{\mathrm{SVM7}}$), and 'new scores: New' (the re-graded scores) across the ten dimensions-of-interest.

Table 7 summarizes the analysis results, and the number with a star indicates that particular correlation is significant ($\alpha = 0.05$). In general, we observe that while these are the same two coaching principals rating the same ten oral presentations just at two different points in time, both their ratings (i.e., original and new) on average correlate with the automatically-derived scores from audio-video more than among themselves. This result is quite intriguing. Human expert can potentially suffer variabilities from undesirable idiosyncratic factors and environment contexts naturally,

TABLE 7
Summary Results in Section 4.2.4

	Ori. versus New	Pred. versus Ori.	Pred. versus New
Dim$_{1o}$	0.09	0.57	0.69*
Dim$_{2o}$	0.46	0.70*	0.74*
Dim$_{3o}$	0.54	0.37	0.50
Dim$_{4o}$	0.34	0.03	0.16
Dim$_{5o}$	0.29	**0.64***	**0.76***
Dim$_{1r}$	0.08	0.41	0.56
Dim$_{2r}$	0.47	0.52	0.74*
Dim$_{3r}$	0.55	0.21	0.25
Dim$_{4r}$	0.26	0.48	0.53
Dim$_{5r}$	0.53	**0.65***	**0.61***
AVG	0.40	0.50	0.55

*'Ori.' indicates the original scores, 'New' indicates the same samples graded again by the same raters, and 'Pred.' denotes the predicted scores. The pairwise correlation is computed using spearman correlation. The number with a * indicates significance level at 0.05*

e.g., grading after watching video tapes versus on-site grading, number of grading needed to be done for a given time, or simply due to changes in one's grading standard over time, etc. Our signal-based assessment score never learns from the re-graded labels but remains robust and reliable across time; for example, the correlation obtained when comparing to experts rating at two different time points for Dim$_{5r}$ remains fairly consistent at 0.65 and 0.61, where the two experts only correlate with each other at 0.53. Our proposed automatic framework can be more reliable and consistent in modeling the quality of an impromptu speech than human experts.

Furthermore, we also observe the best result still comes from Dim$_{5*}$, which reiterates the holistic modeling nature of our profile-based recognition framework. While we only analyze 10 samples, to the best of our knowledge, this is one of the first works that have analyzed the consistency across time in the process of developing recognition framework for subjective human attributes. From Tables 4 and 7, our analyses indicate both the reliability and consistency, i.e., two major metrics in testing the validity of meaningful psychological construct, of our computational framework in assessing candidate principals' oral presentation skills.

4.3 Discussions

In Section 4.2, we present various results in assessing and analyzing our proposed multimodal computational framework toward developing a novel automated impromptu speech rating system using fusion of audio-visual information designed for the NAER's yearly pre-service principals' certification program. A summary of novel results is listed below:

- *Accuracy*: using sample-to-decision boundary distance obtained from the SVM binary classifier by training on distinctive subset of the database achieve improved scoring correlations compared to regression approaches
- *Accuracy*: fusing multimodal behavior profiles outperform single-modality modeling in this context
- *Accuracy*: Dim$_5$ (the final total score) is the dimension that achieves the best results (0.62 & 0.66). It is useful considering the context of the application and reinforcing the *holistic* modeling nature of our framework
- *Consistency*: our automatic scoring system is shown to be possibly more consistent than human experts by demonstrating that the ratings from the same two experts are correlated more to our proposed automatic scoring system than among themselves across two different time points. The most consistent rating that our framework obtains is still Dim$_5$ (the total score)
- *Analysis*: the best distinctive subset of training samples often correspond to the samples with less ambiguity (higher inter-evaluator agreement).

4.3.1 Comparison to Existing Works

There are a couple points to make when comparing approaches of existing works. First, while we do not explicitly compare accuracy to the approach of pre-defining a set of discrete *interpretable* behaviors to be used in the overall assessment, we can compare with those works that are in the similar context. We observe that authors of existing works (i.e., [55], [56]) engineered multimodal features that are more interpretable, e.g., speaking rate and syntactic structure derived from manual transcript, hand and body movement extracted from Kinect, and head orientation, and they, however, obtained an correlation using their automatic holistic assessment to the experts rating approximately in the range of 0.44. Actually, a most recent work done on that particular dataset [57] demonstrated a significant improvement can be obtained by using encoding-based approach on low-level descriptors computed from the speaker's face, which is an approach that is closer in concept to our profile-based approach presented in this work.

Second, authors of work [54] first pre-defined and manually annotated a finite set of *desirable* behaviors of a speaker during public speaking that could be automatically extracted using ensemble trees. The method is quite promising and indeed correlates highly with the overall judgment on the quality of the oral presentation in their corpus. Their purpose was to train an individual's to be good at certain behaviors during public speaking; however it may be difficult to scale up to large-scale assessment when the possible number of discrete behavior types and descriptions can be vast and variable, and further simply having these behaviors annotated first by humans are not feasible.

Although our setups and aims are not directly comparable, we see that our proposed profile-based representation, i.e., directly compute behavior representation from LLDs, possess promising modeling power of a speaker's communicative behavior and classifier-as-regression approach, i.e., handling the ambiguity in the data and subjectivity in the expert labeling, together can achieve reliable and consistent assessment accuracy of oral presentations. Nonetheless, having interpretable high-level behaviors can still be beneficial when moving from assessment to training. In Fig. 6, we plot a truncated (only 150 clusters) summative acoustic behavior profiles (generated with BOW encoding) from top five-rated impromptu speech in the NAER database. While the interpretability of each acoustic individual behavior cluster can be hard to assess immediately, the plot shows that, indeed, certain behavior clusters in the speech modality occur a lot more times than the other clusters in these top-rated speech. Instead of pre-designing a set of behaviors to look for, these data-driven behavior clusters may also

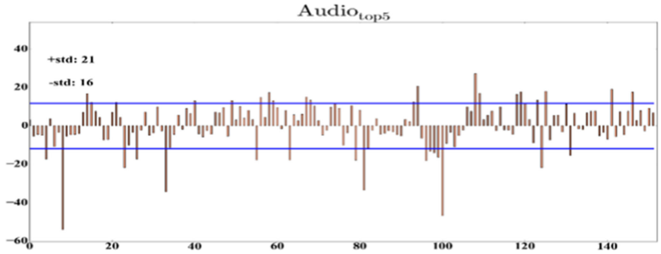

Fig. 6. It shows a summation of truncated (only 150 clusters) of speech behavior profiles generated from the top 5 rated candidate principals.

help in identifying meaningful high-level behaviors that would occur in the well-orchestrated presentation while maintaining robust recognition rates. Future analyses on what those clusters in both audio and video modalities may mean perceptually is beyond the scope of this work, but it is definitely an important direction to explore.

5 Conclusions

Effective leadership bears strong relationship to fundamental attributes, such as emotion contagion, positive mood, and social intelligence. These attributes are also reflected in an individual's communicative behaviors, especially in settings of public speaking. While theoretical conceptualization of leadership has received much attention, little has progressed in terms of quantitative measurement and modeling of these behaviors. In this work, we present a thorough BSP research in the development toward automatic assessment on the qualities of oral presentation in the domain of education, specifically in the real context of impromptu speech assessment during pre-service candidate principals certification program. We propose the use and demonstrate the effectiveness of our *holistic* low-level multimodal behavior profile techniques in automatically scoring these oral presentations. Also, handling the subjectivity in these ratings computationally by utilizing sample-to-decision boundary distance trained on the distinctive subset as the regressed scores further obtains improved and competitive correlations to the expert coaching principals' ratings. The validity is further strengthened by demonstrating the framework's ability to maintain its *reliability* more *consistently* across time as compared to human experts. The unqiue *contextualized* and *in-the-wild* corpus collected in this work will be presented publicly to the community after proper IRB approval; the current work will also help in providing a set of benchmark recognition accuracies obtained on this corpus. There are multiple directions of future works along technical, educational, and scientific directions.

On the technical side, we will continuously work toward improving the overall system's accuracy. There are several directions. One of the immediate direction is the inclusion of additional modality, i.e., the lexical content. The presented multimodal behavior profile can be thought as a quantitative model more on the *delivery* of the speech, where the inclusion of word usage (lexical) information will then be more on the aspect of *content*. Past literatures have indicated that an charismatic leader is in fact proficient in integrating both attributes in order to achieve motivating and emotionally-contagious speech. Further, as we collect more behavior data, e.g., there is an additional 200 impromptu speech recorded in 2015 pre-service principals' certification program, we imagine the state-of-art feature encoding, i.e., generation of behavior profiles, approaches based on deep learning, and/or the time series-based modeling, e.g., long-short term memory neural network, could then have the potential to further improve the robustness of algorithms. Lastly, we will also continue to understand the possible meaning of these data-driven behavior profiles in details and cross-referencing to the known manually pre-defined discrete behaviors. At the same time, it would be interesting to observe what additional behaviors may have been captured in this holistic representations that are absent before.

Aside from the technical directions, since it is a collaborative research effort with the NAER researchers with targeted application in real life, another line of immediate future works is to realize this engineering system in the context of education. The current status quo in the grading of candidate principals (not limited to oral presentations but also other aspects within the certification program) is completely based on a limited number of expert coaching principals. We will first start exploring the possibility of using this system as an additional scores to supplement the current grading structure. At the same time, we will investigate other *outcome* variables, e.g., other instrumented written-tests assessment of these principals within this certification program, to understand quantitatively the validity and the relationship between each principal's communicative behaviors and these other assessment scores. With a focused emphasis on understanding expressive behaviors in this context quantitatively, we may be able to design a better-suited certification training program not only for the candidate principals but also for other educational professionals at scale in the future, e.g., 20,000 per-service teachers going through assessment program in Taiwan every year.

Lastly, many of the low-level multimodal behavior descriptors used in this work have also been shown to be effective in tasks of affective computing-indicating that these behavior representations encompass a wide range of information about various internal states of minds in humans. The automatic framework that we propose in this work can serve as an initial computational building block for further scientific study. One of the directions is to tease apart the aspect of an individual's communicative behaviors that are either emotionally-expressive or emotionally-contagious when giving a public speech. Instead of just annotating emotional states and/or effects of these impromptu speech for further studies, since these are real school principals that are already in office for some time, we plan to investigate their actual leadership effectivenesses and charismatic impressions (i.e., related to emotion contagion) in steering their schools in real life in relation to the communicative behaviors exhibited during public speaking in a longitudinal time span (i.e., from the pre-service certification program, on-boarding speech, to regular public addresses). With the availability of computational methods and the tight integrative collaboration with relevant domain experts, we hope to continue this BSP research to bring in additional scientific understanding to substantiate leadership theories and affect-perception of charisma with quantitative methods that model real-world behaviors.

ACKNOWLEDGMENTS

Thanks to Ministry of Science and Technology, National Academy for Educational Research, and Novatek Fellowship for funding.

REFERENCES

[1] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administ. Sci. Quart.*, vol. 47, no. 4, pp. 644–675, 2002.

[2] A. Erez, V. F. Misangyi, D. E. Johnson, M. A. LePine, and K. C. Halverson, "Stirring the hearts of followers: Charismatic leadership as the transferal of affect," *J. Appl. Psychology*, vol. 93, no. 3, 2008, Art. no. 602.

[3] J. E. Bono and R. Ilies, "Charisma, positive emotions and mood contagion," *Leadership Quart.*, vol. 17, no. 4, pp. 317–334, 2006.

[4] C. M. Brotheridge, R. T. Lee, R. E. Riggio, and R. J. Reichard, "The emotional and social intelligences of effective leadership: An emotional and social skill approach," *J. Managerial Psychology*, vol. 23, no. 2, pp. 169–185, 2008.

[5] J. Gooty, S. Connelly, J. Griffith, and A. Gupta, "Leadership, affect and emotions: A state of the science review," *Leadership Quart.*, vol. 21, no. 6, pp. 979–1004, 2010.

[6] K. J. Levine, R. A. Muenchen, and A. M. Brooks, "Measuring transformational and charismatic leadership: Why isn't charisma measured?" *Commun. Monographs*, vol. 77, no. 4, pp. 576–591, 2010.

[7] C. Baccarani and A. Bonfanti, "Effective public speaking: A conceptual framework in the corporate-communication field," *Corporate Commun.: Int. J.*, vol. 20, no. 3, pp. 375–390, 2015.

[8] R. E. De Vries, A. Bakker-Pieper, and W. Oostenveld, "Leadership= communication? the relations of leaders communication styles with leadership styles, knowledge sharing and leadership outcomes," *J. Bus. Psychology*, vol. 25, no. 3, pp. 367–380, 2010.

[9] S. J. Holladay and W. T. Coombs, "Speaking of visions and visions being spoken an exploration of the effects of content and delivery on perceptions of leader charisma," *Manage. Commun. Quart.*, vol. 8, no. 2, pp. 165–189, 1994.

[10] S. A. Kirkpatrick and E. A. Locke, "Direct and indirect effects of three core charismatic leadership components on performance and attitudes," *J. Appl. Psychology*, vol. 81, no. 1, 1996, Art. no. 36.

[11] R. Awamleh and W. L. Gardner, "Perceptions of leader charisma and effectiveness: The effects of vision content, delivery, and organizational performance," *Leadership Quart.*, vol. 10, no. 3, pp. 345–373, 1999.

[12] K. S. Groves, "Leader emotional expressivity, visionary leadership, and organizational change," *Leadership Org. Develop. J.*, vol. 27, no. 7, pp. 566–583, 2006.

[13] M. C. Bligh, J. C. Kohles, and J. R. Meindl, "Charisma under crisis: Presidential leadership, rhetoric, and media responses before and after the September 11th terrorist attacks," *Leadership Quart.*, vol. 15, no. 2, pp. 211–239, 2004.

[14] M. Koppensteiner, P. Stephan, and J. P. M. Jäschke, "More than words: Judgments of politicians and the role of different communication channels," *J. Res. Personality*, vol. 58, pp. 21–30, 2015.

[15] D. J. Barrett, "Strong communication skills a must for today's leaders," *Handbook Bus. Strategy*, vol. 7, no. 1, pp. 385–390, 2006.

[16] R. W. Picard and R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT press Cambridge, 1997, vol. 252.

[17] A. Vinciarelli, et al., "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 69–87, Jan.–Mar. 2012.

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

[19] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[20] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recog.*, vol. 44, no. 3, pp. 572–587, 2011.

[21] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affective Comput.*, vol. 4, no. 4, pp. 341–359, Oct.–Dec. 2013.

[22] E. Crane and M. Gross, "Motion capture and emotion: Affect detection in whole body movement," in *Affective Computing and Intelligent Interaction*. Berlin, Germany: Springer, 2007, pp. 95–101.

[23] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[24] C. Busso, et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211.

[25] H. Gunes, M. Piccardi, and M. Pantic, *From the Lab to the Real World, Affect Recognition Using Multiple Cues and Modalities*. Rijeka, Croatia: InTech Education and Publishing, 2008.

[26] B. Schuller, et al., "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Comput. Speech Language*, vol. 27, no. 1, pp. 4–39, 2013.

[27] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors," *Comput. Speech Language*, vol. 28, no. 2, pp. 375–391, 2014.

[28] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Auton. Agents Multi-Agent Syst.*, vol. 20, no. 1, pp. 70–84, 2010.

[29] J. H. Jeon, R. Xia, and Y. Liu, "Level of interest sensing in spoken dialog using decision-level fusion of acoustic and lexical evidence," *Comput. Speech Language*, vol. 28, no. 2, pp. 420–433, 2014.

[30] M. P. Black, et al., "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Commun.*, vol. 55, no. 1, pp. 1–21, 2013.

[31] C.-C. Lee, et al., "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Comput. Speech Language*, vol. 28, no. 2, pp. 518–539, 2014.

[32] B. Xiao, et al., "Modeling therapist empathy through prosody in drug addiction counseling," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 213–217.

[33] Z. E. Imel, et al., "The association of therapist empathy and synchrony in vocally encoded arousal," *J. Counseling Psychology*, vol. 61, no. 1, pp. 146–153, 2014.

[34] D. Bone, et al., "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech Language Hearing Res.*, vol. 57, no. 4, pp. 1162–1177, 2014.

[35] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.

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

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

[38] S.-W. Hsiao, H.-C. Sun, M.-C. Hsieh, M.-H. Tsai, H.-C. Lin, and C.-C. Lee, "A multimodal approach for automatic assessment of school prinicipals' oral presentation during pre-service training program," in *Proc. Int. Speech Commun. Assoc.*, 2015, pp. 2529–2533.

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

[40] 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.

[41] D. L. Keith, "Principal desirabilitiy for professional development," *Academy Educational Leadership J.*, vol. 15, no. 2, 2011, Art. no. 95.

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

[43] L. Streeter, J. Bernstein, P. Foltz, and D. DeLand, "Pearsons automated scoring of writing, speaking, and mathematics," vol. 25, p. 2013, 2011.

[44] B. Topol, J. Olson, and E. Roeber, "The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments," *Stanford, CA: Stanford Center Opportunity Policy Edu. Retrieved Aug.*, vol. 2, 2010, Art. no. 2010.

[45] J. Balogh, J. Bernstein, J. Cheng, and B. Townshend, "Automatic evaluation of reading accuracy: Assessing machine scores," in *Proc. ISCA Tutorial Res. Workshop Speech Language Technol. Edu.*, 2007, pp. 112–115.

[46] J. Bernstein, M. Suzuki, J. Cheng, and U. Pado, "Evaluating diglossic aspects of an automated test of spoken modern standard arabic," in *Proc. ISCA Tutorial Res. Workshop Speech Language Technol. Edu.*, 2009, pp. 17–20.

[47] J. Bernstein, A. Van Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[48] J. Bernstein and J. Cheng, "Logic and validation of fully automatic spoken english test," The path of speech technologies in computer assisted language learning: From research toward practice, pp. 174–194, 2007.

[49] M. Worsley, "Multimodal learning analytics: Enabling the future of learning through multimodal data analysis and interfaces," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, 2012, pp. 353–356.

[50] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz, "MLA'14: Third multimodal learning analytics workshop and grand challenges," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 531–532.

[51] X. Ochoa, M. Worsley, N. Weibel, and S. Oviatt, "Multimodal learning analytics data challenges," in *Proc. 6th Int. Conf. Learn. Analytics Knowl.*, 2016, pp. 498–499.

[52] F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Presentation quality assessment using acoustic information and hand movements," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2016, pp. 2812–2816.

[53] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero-towards a multimodal virtual audience platform for public speaking training," in *Proc. Int. Workshop Intell. Virtual Agents*, 2013, pp. 116–128.

[54] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer, "Multimodal public speaking performance assessment," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 43–50.

[55] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee, "Towards automated assessment of public speaking skills using multimodal cues," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 200–203.

[56] L. Chen, C. W. Leong, G. Feng, C. M. Lee, and S. Somasundaran, "Utilizing multimodal cues to automatically evaluate public speaking performance," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 394–400.

[57] V. Ramanarayanan, L. Chen, C. W. Leong, G. Feng, and D. Suendermann-Oeft , "An analysis of time-aggregated and time-series features for scoring different aspects of multimodal presentation data," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1373–1377.

[58] A. Tamrakar, et al., "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3681–3688.

[59] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3169–3176.

[60] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 702–707.

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

[62] R. Hertwig and P. M. Todd, "More is not always better: The benefits of cognitive limits," *Thinking: Psychological Perspectives Reasoning Judgment Decision Making*, pp. 213–231, 2003.

[63] R. M. Hogarth and N. Karelaia, "Ignoring information in binary choice with continuous variables: When is less more?" *J. Math. Psychology*, vol. 49, no. 2, pp. 115–124, 2005.

[64] D. Zhu, B. Ma, and H. Li, "Joint map adaptation of feature transformation and gaussian mixture model for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2009, pp. 4045–4048.

[65] H. D. Kim, C. Zhai, and J. Han, "Aggregation of multiple judgments for evaluating ordered lists," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2010, pp. 166–178.

[66] H. Kaya, A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proc. Int. Speech Commun. Assoc.*, 2015, pp. 909–913.

[67] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. Int. Speech Commun. Assoc*, 2016, pp. 495–499.

[68] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[69] P. M. Müller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, "Emotion recognition from embedded bodily expressions and speech during dyadic interactions," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 663–669.

[70] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3551–3558.

[71] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1817–1824.

[72] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2593–2600.

[73] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understanding*, vol. 150, pp. 109–125, 2016.

[74] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[75] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. British Mach. Vis. Conf.*, 2011, vol. 2, no. 4, pp. 8–19.

[76] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.

[77] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[78] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 94–101.

Shan-Wen Hsiao (S'15) received the BS degree in electrical engineering from the National Central University (NCU), Taiwan, in 2014. He is working toward the MS degree in the Electrical Engineering Department, the National Tsing Hua University (NTHU), Taiwan. He had Presidential Award: Top 5 percent GPA of the class in 2014. His research interests include human-centered behavioral signal processing (BSP), machine learning, and multimodal signal processing. He was awarded with the NOVATEK Fellowship (2015-2016). He is a student member of the IEEE.

Hung-Chin Sun (S'15) received the BS degree in electrical engineering from the National Chung Cheng University (NCCU), Taiwan, in 2014. He is working toward the MS degree in the Electrical Engineering Department, the National Tsing Hua University (NTHU), Taiwan. His research interests include human-centered behavioral signal processing (BSP), machine learning, and multimodal signal processing. He is a student member of the IEEE.

Ming-Chuan Hsieh received the PhD degree in measurement and statistics from the University of Iowa, Iowa city, in 2007. She is an associate research fellow at National Academy for Educational Research, Taiwan. Her research explores the uses and interpretations of educational assessment, with an emphasis on evaluating the accountability using modern test theory. Her work has investigated a variety of technical and policy issues in the uses of test data, including performance assessment designs and the influence of tests on policy making.

Ming-Hsueh Tsai received the PhD degree in education policy and management from the University of Taipei University of Education, in 2008. He is an associate research fellow at National Academy for Educational Research, Taiwan. His research explores of education administrative staff professional development and school quality of the relationship. His work has investigated a variety of education administrative staff professional development and performance assessments.

Yu Tsao (M'09) received the BS and MS degrees in electrical engineering from National Taiwan University, in 1999 and 2001, respectively, and the PhD degree in electrical and computer engineering from the Georgia Institute of Technology, in 2008. From 2009 to 2011, he was a researcher in the National Institute of Information and Communications Technology (NICT), Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. Currently, he is an associate research fellow of the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taiwan. His research interests include speech and speaker recognition, acoustic and language modeling, multimedia signal and information processing, pattern recognition, and machine learning. He is a member of the IEEE.

Chi-Chun Lee (M'13) received the PhD degree in electrical engineering from USC, in 2012. He is an assistant professor in the Electrical Engineering Department, the National Tsing Hua University (NTHU), Taiwan. He was a data scientist at id:a lab at ID Analytics, Inc. from Feb. 2013-Dec. 2013. His research interests include interdisciplinary human-centered behavioral signal processing. He was awarded with the USC Annenberg Fellowship (2007-2009). He led a team to participate and win the Emotion Challenge Classifier Sub-Challenge in Interspeech 2009. He is a coauthor on the best paper award in InterSpeech 2010. He is a member of Tau Beta Pi, Phi Kappa Phi and Eta Kappa Nu. He is also a member of ISCA. He has been a reviewer for multiple internationally-renowned journals and technical conferences. He is a member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.