Multimodal Public Speaking Performance Assessment

Torsten Wörtwein
KIT Institute of Anthropomatics and Robotics
Karlsruhe, Germany
uncwt@student.kit.edu

Mathieu Chollet
USC Institute for Creative Technologies
Playa Vista, CA, USA
chollet@ict.usc.edu

Boris Schauerte
KIT Institute of Anthropomatics and Robotics
Karlsruhe, Germany
boris.schauerte@kit.edu

Louis-Philippe Morency
CMU Language Technologies Institute
Pittsburgh, PA, USA
morency@cs.cmu.edu

Rainer Stiefelhagen
KIT Institute of Anthropomatics and Robotics
Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

Stefan Scherer
USC Institute for Creative Technologies
Playa Vista, CA, USA
scherer@ict.usc.edu

ABSTRACT

The ability to speak proficiently in public is essential for many professions and in everyday life. Public speaking skills are difficult to master and require extensive training. Recent developments in technology enable new approaches for public speaking training that allow users to practice in engaging and interactive environments. Here, we focus on the automatic assessment of nonverbal behavior and multimodal modeling of public speaking behavior. We automatically identify audiovisual nonverbal behaviors that are correlated to expert judges' opinions of key performance aspects. These automatic assessments enable a virtual audience to provide feedback that is essential for training during a public speaking performance. We utilize multimodal ensemble tree learners to automatically approximate expert judges' evaluations to provide post-hoc performance assessments to the speakers. Our automatic performance evaluation is highly correlated with the experts' opinions with r = 0.745 for the overall performance assessments. We compare multimodal approaches with single modalities and find that the multimodal ensembles consistently outperform single modalities.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities

General Terms

Algorithms, Human Factors, Measurement, Performance, Experimentation

Keywords

Public Speaking; Machine Learning; Nonverbal Behavior; Virtual Human Interaction

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

ACM 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2820762 .

1. INTRODUCTION

Recent developments in nonverbal behavior tracking, machine learning, and virtual human technologies enable novel approaches for interactive training environments [7, 3, 30]. In particular, virtual human based interpersonal skill training has shown considerable potential in the recent past, as it proved to be effective and engaging [15, 25, 2, 12, 4]. Interpersonal skills such as public speaking are essential assets for a large variety of professions and in everyday life. The ability to communicate in social and public environments can greatly influence a person's career development, help build relationships, resolve conflict, or even gain the upper hand in negotiations. Nonverbal communication expressed through behaviors, such as gestures, facial expressions, and prosody, is a key aspect of successful public speaking and interpersonal communication. This was shown in many domains including healthcare, education, and negotiations where nonverbal communication was shown to be predictive of patient and user satisfaction as well as negotiation performance [27, 8, 22]. However, public speaking with good nonverbal communication is not a skill that is innate to everyone, but can be mastered through extensive training [12].

We propose the use of an interactive virtual audience for public speaking training. Here, we focus primarily on the automatic assessment of nonverbal behavior and multimodal modeling of public speaking behavior. Further, we assess how nonverbal behavior relates to a number of key performance aspects and the overall assessment of a public speaking performance. We aim to identify nonverbal behaviors that are correlated to expert judges' opinions automatically in order to enable a virtual audience to provide feedback during a public speaking performance. Further, we seek to model the performance using multimodal machine learning algorithms to enable the virtual audience to provide post-hoc performance assessments after a presentation in the future. Lastly, we compare subjective expert judgements with objective manually transcribed performance aspects of two key behaviors, namely the use of pause fillers and the ability to hold eye contact with the virtual audience. In particular, we investigate three main research questions:

Q1: What nonverbal behaviors are correlated with expert judgements of certain performance aspects of public speaking to enable online feedback to the trainee?

Q2: Is it possible to automatically approximate public speaking performance assessments of expert judges using multimodal machine learning to provide post-hoc feedback to the trainee?

Q3: Are expert judges objective with their assessments, when compared to ground truth manual annotations of two key behaviors and how do automatic behavior assessments compare?

2. RELATED WORK

In general, excellent and persuasive public speaking performances, such as giving a presentation in front of an audience, are not only characterized by decisive arguments or a well structured train of thoughts, but also by the nonverbal characteristics of the presenter's performance, i.e. the facial expressions, gaze patterns, gestures, and acoustic characteristics. This has been investigated by several researchers in the past using political speakers' performances. For example researchers found that vocal variety, as measured by fundamental frequency (f_0) range and maximal f_0 of focused words are correlated with perceptual ratings of a good speaker within a dataset of Swedish parliamentarians [29, 24]. Further, manual annotations of disfluencies were identified to be negatively correlated with a positive rating.

In [26], the acoustic feature set, used in [29], was complemented by measures of pause timings and measures of tense voice qualities. The study shows that tense voice quality and reduced pause timings were correlated with overall good speaking performances. Further, the authors investigated visual cues, in particular motion energy, for the assessment of the speakers' performances. They found that motion energy is positively correlated with a positive perception of speakers. This effect is increased when only visual cues are presented to the raters.

A specific instance of public speaking are job interviews. In [21] researches have tried to predict the *hireability* in job interviews. Based on a dataset of 43 job interviews, the following non-verbal behaviors were used to estimate the *hireability*: manual annotations of body activity (gestures and self-touches), hand speed and position (on table height or on face height), and the speaking status to mask the temporal features. The most useful feature was the activity histogram from the hand position and speed.

Within this work we employ a virtual audience for public speaking training. Virtual humans are used in a wide range of social and interpersonal skill training environments, such as job interview training [2, 14], public speaking training [4, 23], and intercultural communicative skills training [18].

In [4], the use of a virtual audience and the automatic assessment of public speaking performances was investigated for the first time. A proof-of-concept non-interactive virtual public speaking training platform named Cicero was introduced. Three main findings were reported: nonverbal behaviors of only 18 subjects, such as flow of speech, vocal variety, were significantly correlated with an overall assessment of a presenter's performance as assessed by public speaking experts. A simple support vector regression approach showed promising results of automatic approximation of the experts' overall performance assessment with a significant correlation of $r = 0.617$ ($p = 0.025$), which approaches the correlation between the experts' opinions (i.e. $r = 0.648$).

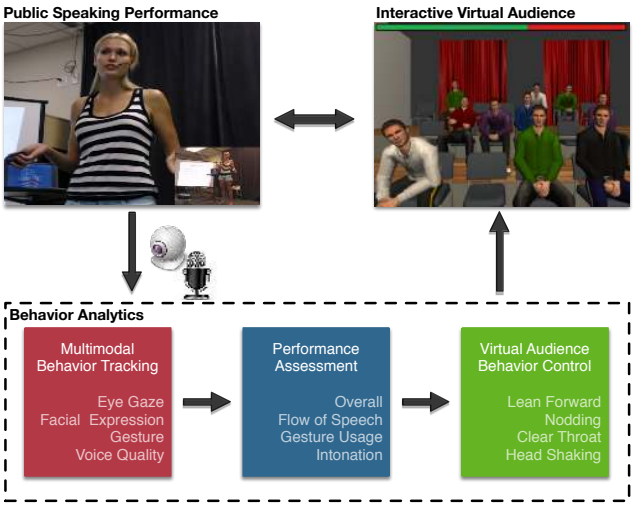

Figure 1: Depiction of the virtual human driven public speaking training interaction loop. The trainee's performance is automatically assessed within the behavior analytics framework. Multimodal nonverbal behavior is tracked, the performance is automatically assessed using machine learning approaches and the virtual audience provides audiovisual feedback to the trainee based on training strategies.

Within the present work, we aim to further improve these promising results using more sophisticated machine learning approaches. In addition, the present work is concerned with the assessment of *individual speaker improvement after training* rather than how well they present in comparison to other speakers. The present work is to the best of our knowledge the first to identify within-subject improvements and how they relate to changes in nonverbal behavior from one presentation to the other. Previous work did either not focused specifically on public speaking performance, did not investigate a data-driven comparison of presentations, or did not solely rely on automatically extracted features. Therefore, we focus on automatic features indicating differences in public speaking performances between presentations.

3. INTERACTIVE VIRTUAL AUDIENCE

We developed an interactive learning framework based on audiovisual behavior sensing and learning feedback strategies (cf. Figure 1). In particular, the speaker's audiovisual nonverbal behavior was registered in the architecture and feedback was provided to the speaker via the learning feedback strategies. We investigated three such strategies: (1) no feedback, i.e. the control condition, (2) direct visual feedback, and (3) nonverbal feedback from the virtual audience itself. Within this work, however, we solely focus on the performance of the speakers rather than the effect of learning strategies.

During training, the interactive virtual characters were configured with a feedback profile as our learning strategies. These profiles define behaviors the virtual characters will enact when specific conditions were met. Thus, the virtual characters can be used to provide natural, nonverbal feedback to the users according to their performance [5]. In our case, the characters can change their postures (leaning forward and being attentive, standing straight on their chairs, and

leaning backwards in a very relaxed manner), and also nod or shake their head regularly. Positive behaviors (leaning forward, nodding regularly) were assigned to trigger when the speaker's performance is good, while negative behaviors (leaning backwards, head shake) would trigger when the speaker's performance is bad. Threshold values for these triggers were randomly sampled for each virtual character so that the virtual characters would not behave simultaneously. Alternatively, we can provide color-coded *direct visual feedback*. A color-coded bar allows us to display the internal value of a behavioral descriptor directly, giving immediate feedback to the speaker about his or her performance.

To provide the learning framework with perceptual information on the speaker's performance, we made use of a wizard of Oz interface in order to ensure correct detection of the target behaviors (i.e. eye contact and pause fillers). In the future, we will utilize automatic audiovisual behavior tracking and machine learned models, investigated in this work, to automatically assess the speaker's performance, as suggested in [4]. Within the architecture, the perceptual information was aggregated into public speaking performance descriptors that directly influenced either the virtual audience's nonverbal behavior as feedback or the direct visual overlays.

4. METHODS

4.1 Experimental Design

As an initial study, we had users train with our virtual audience prototype with a pre- to post-training test paradigm, i.e. we compare learning outcomes between a pre-training performance and a post-training performance. By following this paradigm, we can assess speakers' relative performance improvement while compensating for their initial public speaking expertise.

4.1.1 Study Protocol

Participants were instructed they would be asked to present two topics during 5-minute presentations and were sent material (i.e. abstract and slides) to prepare the day of the study. Before recording the first presentation, participants completed questionnaires on demographics, self-assessment, and public-speaking anxiety. Each participant gave four presentations. The first and fourth consisted of the pre-training and post-training presentations, where the participants were asked to present the same topic in front of a passive virtual audience. Between these two tests, the participants trained for *eye contact* and *avoiding pause fillers* in two separate presentations, using the second topic. We specifically chose these two basic behavioral aspects of good public speaking performances following discussions with Toastmasters[1] experts. In addition, these aspects are clearly defined and can be objectively quantified using manual annotation enabling our threefold evaluation. In the second and third presentations, the audience was configured to use condition dependent different feedback strategies: no feedback, direct feedback through a red-green bar-indicator, or through non-verbal behavior of the audience. The condition was randomly assigned to participants when they came in. Please note, this work does not rely on these conditions. All three research questions focus on the pre- and post-training presentation.

[1] `http://www.toastmasters.org/`

The virtual audience was displayed using two projections to render the audience in life-size. The participants were recorded with a head mounted microphone, with a Logitech web camera capturing facial expressions, and a Microsoft Kinect placed in the middle of the two screens capturing the body of the presenter.

4.2 Participants and Dataset

Participants were recruited from Craigslist[2] and paid USD 25. In total, 47 people participated (29 male and 18 female) with an average age of 37 years ($SD = 12.05$). Out of the 47 participants 30 have some college education.
Two recordings had technical problems leaving a total of 45 participants On average the pre-training presentations lasted for 237 seconds ($SD = 116$) and the post-training presentation 234 seconds ($SD = 137$) respectively. Overall, there is no significant difference in presentation length between pre- and post-training presentations.

Experts

To compare the pre- with the post-training presentations, three experienced experts of the worldwide organization of Toastmasters were invited and paid USD 125. Their average age is 43.3 year ($SD = 11.5$), one was female and two were male. The experts rated their public speaking experience and comfort on 7-point Likert scales. On average they felt very comfortable presenting in front of a public audience ($M = 6.3$, with 1 - not comfortable, 7 - totally comfortable). They have extensive training in speaking in front of an audience ($M = 6$, with 1 - no experience, 7 - a lot of experience).

4.3 Measures

To answer our research questions we need different measures. For **Q1** as well as **Q2** we need an expert assessment and automatically extracted features. In addition to an expert assessment, **Q3** requires manually annotated behaviors.

4.3.1 Expert Assessment

Three Toastmasters experts, who were blind to the order of presentation (i.e. pre-training vs. post-training), evaluated whether participants improved their public speaking skills. Experts viewed videos of the presentations. The videos are presented pairwise for a direct comparison in a random order. Each video showed both the participant's upper body and facial expressions (cf. Figure 1). Each expert evaluated the performance differences for all pairs on 7-point Likert scales for all ten aspects. In particular, they assessed performance aspects derived from prior work on public speaking assessment [27, 4, 26, 24] and targeted discussions with experts apply more to the pre- or post-training presentation[3]:

1. Eye Contact
2. Body Posture
3. Flow of Speech
4. Gesture Usage
5. Intonation
6. Confidence Level
7. Stage Usage
8. Avoids pause fillers
9. Presentation Structure
10. Overall Performance

The pairwise agreement between the three experts is measured by the absolute distance between the experts' Likert

[2] `http://www.craigslist.org/`

[3] Aspect definitions and a dummy version of the questionnaire are available: `http://tinyurl.com/ovtp67x`

scale ratings. The percentage of agreement with a maximal distance of 1 ranges between 63.70% and 81.48% for all 10 aspects, indicating high overall agreement between raters.

4.3.2 Objective measures

To complement expert ratings, we evaluated public speaking performance improvement using two objective measures, namely *eye contact* and the *avoidance of pause fillers*. The presenters were specifically informed about these two aspects in the training presentations for all three conditions. In order to create objective individual baselines, we annotated both measures for all pre-training and post-training test presentations. Two annotators manually marked periods of *eye contact* with the virtual audience and the occurrence of *pause fillers* using the annotation tool ELAN [28]. For both aspects we observed high inter-rater agreement for a randomly selected subset of four videos that both annotators assessed. The Krippendorff α for eye contact is $\alpha = 0.751$ and pause fillers $\alpha = 0.957$ respectively. Krippendorff's α is computed on a frame-wise basis at 30 Hz.

For eye contact we computed a ratio for looking at the audience $\in [0, 1]$, with 0 = never looks at the audience and 1 = always looks at the audience, over the full length of the presentation based on the manual annotations. The number of pause filler words were normalized by the duration of the presentation in seconds.

The improvement is measured by the normalized difference index *ndi* between the pre-training and post-training test presentations for both objectively assessed behaviors and was calculated by

$$ndi = \frac{post - pre}{post + pre}. \quad (1)$$

4.3.3 Automatic Behavior Assessment

In this section the automatically extracted features and the used machine learning algorithms are introduced. The following features of the pre- and post-training presentations are combined with equation 1 to reflect improvement or the lack thereof.

Acoustic Behavior Assessment.

For the processing of the audio signals, we use the freely available COVAREP toolbox (v1.2.0), a collaborative speech analysis repository [6]. COVAREP provides an extensive selection of open-source robust and tested speech processing algorithms enabling comparative and cooperative research within the speech community.

All following acoustic features are masked with voiced-unvoiced (VUV) [9], which determines whether the participant is voicing, i.e. the vocal folds are vibrating. After masking, we use the average and the standard deviation of the temporal information of our features. Not affected by this masking is VUV itself, i.e. the average of VUV is used as an estimation of the ratio of speech to pauses.

Using COVAREP, we extract the following acoustic features: the maxima dispersion quotient (MDQ) [17], peak slope (PS) [16], normalized amplitude quotient (NAQ) [1], the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2) [31], and the estimation of the R_d shape parameter of the Liljencrants-Fant glottal model (RD) [11]. Beside these features we also use the fundamental frequency (f_0) [9] and the first two KARMA filtered formants (F_1, F_2) [20]. Additionally, we use the first four Mel-frequency cepstral coefficients ($MFCC_{0-3}$) and extract the voice intensity in dB.

Visual Behavior Assessment.

Gestures are measured by the change of upper body joints' angles from the Microsoft Kinect. Therefore, we take the sum of differences in angles (from the following joints: shoulder, elbow, hand, and wrist). To eliminate noise, we set the difference to zero when not both hands are above the hips. To avoid assigning too much weight to voluminous gestures, we truncate the differences when the difference is higher than a threshold, which we calculated from manual gesture annotations of 20 presentations. In the end, we use the mean of the absolute differences as an indicator for gesturing during the presentation.

We evaluate eye contact with the audience based on two eye gaze estimations. The eye gaze estimation from OKAO [19] head orientation from CLNF [3] are used separately to classify whether a participant is looking at the audience or not. The intervals for the angles, which we use to attest eye contact, are calculated from our eye contact annotations. Finally, we use the ratio of looking at the audience relative to the length of the presentation as a feature.

Emotions, such as anger, sadness, and contempt, are extracted with FACET [10]. After applying the confidence provided by FACET, we take the mean of the emotions' intensity as features.

Machine Learning.

To approximate the experts' combined performance assessments we utilize a regression approach with the experts' averaged ratings as targets for all ten aspects. For this, we use Matlab's implementation of a least squared boosted regression ensemble tree. We evaluate our predictions with a leave one speaker out cross-evaluation. Since we have relatively many features with respect to the number of participants, we use the forward feature selection to find a subset of features using the speaker independent validation strategy. This kind of feature selection starts with an empty set of features and iteratively adds the feature that decreases together with the chosen features a criterion function the most. The resulting feature subset might not be optimal. As a criterion function we use $(1 - \mathrm{corr}(\hat{y}, y))^2$, where $\hat{y}$ are the predictions of the leave one speaker out cross-evaluation and y the ground truth.

5. RESULTS

5.1 Q1 - Behavioral Indicators

We report correlation results with the linear Pearson correlation coefficient along with the degrees of freedom and the p-value. Table 1 summarizes the features, which change between the pre- and post-training presentation measured by *ndi* (Eq. 1) strongly correlates with aspects of public speaking proficiency. We list only detail statistical findings for a reduced number of aspects due to space restrictions. Please note, this table and the following paragraphs are not complete in respect to all aspects nor do they mention all features. For a complete list of all aspects and correlating features see Table 1 of the supplementary material.

Eye Contact: Experts' eye contact assessments slightly correlate with an increase of the average eye contact as

Table 1: Overview of correlating features for each improvement aspect. Arrows indicating the direction of the correlation, i.e. ↑ means positive correlation and ↓ means negative correlation.

Improvement Aspect	Correlating Behavior
Eye Contact	↑ eye contact ↓ contempt ↓ VUV std ↓ H1H2 std
Flow of Speech	↑ voice intensity variation ↑ peak slope std ↑ $MFCC_0$ std ↓ H1H2 std ↓ $MFCC_0$ mean ↓ VUV std
Gesture Usage	↑ ratio of speech and pauses ↑ gesture ↑ MDQ mean
Intonation	↑ pitch mean ↑ peak slope std ↑ vocal expressivity ↓ peak slope mean
Confidence	↑ F_1 mean ↑ vocal expressivity ↑ ratio of speech and pauses ↑ pitch mean ↓ $MFCC_0$ mean ↓ peak slope mean
Pause Fillers	↑ peak slope std ↓ contempt ↓ VUV std
Overall Performance	↑ peak slope std ↓ H1H2 std ↓ contempt ↓ VUV std

assessed by OKAO ($r(43) = 0.24$, $p = 0.105$). Also, assessed contempt facial expressions correlate negatively with the eye contact assessment ($r(43) = -0.33$, $p = 0.029$). We identify two negatively correlating acoustic features, namely a decrease of the standard deviation of VUV ($r(43) = -0.34$, $p = 0.024$) and a decrease of the standard deviation of H1H2 ($r(43) = -0.45$, $p = 0.002$).

Flow of Speech: Experts' flow of speech assessment correlates only with acoustic features. They include an increase of the standard deviation of the voice intensity ($r(43) = 0.31$, $p = 0.039$), a decrease of the average of $MFCC_0$ ($r(43) = -0.36$, $p = 0.015$) as well as an increase of the standard deviation of it ($r(43) = 0.35$, $p = 0.019$). Beside these features also a decrease in the standard deviation of VUV ($r(43) = -0.39$, $p = 0.008$), a decrease in the standard deviation of H1H2 ($r(43) = -0.33$, $p = 0.026$), and an increase in the standard deviation of PS ($r(43) = 0.31$, $p = 0.040$) correlated with *flow of speech*.

Confidence: Experts' confidence assessment correlates with several acoustic features. It correlates with an increasing average pitch ($r(43) = 0.32$, $p = 0.034$), an increase of mean of VUV ($r(43) = 0.32$, $p = 0.031$), a decrease of VUV standard deviation ($r(43) = -0.40$, $p = 0.006$), the decrease of the standard deviation of H1H2 ($r(43) = -0.36$, $p = 0.016$), and a decrease of the average PS ($r(43) = -0.32$, $p = 0.034$). In addition to these features, an increase of the standard deviation of $MFCC_2$ ($r(43) = 0.30$, $p = 0.044$) correlate with *confidence*, too. Lastly, the following formants correlate with *confidence* of the first formant an increase in the average ($r(43) = 0.34$, $p = 0.021$) as well as an increase of the standard deviation ($r(43) = 0.31$, $p = 0.037$) and of the bandwidth from the second formant a decrease of the average ($r(43) = -0.32$, $p = 0.033$) as well as a decrease of the standard deviation ($r(43) = -0.40$, $p = 0.007$).

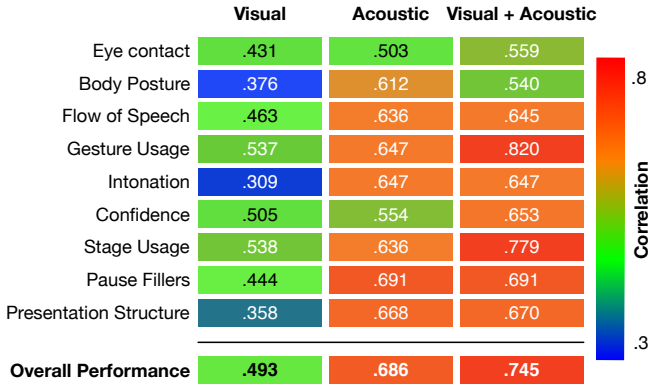

Figure 2: Color coded visualization of the Pearson correlation between the expert assessments of all evaluated aspects and the automatic prediction using both single modalities and both combined.

Avoids Pause Fillers The assessed avoidance of pause fillers correlates with a decrease of the standard deviation of VUV ($r(43) = -0.52$, $p < 0.001$) and an increase of PS's standard deviation ($r(43) = 0.32$, $p = 0.030$). Furthermore, it correlates with two visual features, namely being less contempt ($r(43) = -0.36$, $p = 0.016$) and having more neutral facial expressions ($r(43) = 0.35$, $p = 0.019$).

Overall Performance: Finally, the assessed overall performance correlates with showing less contempt facial expressions ($r(43) = -0.32$, $p = 0.030$) and the following acoustic features: a decrease of the standard deviation of VUV ($r(43) = -0.46$, $p = 0.002$), a decrease of the standard deviation of H1H2 ($r(43) = -0.31$, $p = 0.039$), an increase in PS' standard deviation ($r(43) = 0.36$, $p = 0.015$), and a decrease of the bandwidth from the second formant ($r(43) = -0.30$, $p = 0.042$).

5.2 Q2 - Automatic Assessment

In Figure 2 we report Pearson's correlation results for every aspect using the three feature sets (*visual*, *acoustic*, and *acoustic+visual*). As seen in Figure 2, the acoustic+visual feature set outperforms the other two modalities consistently. We present the p-values of two-tailed t-tests and Hedges' g values as a measure of the effect size. The g value denotes the estimated difference between the two population means in magnitudes of standard deviations [13].

In addition to correlation assessments, we investigate mean absolute errors of our automatic assessment using tree ensembles. As a comparison performance we use the mean over all expert ratings for every aspect and all participants as a constant baseline and compare it with the automatic assess-

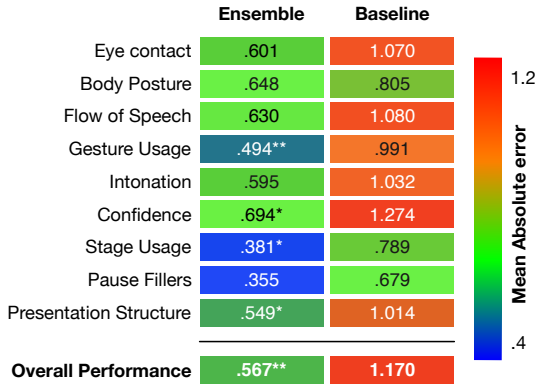

Figure 3: Color coded visualization of the mean absolute error of the ensemble tree approach and the baseline assessment. Significant improvements are marked with * ($p < 0.05$) and ** ($p < 0.01$).

ment using the mean absolute error. Our prediction errors ($M = 0.55$, $SD = 0.42$) are consistently lower compared to the baseline errors ($M = 0.74$, $SD = 0.57$) and significantly better ($t(898) = 0.50$, $p < 0.001$, $g = -0.372$) across all aspects. Additionally, for *overall performance* alone the automatic assessment ($M = 0.57$, $SD = 0.46$) is also significantly better than the baseline ($M = 0.88$, $SD = 0.61$; $t(88) = 0.54$, $p = 0.008$, $g = -0.566$). For a full comparison of all aspects between our prediction errors and the constant prediction errors see Figure 3.

When we compare the different modalities, i.e. acoustic and visual features separately as well as jointly, we identified the following results: *Acoustic* features only ($M = 0.59$, $SD = 0.45$; $t(898) = 0.50$, $p = 0.002$, $g = -0.202$) and *acoustic+visual* features ($M = 0.55$, $SD = 0.42$; $t(898) = 0.48$, $p < 0.001$, $g = -0.297$) significantly outperform the *visual* ($M = 0.70$, $SD = 0.54$) features. We do not observe a significant difference between *acoustic* and *acoustic+visual* ($t(898) = 0.44$, $p = 0.140$, $g = 0.099$).

5.3 Q3 - Expert Judgments

We annotated eye contact and pause filler words in all pre- and post-training presentations, see Section 4.3. For the eye contact annotation, we use the normalized value of the annotated aspect, the annotated period divided by the length of the presentations. The count of pause filler words is used directly as a score. Thereby, we have scores of the annotated aspects representing the annotated behavior in the pre- and post-training presentation.

To have a score comparable to the expert assessment, we use equation 1 to combine the scores of the pre- and post-training presentation. We do not observe a significant correlation between the experts' eye contact assessment and the annotated eye contact ($r(43) = 0.20$, $p = 0.197$) as well as no correlation between the pause filler annotations and the assessed pause fillers ($r(43) = -0.05$, $p = 0.739$). However, we observe a strong correlation between the automatic measures of eye contact using OKAO ($r(88) = 0.62$, $p < 0.001$) and CLNF ($r(88) = 0.66$, $p < 0.001$) and the manually annotated eye contact.

6. DISCUSSION

6.1 Q1 - Behavioral Indicators

Our first research question aims at identifying within-subject nonverbal behavioral changes between pre- and post-training presentations and how they relate to expert assessments. All reported behaviors are automatically assessed and changes are identified using the *ndi* (see Section 4). We correlate the observed behavioral changes with expert assessed performances in all ten aspects. Our findings confirm that large portions of a public speaking performance improvements are covered by nonverbal behavior estimates.

As summarized in Table 1, we can identify a number of multimodal nonverbal behaviors that are correlated with expert assessments for the investigated aspects as well as for overall performance improvement (due to space restrictions we excluded three aspects from our analysis). For several aspects, such as *intonation* and *gesture usage* prior intuitions are confirmed. For example vocal expressivity, increased tenseness (as measured by peak slope), and pitch are correlated with an improvement in the assessment of intonation. Further, the increased usage of gestures from pre- to post-training presentations is correlated with an assessed improvement of gesture usage. In addition, multifaceted aspects such as *confidence* show a broad spectrum of automatic measures that are correlated, both positively and negatively. For the assessment of *overall performance*, we identified that the use of contempt facial expressions is negatively correlated with performance improvement. This could be interpreted that if the presenter accepted the virtual audience and engaged in training they gained more from the experience. Further, changes in pause to speech ratio and increased variability in the voice as measured with peak slope correlate with overall performance improvement. Change of the standard deviation of VUV negatively correlates most prominently with a number of aspects. As VUV is a logical vector, i.e. whether a person is voicing or not, the standard deviation is similar to the entropy. This indicates that a decrease in speaking variety or the development of a more regular flow of speech, is a key feature of public speaking performance improvement.

As shown in Figure 1, we plan to incorporate the identified automatic measures of behavioral change as input to control the virtual audience online feedback. We anticipate to use both targeted feedback behavior, such as a virtual audience member clears its throat to signify that it feels neglected by lack of eye contact, as well as complex feedback covering multiple aspects. The exact types of behavioral online feedback need to be further investigated in future studies; it is important that the behaviors of the virtual audience are clearly identifiable. Only then will the trainee be able to reflect on his or her behavior during training and ultimately improve.

6.2 Q2 - Automatic Assessment

In order to answer the second research question, we conducted extensive unimodal and multimodal experiments and investigate ensemble trees to automatically approximate the experts' assessments on the ten behavioral aspects. Figure 2 summarizes the observed correlation performance of our automatic performance assessment ensemble trees. In addition, we investigate mean absolute errors for the regression output of the ensemble trees compared to an average baseline. For both performance measures, i.e. correlation and mean abso-

lute error, we observe that multimodal features consistently outperform unimodal feature sets. In particular, complex behavioral assessments such as the overall performance and confidence of the speaker benefit from features of multiple modalities. Out of the single modalities the acoustic information seems to be most promising for the assessment of performance improvement. However, we are confident that with the development of more complex and tailored visual features similar success can be achieved. When compared to the baseline, the ensemble tree regression approach significantly improves baseline assessment for several aspects including overall performance.

One of the main reasons for choosing ensemble trees as the regression approach of choice is the possibility to investigate the selected features to achieve the optimal results. This enables us to investigate behavioral characteristics of public speaking performance improvement in detail. For the overall performance estimation the multimodal ensemble tree selected negative facial expressions, pause to speech ratio, average RD measure, average second and third formants, as well as the second formant's bandwidth. This selection shows the importance for both the facial expressions and the voice characteristics for the assessment of performance improvement. Overall the ensemble trees' output is correlated with the experts' assessment at $r > 0.7$, which is a considerably high correlation and a very promising result.

In the future, we plan to use these automatic performance improvement estimates to give the trainees targeted post-hoc feedback on what aspects did improve and which need further training. Similar to the work in [14], we plan to provide a visual performance report to the trainee as a first step. In addition, such visual reports can be enhanced by a virtual tutor that guides the trainee through the assessment and possibly replays several scenes from the presentation and provides motivational feedback to the trainee.

6.3 Q3 - Expert Judgments

When investigating the third research question, we identified a considerable difference between the manual ground truth labels of two key behaviors and experts' subjective opinions. In particular, we found that there is no correlation between the experts' assessment of the used number of pause fillers and the actual number of pause fillers uttered by the presenters. We also found no correlation between the assessment of eye contact and the manually annotated times of actual eye contact of the presenter with the virtual audience. However, we found a strong correlation between the automatic measures of eye contact with the manual annotation. This finding could explain why the automatically assessed eye contact only slightly correlates with the experts' assessment (see Section 5.1).

It is possible to argue that the experts did not accurately count the number of pause fillers during their assessments and if they had been provided with the exact numbers their assessment would have changed, however, this probably does not hold for the more complex behavior of eye contact. While both the manual annotation and the automatic assessment of eye contact are crude measures of the time a presenter looks at an audience, the experts might inform their decision on a much more complex basis. In particular, we believe that eye gaze patterns such as slowly swaying the view across the entire audience vs. staring at one single person for an entire presentation might be strong indicators of excellent or poor gaze behaviors. A distinction between such behaviors is not possible using the utilized crude measures. As mentioned earlier, we plan to investigate more tailored behavioral descriptors in the near future.

7. CONCLUSION

Based on the three research questions investigated in this work we can identify the following main findings: **Q1** We could identify both visual and acoustic nonverbal behaviors that are strongly correlated with pre- to post-training presentation performance improvement and the lack thereof. In particular, facial expressions, gestures, and voice characteristics are identified to correlate with performance improvement as assessed by public speaking experts from the Toastmasters organization. **Q2** Based on the automatically tracked behaviors, we investigated machine learning approaches to approximate the public speaking performance on ten behavioral aspects including the overall performance. We showed that the multimodal approach utilizing both acoustic and visual behaviors consistently outperformed the unimodal approaches. In addition, we found that our machine learning approach significantly outperforms the baseline. **Q3** Lastly, we investigated manual ground truth assessments for eye contact and number of pause fillers used in pre- and post-training presentations and how they relate to expert assessments. We could identify a considerable difference between expert assessments and actual improvement for the two investigated behaviors. This indicates that experts base their assessments on more complex information than just the amount spent looking at the audience or the number of pause fillers uttered, but rather identify patterns of behaviors showing proficiency. We plan to investigate such more complex patterns of behaviors in the near future and confer with our public speaking experts to inform us on their decision process.

Overall, our findings and results are promising and we believe that this accessible technology has the potential to impact training focused on the nonverbal communication aspects of in fact a wide variety of interpersonal skills, including but not limited to public speaking.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1421330 and U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government, and no official endorsement should be inferred.

8. REFERENCES

[1] P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[2] K. Anderson and et al. The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In *Proceedings of International Conference on Advances in Computer Entertainment*, pages 476–491, 2013.

[3] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial

landmark detection in the wild. In *International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[4] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. Cicero - towards a multimodal virtual audience platform for public speaking training. In *Proceedings of Intelligent Virtual Agents 2013*, pages 116–128. Springer, 2013.

[5] M. Chollet, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer. An interactive virtual audience platform for public speaking training. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems*, pages 1657–1658, 2014.

[6] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. Covarep - a collaborative voice analysis repository for speech technologies. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 960–964, 2014.

[7] D. DeVault and et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of Autonomous Agents and Multiagent Systems*, pages 1061–1068, 2014.

[8] M. R. DiMatteo, R. D. Hays, and L. M. Prince. Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. *Health Psychology*, 5(6):581, 1986.

[9] T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976, 2011.

[10] Emotient. FACET SDK, 2014. `http://www.emotient.com/products`.

[11] G. Fant, J. Liljencrants, and Q. Lin. The LF-model revisited. transformations and frequency domain analysis. *Speech Transmission Laboratory, Quarterly Report, Royal Institute of Technology*, 2(1):119–156, 1995.

[12] J. Hart, J. Gratch, and S. Marsella. *How Virtual Reality Training Can Win Friends and Influence People*, chapter 21, pages 235–249. Human Factors in Defence. Ashgate, 2013.

[13] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981.

[14] M. Hoque, M. Courgeon, J.-C. Martin, M. Bilge, and R. Picard. Mach: My automated conversation coach. In *Proceedings of International Joint Conference on Pervasive and Ubiquitous Computing*, 2013.

[15] W. L. Johnson, J. W. Rickel, and J. C. Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11(1):47–78, 2000.

[16] J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180, 2011.

[17] J. Kane and C. Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(6):1170–1179, 2013.

[18] H. C. Lane, M. J. Hays, M. G. Core, and D. Auerbach. Learning intercultural communication skills with virtual humans: Feedback and fidelity. *Journal of Educational Psychology Special Issue on Advanced Learning Technologies*, 105(4):1026–1035, 2013.

[19] S. Lao and M. Kawade. Vision-based face understanding technologies and their applications. In *Proceedings of the Conference on Advances in Biometric Person Authentication*, pages 339–348, 2004.

[20] D. D. Mehta, D. Rudoy, and P. J. Wolfe. KARMA: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *Journal of the Acoustical Society of America*, 132(3):1732–1746, 2011.

[21] L. S. Nguyen, A. Marcos-Ramiro, M. Marrón Romera, and D. Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *Proceedings of the International Conference on Multimodal Interaction*, pages 437–444, 2013.

[22] S. Park, P. Shoemark, and L.-P. Morency. Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization. In *Proceedings of the 18th International Conference on Intelligent User Interfaces*, pages 37–46. ACM, 2014.

[23] D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, 11(1):68–78, 2002.

[24] A. Rosenberg and J. Hirschberg. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Interspeech 2005*, pages 513–516, 2005.

[25] J. Rowe, L. Shores, B. Mott, and J. C. Lester. Integrating learning and engagement in narrative-centered learning environments. In *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems*, 2010.

[26] S. Scherer, G. Layher, J. Kane, H. Neumann, and N. Campbell. An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 1114–1120. ELRA, 2012.

[27] L. M. Schreiber, D. P. Gregory, and L. R. Shibley. The development and test of the public speaking competence rubric. *Communication Education*, 61(3):205–233, 2012.

[28] H. Sloetjes and P. Wittenburg. Annotation by category: ELAN and ISO DCR. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

[29] E. Strangert and J. Gustafson. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Proceedings of Interspeech 2008*, pages 1688–1691, 2008.

[30] W. Swartout, R. Artstein, E. Forbell, S. Foutz, H. C. Lane, B. Lange, J. Morie, A. Rizzo, and D. Traum. Virtual humans for learning. *AI Magazine*, 34(4):13–30, 2013.

[31] I. Titze and J. Sundberg. Vocal intensity in speakers and singers. *Journal of the Acoustical Society of America*, 91(5):2936–2946, 1992.