A Multimodal Corpus for the Assessment of Public Speaking Ability and Anxiety

Mathieu Chollet[1], Torsten Wörtwein[2], Louis-Philippe Morency[3], Stefan Scherer[1]

[1] Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
[2] Institute of Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{mchollet, scherer}@ict.usc.edu, uncwt@student.kit.edu, morency@cs.cmu.edu

Abstract

The ability to efficiently speak in public is an essential asset for many professions and is used in everyday life. As such, tools enabling the improvement of public speaking performance and the assessment and mitigation of anxiety related to public speaking would be very useful. Multimodal interaction technologies, such as computer vision and embodied conversational agents, have recently been investigated for the training and assessment of interpersonal skills. Once central requirement for these technologies is multimodal corpora for training machine learning models. This paper addresses the need of these technologies by presenting and sharing a multimodal corpus of public speaking presentations. These presentations were collected in an experimental study investigating the potential of interactive virtual audiences for public speaking training. This corpus includes audio-visual data and automatically extracted features, measures of public speaking anxiety and personality, annotations of participants' behaviors and expert ratings of behavioral aspects and overall performance of the presenters. We hope this corpus will help other research teams in developing tools for supporting public speaking training.

Keywords: Multimodal corpus; public speaking training; virtual audiences

1. Introduction

Modern life often involves situations where we are required to speak in public, both in our professional lives, for instance when presenting results of our work in front of colleagues, and in our personal lives, such as when giving a toast at a wedding. A proficient public speaker mobilizes a vast array of skills to perform a good speech, ranging from the selection and arrangement of appropriate and convincing arguments to the efficient vocal and non-verbal delivery of the speech. Considering how prevalent public speaking situations are in modern professional and personal life, it is only natural some individuals would desire to improve their ability to speak in public. Additionally, public speaking anxiety is an extremely common fear (Furmark et al., 2000; Bodie, 2010), and some people experience an unmanageable amount of stress when preparing or undergoing public speaking. These two situations warrant for the development of tools and methods to support the assessment of public speaking ability, the training of public speaking skills and the reduction of public speaking anxiety.

Multimodal interaction technologies (*e.g.* social signals processing, virtual humans) have been deployed in many types of social skills training applications, from job interview training (Hoque et al., 2013; Ben Youssef et al., 2015) to intercultural skills training (Lane et al., 2013) or public speaking skills training (Damian et al., 2015; Chollet et al., 2015). Moreover, virtual audiences have been used for supporting people suffering from severe public speaking anxiety (North et al., 1998; Pertaub et al., 2002). Finally, recent works have proposed to automatically assess the public speaking ability of politicians (Rosenberg and Hirschberg, 2005; Scherer et al., 2012; Brilman and Scherer, 2015) or job applicants (Nguyen et al., 2013). The implementation of such technologies often require the use of multimodal corpora, either as data for training the models that will recognize multimodal behaviors (*e.g.* smiles, gestures) or higher level variables (*e.g.* emotions of the user, performance of a speaker), or for building the repertoire of behaviors of a virtual character.

In this paper, we present and share a multimodal corpus of public speaking presentations that we collected while studying the potential of interactive virtual audiences for public speaking skills training. This corpus includes audio-visual data and automatically extracted multimodal features, measures of public speaking anxiety and personality, annotations of participants' behaviors and expert ratings of behavioral aspects and overall performance of the presenters. We hope this corpus will help other research teams in developing tools for supporting public speaking training.

In the next section, after briefly presenting the interactive virtual audience framework and the study we conducted to gather data, we present our multimodal corpus of public speaking presentations. In section 3, we outline previous studies that were realized using this corpus, on the evaluation of public speaking improvement and the automatic assessment of public speaking ability and public speaking anxiety, in order to demonstrate that this corpus can be used for a variety of different purposes. Finally, we present future directions of research as well as current extension work on our corpus.

2. Multimodal Corpus

Our corpus was collected in the context of a study on the use of virtual audiences for public speaking training. During that study, we explored different feedback strategies of virtual audiences. To this effect, we compared learning outcomes of users training with a virtual audience providing feedback according to one of three investigated strategies using a pre- to post-training test paradigm. We present our system and these feedback strategies in the next section.

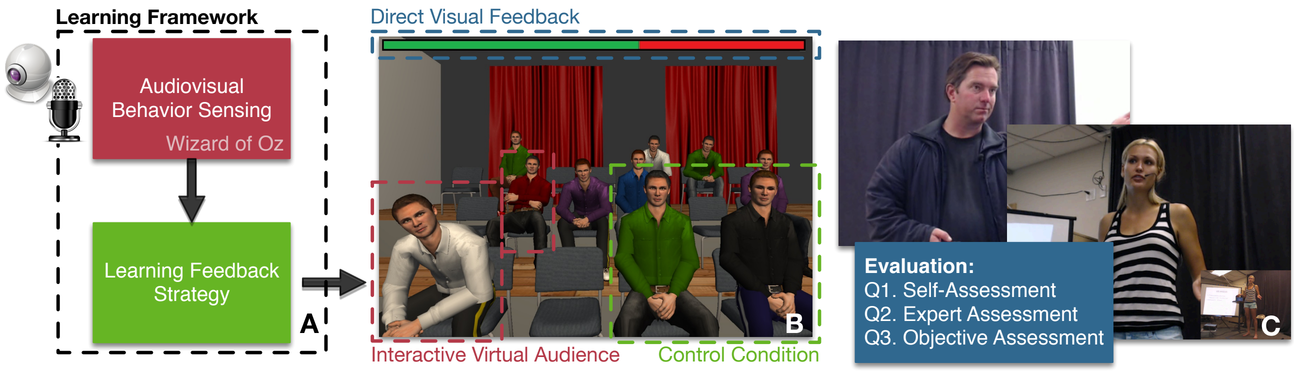


Figure 1: (A) Our architecture automatically provides multimodal realtime feedback based on the speaker's audiovisual behavior. (B) We evaluated three feedback strategies: (IVA) an interactive virtual audience, (DF) direct visual feedback, and (Non-IVA) a non-interactive virtual audience (control). (C) We evaluated the participants' performance improvement in a pre- vs. post-training evaluation paradigm with three assessment perspectives: (**Q1**) the presenters themselves, (**Q2**) public speaking experts, and (**Q3**) objectively quantified data.

2.1. Public Speaking Training System

We developed a public speaking training framework based on audiovisual behavior sensing and virtual audience feedback strategies: the speaker's vocal and nonverbal behaviors are detected and feedback is provided in return according to pre-defined strategies (Chollet et al., 2015). We investigated three such strategies, presented in Figure 1B.

In the *direct visual feedback* (DF) strategy, colored gauges were configured to give immediate feedback to the speaker about his/her performance. For instance, when training gaze behavior (*e.g.* increase eye contact with the audience), a full green bar would indicate to the participant that his/her performance is very good. However, the virtual characters only adopted a neutral posture and did not provide additional feedback. When using the *interactive virtual audience* (IVA) feedback strategy, the virtual characters would display behaviors when specific conditions were met. In our study, the characters would nod or lean forward when the participant's performance was good, and they would lean backwards or shake their head when it was poor. The thresholds used for triggering these behaviors were configured manually so that the different characters in the audience would not behave simultaneously and so that the ratio of positive *vs* negative behaviors would reflect the current performance of the user (*e.g.* if the participant was performing very well, all characters would be leaning forward and nodding regularly). Finally, in the *control condition* (Non-IVA), the virtual characters adopted a neutral posture and no gauge was displayed.

2.2. Study Protocol

A few days before their participation in the study, participants were instructed that they would have to present two topics during 5-minute presentations. The first topic was a presentation of Los Angeles, California, the city in which the study was performed. The second topic was a sales pitch for a beauty product. They were sent material about those presentations (*i.e.* abstract and slides) in advance to prepare before the day of the study. On the day of the study, participants first completed questionnaires on demographics, personality, and public speaking anxiety (*cf.* section 2.4.1.). Each participant then gave four presentations (*cf.* Figure 2). Presentations (1) and (4) consisted of pre- and post-training presentations where the participants presented the city of Los Angeles in front of a non-interactive virtual audience (*i.e.* configured in the *control condition*). Between these two tests, *i.e.* during presentations (2) and (3), the participants trained with our system using the sales pitch topic. In presentation (2), the training was targeted at reducing the amount of *pause fillers* they produced while speaking. In the second training presentation, *i.e.* presentation (3), the aim was to improve the participants' *eye contact* with the audience. Every participant was given an information sheet with quotes from public speaking experts of the Toastmasters organization[1] about how gaze and pause fillers impact a public speaking performance[2]. These two basic behavioral aspects of good public speaking performances were specifically chosen following discussions with Toastmasters experts. In addition, these aspects can be clearly defined and objectively quantified using manual annotation enabling our threefold evaluation. During the training presentations, *i.e.* presentations (2) and (3), the au-

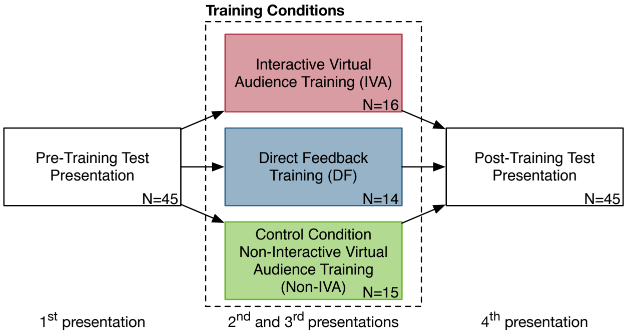

Training Conditions
Pre-Training Test Presentation
N=45
Interactive Virtual Audience Training (IVA)
N=16
Direct Feedback Training (DF)
N=14
Control Condition Non-Interactive Virtual Audience Training (Non-IVA)
N=15
Post-Training Test Presentation
N=45
1st presentation
2nd and 3rd presentations
4th presentation

Figure 2: Study protocol.

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

[2] Written hints provided before training: http://tinyurl.com/m4t6l62

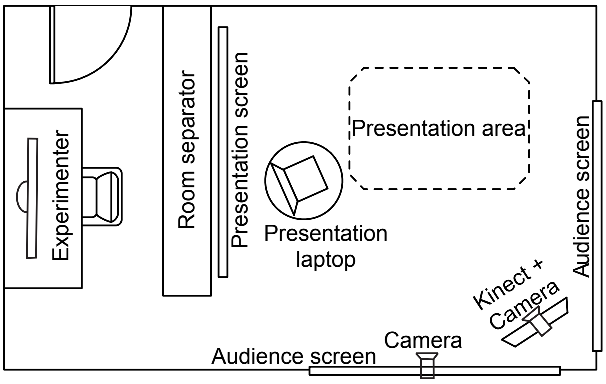


Figure 3: Study room setup.

dience was configured according to one of the three feedback conditions presented in the previous section, that is the use of an interactive virtual audience (IVA), direct visual feedback (DF), and non-interactive virtual audience (Non-IVA). The condition was randomly assigned to participants, and a *Wizard of Oz* setup was used: unbeknownst to the participants, a confederate was watching their presentation remotely, and pressed a button when they were uttering a pause filler or looking away. This allowed us to provide the virtual audience with real-time information about the speaker's performance on the trained behavior After the last presentation, participants completed questionnaires about their experience (*cf.* section 2.4.1.), were debriefed, paid, and escorted out.

2.3. Participants and Dataset

Participants were native English speakers of the Los Angeles area recruited from the classified ads website Craigslist[3]. In total, 47 people participated (29 M, 18 F) with an average age of 37 years ($SD = 12.05$). Two recordings had technical problems leaving a total of 45 participants, with 15 participants assigned to the control condition, 14 to the direct feedback condition, and 16 to the interactive virtual audience condition. Thus, our multimodal corpus constitutes a collection of 180 public speaking presentations. On average the pre-training presentations lasted for 3:57 minutes (SD=1:56 minutes) and the post-training presentation 3:54 minutes (SD=2:17 minutes) respectively, with no significant difference in presentation length.

For each presentation, the participants were recorded with a headset microphone, a Microsoft Kinect capturing the whole scene and two webcams positioned at different angles zoomed on the participant's upper body. Figure 3 provides an overview of the placement of the sensors.

2.4. Measures

In addition to the raw audio-visual and depth data recorded by our sensors, our corpus contains measures obtained from participants' questionnaires, expert assessments, manual annotation and automatic multimodal behavior assessment.

2.4.1. Participant Questionnaires

All participants completed questionnaires before the pre-training presentation: a demographics questionnaire, the 'Big Five Inventory' personality questionnaire (Rammstedt and John, 2007) and the 'Personal Report of Confidence as a Speaker (PRCS)' questionnaire (Paul, 1966), used to estimate public speaking anxiety (Hook et al., 2008). After the last presentation, participants completed a self assessment questionnaire, the 'Positive and Negative Affect Schedule' questionnaire (Crawford and Henry, 2004), and the immersive experience questionnaire (Jennett et al., 2008).

2.4.2. Expert Assessments

To compare the pre- with the post-training presentations, three experts of the Toastmasters organization evaluated whether participants improved their public speaking skills after training. They were presented the pre- and post-training videos alongside for direct comparison. Each video showed both the participant's upper body as well as facial expressions (*cf.* Figure 1 (C)). The position of the pre- and post-training video, *i.e.* left or right, was randomized for each pair, as well as the order of participants. Additionally, experts were unaware of the participant's training condition. They assessed whether 10 performance aspects - derived from prior work on public speaking assessment (Schreiber et al., 2012; Batrinca et al., 2013; Scherer et al., 2012; Rosenberg and Hirschberg, 2005) and rated on 7-point Likert scales - applied more to the pre- or post-training presentation, allowing us to evaluate whether the participants' skill improved or not after training[4] :

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

2.4.3. Objective Measures

To complement the expert ratings, two annotators manually marked periods of *eye contact* with the audience and the occurrence of *pause fillers* using the annotation tool ELAN (Sloetjes and Wittenburg, 2008). We observed high inter-rater agreement for a randomly selected subset of four videos that both annotators assessed: Krippendorff α for eye contact is $\alpha = 0.751$ and pause fillers $\alpha = 0.957$ respectively (α computed on a frame-wise basis at 30 Hz).

2.4.4. Automatic Acoustic Behavior Assessment

We used the freely available COVAREP toolbox, a collaborative speech analysis repository (Degottex et al., 2014), to automatically extract audio features. COVAREP provides an extensive selection of open-source robust and tested speech processing algorithms enabling comparative and cooperative research within the speech community. All the following acoustic features are masked with voiced-unvoiced (VUV) (Drugman and Alwan, 2011), 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 extracted the following acoustic features: the maxima dispersion quotient

[3] http://www.craigslist.org/

[4] Aspect definitions and an online version of the questionnaire are available: http://tinyurl.com/ovtp67x

(MDQ) (Kane and Gobl, 2013), peak slope (PS) (Kane and Gobl, 2011), normalized amplitude quotient (NAQ) (Alku et al., 2002), the amplitude difference betzeen the first two harmonics of the differentiated glottal source spectrum (H1H2) (Titze and Sundberg, 1992), and the estimation of the Rd shape parameter of the Liljencrants-Fant glottal model (RD) (Gobl and Chasaide, 2003). Beside these features we also extracted the fundamental frequency (f0) (Drugman and Alwan, 2011) and the first two KARMA filtered formants (F1, F2) (Mehta et al., 2011). Additionally, we extracted the first four Mel-frequency cepstral coefficients (MFCC 0 - 3) and the voice intensity in dB.

2.4.5. Automatic Visual Behavior Assessment

Gestures were measured by the change of upper body joints' angles obtained from the Microsoft Kinect: we summed the differences in angles from the shoulder, elbow, hand, and wrist joints, and compared them to manually annotated gestures from 20 presentations to automatically detect gesture occurrences. We evaluated eye contact with the audience using two eye gaze estimations from the OKAO (Lao and Kawade, 2005) and CLNF (Baltrusaitis et al., 2013) softwares. We also extracted the audience eye contact ratio relative to the length of the presentation as a feature. Emotions, such as anger, sadness, and contempt, were extracted with FACET[5]. After applying the confidence provided by FACET, we extracted the mean of the emotions' intensity as another set of features.

3. Corpus Use Cases

In this section, we present these 3 studies to demonstrate the breadth of investigations on public speaking that are enabled by our multimodal corpus. Our multimodal corpus was originally created to evaluate the potential of our interactive virtual audience system for training and the impact of different feedback strategies on training efficacy: we present this study in section 3.1. We realized two additional studies using the corpus: the first one investigated the automatic assessment of public speaking ability is presented in section 3.2. The second one was focused on the automatic assessment of public speaking anxiety, and is presented in section 3.3.

3.1. Evaluation of Virtual Audience Feedback Strategies

Our original research goal was to investigate if virtual audiences can be beneficial for improving public speaking performance, and which feedback strategy provides the best improvement (Chollet et al., 2015). To that end, we had experts assess whether the study participants's performance on 10 behavioral categories (*cf.* section 2.4.2.) was better *before* training or *after* training with our virtual audience By comparing the performances of pre- and post-training presentations, we can compensate for both the presenters' level of expertise and the experts' critical opinion. Additionally, the experts were blind to the condition in which the participants had trained.

We observe that overall, all the considered performance aspects improved across all training conditions, although the effect is only moderate. The overall performance improvement was the strongest for the interactive virtual audience condition. The effect is approaching significance with $p = 0.059$ when compared to the direct feedback condition. When comparing all the assessed aspects together, the interactive virtual audience ($\mu = 0.39$, $\sigma = 0.83$; $t(298) = 0.86$, $p = 0.001$, $g = 0.395$) and control conditions ($\mu = 0.35$, $\sigma = 1.05$; $t(288) = 0.98$, $p = 0.010$, $g = 0.305$) both lead to statistically significantly better expert ratings than the direct feedback condition (*cf.* Table 1).

Aspect	Non-IVA	DF	IVA
Eye Contact	0.40 (1.37)	0.02 (1.32)	0.27 (1.27)
Body Posture	0.29 (1.12)	0.00 (1.13)	0.19 (1.12)
Flow of Speech	0.16 (1.33)	0.17 (1.25)	0.40 (1.30)
Gesture Usage	0.42 (1.39)	0.26 (1.15)	0.33 (1.24)
Intonation	0.29 (1.38)	-0.02 (1.09)	0.50 (1.35)
Confidence Level	0.33 (1.49)	0.05 (1.45)	0.44 (1.58)
Stage Usage	0.42 (1.25)	-0.12 (0.99)	0.40 (0.89)
Avoids pause fillers	0.47 (1.01)	-0.07 (0.84)	0.35 (0.76)
Presentation Structure	0.22 (1.35)	0.17 (1.38)	0.42 (1.15)
Overall Performance	0.49 (1.42)	0.05 (1.45)	0.60 (1.32)
Combined Aspects	0.35 (1.05)	0.05 (0.89)	0.39 (0.83)

Table 1: Expert Assessments. Mean values and standard deviation (in brackets) for all aspects for all three conditions, namely non-interactive virtual audience (Non-IVA), direct feedback (DF), and interactive virtual audience (IVA).

In addition, we found significant differences on some particular aspects across conditions: a significant difference is observed for the *stage usage* aspect between conditions ($F(2, 132) = 3.627$, $p = 0.029$). Stage usage improves significantly more for the interactive virtual audience condition ($\mu = 0.40$; $t(88) = 0.94$, $p = 0.011$, $g = 0.543$) and the control condition ($\mu = 0.42$; $t(85) = 1.13$, $p = 0.029$, $g = 0.473$) respectively, when compared to the direct feedback condition ($\mu = -0.12$). For the *avoids pause fillers* aspect a significant difference is observed between conditions ($F(2, 132) = 4.550$, $p = 0.012$). Participants improve significantly more on average in the interactive virtual audience condition ($\mu = 0.35$; $t(88) = 0.80$, $p = 0.013$, $g = 0.530$) and control condition ($\mu = 0.47$; $t(85) = 0.93$, $p = 0.009$, $g = 0.572$) respectively as assessed by experts, when compared to the improvement in the direct feedback condition ($\mu = -0.07$).

In conclusion, the system generally shows promise for improving presenters' public speaking skills across all investigated aspects. It seems however that direct visual feedback performed poorly compared to the other conditions. This effect can be explained in a way that the additional visual stimuli (*i.e.* color coded gauges) proved to be more of a distraction than a benefit for the participants. This finding is in line with prior findings in related work where researchers found that users' preferred sparse direct visual feedback that is only available at some instances during a presentation rather than continuously (Tanveer et al., 2015). The interactive virtual audience condition producing nonverbal feedback was not significantly better than the control condition after the investigated minimal training of

[5] http://www.emotient.com/products

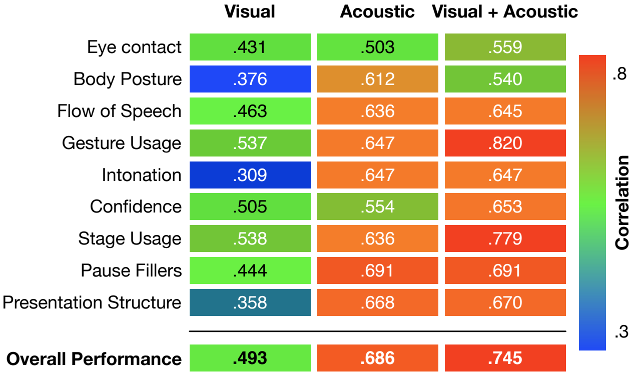
only two short presentations. However, we found using participants' questionnaires that the interactive virtual audience was perceived as more engaging ($\mu_{IVA} = 4.50$, $\mu_{Non-IVA} = 3.44$; $t(30) = 0.86, p = 0.001, g = 1.211$) and challenging ($\mu_{IVA} = 2.94$, $\mu_{Non-IVA} = 2.00$; $t(28) = 1.02$, $p = 0.025$, $g = 0.801$) than the control condition, which could prove pivotal in the long run and keep the learner engaged and present a more challenging task.

3.2. Automatic Assessment of Public Speaking Performance

In order to study how to automatically evaluate a public speaking performance, we conducted extensive unimodal and multimodal experiments and investigated regression ensemble trees to automatically predict the experts' assessments on the ten behavioral aspects presented in section 2.4. with automatically extracted audiovisual features. Full details can be found in (Wörtwein et al., 2015a). For instance, the expert 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 ($r(43) = -0.30$, $p = 0.042$). Figure 4 summarizes the observed correlation performance of our automatic performance assessment ensemble trees. 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 a baseline (the mean over all expert ratings for every aspect and all participants as a constant), the ensemble tree regression approach significantly improves baseline assessment for several aspects including overall performance: the prediction errors ($\mu = 0.55, \sigma = 0.42$) are consistently lower compared to the baseline errors ($\mu = 0.74, \sigma = 0.57$) and significantly better ($t(898) = 0.50, p < 0.001, g = -0.372$) across all aspects. Additionally, for *overall performance* alone the automatic assessment ($\mu = 0.57, \sigma = 0.46$) is also significantly better than the baseline ($\mu = 0.88, \sigma = 0.61; t(88) = 0.54, p = 0.008, g = -0.566$). For a full comparison of all aspects between our prediction errors and the constant prediction errors see (Wörtwein et al., 2015a).

Ensemble trees enable us to investigate the selected features that achieve optimal regression results, and thus 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 second and third formants, as well as the second formant's bandwidth. This shows the importance of both nonverbal and vocal 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.



	Visual	Acoustic	Visual + Acoustic
Eye contact	.431	.503	.559
Body Posture	.376	.612	.540
Flow of Speech	.463	.636	.645
Gesture Usage	.537	.647	.820
Intonation	.309	.647	.647
Confidence	.505	.554	.653
Stage Usage	.538	.636	.779
Pause Fillers	.444	.691	.691
Presentation Structure	.358	.668	.670
Overall Performance	**.493**	**.686**	**.745**

Figure 4: 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.

3.3. Prediction of Public Speaking Anxiety

In a third study, we tested whether it is possible to automatically assess public speaking anxiety with acoustic and visual features. Full details can be found in (Wörtwein et al., 2015b). First, we investigated which features correlated the most with the PRCS anxiety score. The most correlated feature from the acoustic features is the vocal expressivity measured by the standard deviation of the first formant: ARMA-filtered ($r(43) = -0.30$, $p < 0.05$), KARMA-filtered ($r(43) = -0.41$, $p < 0.01$). Additionally, the standard deviation of MFCC0 negatively correlates with the PRCS anxiety score ($r(43) = -0.36$, $p < 0.05$). Lastly, the pause time estimated by the ratio of unvoiced phonemes and voicing correlates positively with the anxiety score ($r(43) = 0.35$, $p < 0.05$). For the visual features, FACET's average facial fear expression intensity significantly correlates with the PRCS anxiety score ($r(43) = 0.41, p < 0.01$). Furthermore, both automatically extracted eye contact scores and the annotated eye contact score negatively correlate with the PRCS anxiety score: eye contact score based on CLNF ($r(43) = -0.41, p < 0.01$), based on OKAO ($r(43) - 0.54, p < 0.001$), and the annotated eye contact score ($r(43) = -0.32, p < 0.05$).

We then tried to automatically predict the PRCS anxiety score using a regression approach. We used the same method for this as for automatic performance assessment, regression ensemble trees. Regression trees were trained for unimodal audio and visual feature sets and for the multimodal features together. We found that using both acoustic and visual features ($r = 0.825, MAE = 0.118$) increased performance compared to using visual features only ($r = 0.640, MAE = 0.154$) or audio features only ($r = 0.653, MAE = 0.148$) both with respect to mean absolute error (MAE) and Pearson's correlation. The features selected by the multimodal regression ensemble tree are comprised of closely related features such as express-

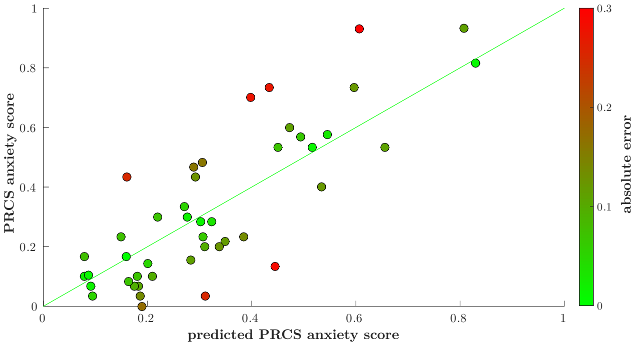


Figure 5: Scatter plot of the predicted PRCS anxiety score against its ground truth.

ing sadness, being more quiet, speaking at a slower pace, gesturing less, and a lack of orientation towards to the audience.

In addition to comparisons between modalities, we compared our automatic assessment errors with the error of a constant mean prediction. Our multimodal prediction ($\mu = 0.12$, $\sigma = 0.09$) is significantly more accurate than the constant prediction ($\mu = 0.21, \sigma = 0.14; t(88) = 0.11, p < 0.001, g = -0.840$). Figure 5 shows a scatter plot of the ground truth against the predicted anxiety with both modalities.

4. Future Research Directions

Our current work is focused on overhauling the system into a fully automatic public speaking training framework. In the study we presented, the assessment of participants' behavior during training was done by a confederate using a *wizard of Oz* paradigm. In the future, the assessment of participants' performance will be done automatically by machine learning trained on our multimodal corpus, using audiovisual features extracted in real-time. Additionally, our virtual audience is being improved: more varied character models have been introduced and more animations and behaviors are being added to allow the audience to display various states such as boredom, engagement or disagreement (*cf.* Figure 6). We plan to use the new system to investigate how training with it improves public speaking ability over time. We will perform a longitudinal study in which participants will train several times over a few weeks, and then be evaluated in an actual public speaking presentation in front of the other participants. This will allow us to measure if training with a virtual audience transfers into improvement of public speaking ability in real situations.

Additionally, this multimodal corpus will be growing as we perform more studies related to studying public speaking. We present here current extension work on the corpus.

4.1. Non-native English Speakers Dataset

Performing a public speaking task is more challenging in a foreign language than in one's mother tongue. We are interested in studying how this impacts the different behavioral aspects of public speaking, and whether our interactive virtual audience system can also help non-native speakers. To this end, we recruited 13 subjects originating from various non-English speaking countries (Japan, China, Portugal, *etc.*) with an intermediate level in English, and had them participate in a study following the protocol we presented in section 2.2., in the interactive virtual audience condition. The resulting videos are being annotated and processed, and will be compared with our data of native speakers. In particular, we will investigate whether non-native participants benefit as much from the training as native participants.

4.2. Crowdsourced Measures of Public Speaking Performance

We are collecting ratings of public speaking performance of our data by using Amazon's Mechanical Turk[6] crowdsourcing tool to obtain laypersons' evaluations of the participants' performance. We intend to compare the experts' assessments with crowdsourced assessments to investigate whether the "wisdom of the crowd" is in agreement with experts' opinions. Additionally, we will compare the improvement scores obtained with our system when pre- and post-training videos are evaluated side by side (comparison rating) to an evaluation where videos are rated individually by annotators (absolute rating).

4.3. Transcriptions and Verbal Behavior Analysis

An important aspect of successful public speaking is shown in the verbal behaviors of the speakers. While the initial automatic analysis focused on nonverbal behaviors, a key component of the research surrounding this dataset will be around the study of language during public speaking. We plan to explore both manual transcriptions as well as automatic speech recognition.

5. Conclusion

In this paper, we presented a multimodal corpus of public speaking presentations. This corpus was collected using an interactive virtual audience system that gave feedback to the participants regarding their performance. We recorded 4 presentations by 45 participants. Half of these 180 presentations (pre- and post-training presentations) have been rated by experts and annotated for two behaviors, gaze and pause fillers. Additionally, multimodal features were automatically extracted, and questionnaires were collected about demographics, assessment of the system, personality and public speaking anxiety. We outlined three studies realized with this corpus two automatically assess public speaking ability and anxiety, and to evaluate its potential for public speaking training. Our corpus will be made available upon request for academic research.

6. Acknowledgments

We would like to thank the Toastmasters group Funny Bones of Culver City, CA for their engagement and their helpful comments and discussions. This material is based upon work supported by the National Science Foundation under Grants No. IIS-1421330 and No. IIS-1118018 and U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Any opinions, findings, and

[6] https://www.mturk.com

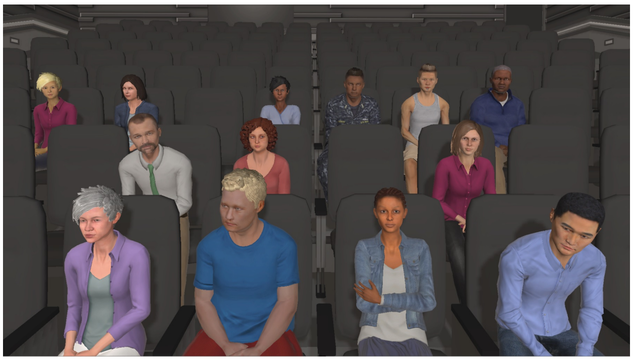


Figure 6: Overhauled version of our virtual audience system.

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.

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

Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 354–361. IEEE.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. (2013). Cicero - towards a multimodal virtual audience platform for public speaking training. In *Intelligent Virtual Agents*, page 116–128.

Ben Youssef, A., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., and Ochs, M. (2015). Towards a socially adaptive virtual agent. In Willem-Paul Brinkman, et al., editors, *Intelligent Virtual Agents*, volume 9238 of *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing.

Bodie, G. D. (2010). A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education*, 59(1):70–105.

Brilman, M. and Scherer, S. (2015). A multimodal predictive model of successful debaters or how i learned to sway votes. In *to appear in Proceedings of ACM Multimedia 2015*.

Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A., and Scherer, S. (2015). Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1143–1154, New York, NY, USA. ACM.

Crawford, J. R. and Henry, J. D. (2004). The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43(3):245–265.

Damian, I., Tan, C. S. S., Baur, T., Schöning, J., Luyten, K., and André, E. (2015). Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 565–574, New York, NY, USA. ACM.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). Covarep - a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964, May.

Drugman, T. and Alwan, A. (2011). Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976.

Furmark, T., Tillfors, M., Stattin, H., Ekselius, L., and Fredrikson, M. (2000). Social phobia subtypes in the general population revealed by cluster analysis. *Psychological Medicine*, 30(6):1335–1344.

Gobl, C. and Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212.

Hook, J., Smith, C., and Valentiner, D. (2008). A short-form of the personal report of confidence as a speaker. *Personality and Individual Differences*, 44(6):1306–1313.

Hoque, M., Courgeon, M., Martin, J.-., Mutlu, B., and Picard, R. (2013). Mach: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 697–706, New York, NY, USA. ACM.

Jennett, C., Cox, A., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., and Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66:641–661.

Kane, J. and Gobl, C. (2011). Identifying regions of non-modal phonation using features of the wavelet transform. In *INTERSPEECH*, pages 177–180.

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

Lane, C., Hays, J., Core, M., and Auerbach, D. (2013). Learning intercultural communication skills with virtual humans: Feedback and fidelity. *Journal of Educational Psychology*, 105(4):1026–1035, November.

Lao, S. and Kawade, M. (2005). Vision-based face understanding technologies and their applications. In S.. Li, et al., editors, *Advances in Biometric Person Authentication*, volume 3338 of *Lecture Notes in Computer Science*, pages 339–348. Springer Berlin Heidelberg.

Mehta, D., Rudoy, D., and Wolfe, P. (2011). Karma: Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *arXiv preprint arXiv:1107.0076*.

Nguyen, L., Marcos-Ramiro, A., R., M., and Gatica-Perez, D. (2013). Multimodal analysis of body communication cues in employment interviews. In *Proceedings of the 15th ACM on International Conference on Multimodal*

Interaction, ICMI '13, pages 437–444, New York, NY, USA. ACM.

North, M., North, S., and Coble, J. (1998). Virtual reality therapy: An effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, 3:2–6.

Paul, G. (1966). *Insight vs. Desensitization in Psychotherapy: An Experiment in Anxiety Reduction*. Stanford University Press.

Pertaub, D., Slater, M., and Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments*, 11:68–78.

Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

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

Scherer, S., Layher, G., Kane, J., Neumann, H., and Campbell, N. (2012). An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1114–1120. ELRA.

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

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: Elan and iso dcr. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Tanveer, M., Lin, E., and Hoque, M. E. (2015). Rhema: A real-time in-situ intelligent interface to help people with public speaking,. In *ACM International Conference on Intelligent User Interfaces (IUI)*.

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

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., and Scherer, S. (2015a). Multimodal Public Speaking Performance Assessment. In *to appear in Proceedings of ICMI 2015*, Seattle, WA, USA, October. ACM.

Wörtwein, T., Morency, L.-P., and Scherer, S. (2015b). Automatic Assessment and Analysis of Public Speaking Anxiety: A Virtual Audience Case Study. In *Proceedings of ACII 2015*, Xi'an, China, September. IEEE.