# Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training

Ligia Batrinca[1], Giota Stratou[2], Ari Shapiro[2], Louis-Philippe Morency[2], and Stefan Scherer[2]

[1] FBK-IRST, Via Sommarive 18, 38050 Trento, Italy.
batrinca@fbk.eu
[2] University of Southern California
Institute for Creative Technologies, Los Angeles, California
stratou|shapiro|morency|scherer|@ict.usc.edu

**Abstract.** Public speaking performances are not only characterized by the presentation of the content, but also by the presenters' nonverbal behavior, such as gestures, tone of voice, vocal variety, and facial expressions. Within this work, we seek to identify automatic nonverbal behavior descriptors that correlate with expert-assessments of behaviors characteristic of good and bad public speaking performances. We present a novel multimodal corpus recorded with a virtual audience public speaking training platform. Lastly, we utilize the behavior descriptors to automatically approximate the overall assessment of the performance using support vector regression in a speaker-independent experiment and yield promising results approaching human performance.

**Keywords:** Virtual Reality, Behavioral Modification, Multimodal Perception, Public Speaking, Training

## 1 Introduction

Public speaking is an essential skill for a large variety of professions and in everyday life. The quality of a presentation can greatly influences the presenters career development or the likelihood to close a deal. However, public speaking itself is not a skill that is innate to everyone, but can be mastered through extensive training[3]. Further, mild forms of public speaking anxiety may be controlled via frequent exposure to presentation scenarios (even virtual ones) [10, 11]. The best form of training often is to present in familiar and forgiving environments and by receiving the audiences feedback during and after the presentation. Audiences provide indirect feedback during presentations by signaling nonverbal feedback, while they continuously rate and sense the presenters speaking style. While audiences show signs of high attention (e.g. mutual gaze or forward leaning posture) and cues of rapport (e.g. nodding or smiling) in presentations they enjoy, they often show no interest (e.g. averted gaze or lack of backchannel behavior) or

[3] http://www.toastmasters.org/tips.asp

disagreement otherwise. By consciously perceiving and adapting ones speaking style with respect to these feedback behaviors the presenter can greatly improve.

Unfortunately, often no human audience is available or the fear of presenting in front of a human audience is difficult. The present work provides preliminary steps towards an artificial virtual audience capable of providing such nonverbal feedback during presentations, based on the perceived multimodal presenter's speaking style. Such a virtual audience would be available at any time and could help improve public speaking skills in an efficient and non-threatening way. Within this paper, we present our prototype virtual human platform for public speaking training, called Cicero. In the future, Cicero will serve as such a virtual audience that will provide users with helpful feedback.

We analyze a preliminary dataset of 14 subjects giving a 5-15 minute presentation in front of a virtual audience. Public speaking experts from the worldwide organization of Toastmasters, assess both public speaking related behaviors and observations and estimate the presenters' overall performance in a viewing study. We then correlate automatically observed multimodal nonverbal behaviors with expert assessments of the assessed behaviors and try to automatically approximate the experts' overall assessment of the presenters' performance in a speaker-independent regression task. In particular, our research goals for the present work are:

**R1 Expert Assessment:** We aim to identify expert estimates of nonverbal behaviors, including flow of speech, clarity of intonation, correct use of gestures, and gaze patterns, that correlate with the experts' overall assessment of the presenter's performance.

**R2 Automatic Behavior Descriptors:** We seek to identify basic automatic multimodal behavior descriptors that strongly correlate with the experts' assessment of the presenters' audiovisual nonverbal behavior. These automatic measures are extracted from three independent sensors and comprise basic estimates of speech characteristics, gestures, and gaze.

**R3 Automatic Performance Assessment:** We further estimate the presenters' performance in a preliminary presenter-independent classification experiment using the automatically estimated nonverbal behaviors as input for the support vector regression.

The remainder of this paper is organized as follows: Section 2 discusses some related work on virtual audiences and speaking performance. Section 3 then introduces our experimental setup, the investigated dataset. In Section 4, we discuss the details of the expert assessment study. The automatic behavior descriptors using audiovisual information are then introduced and their correlations with expert opinions are studied in Section 5. In Section 5.4, we investigate how well the automatic behavior descriptors can be used to approximate the expert assessments. In Section 6, we discuss our findings and outline future paths. Finally, Section 7 concludes the paper.

## 2 Related Work

**Virtual Audiences.** Virtual audiences have been investigated in the past to treat public speaking anxiety. One of the first works on virtual reality (VR) used to treat public speaking anxiety was done by [10]. This study suggested that VR could indeed be useful in treating public speaking anxiety. At the end of the study, self-reported levels of anxiety were reduced. In a study by [11] participants were asked to give a 5 minutes long presentation to 3 different types of audiences: a neutral, a positive, and a negative audience. The virtual audience consisted of 8 virtual characters. The study showed that all three settings mentioned above, have an influence on the subject, generating anxiety in participants which scored high on the Personal Report of Confidence as a Public Speaker (PRCS). In the same year, another study, by [3], focused on university students with prominent public speaking anxiety. One group was exposed to Virtual reality Exposure Therapy (VRET) while another group were put in a wait-list control group. The results of this study are in line with previous findings: virtual reality treatment sessions are effective in reducing public speaking anxiety.

Although lately, more researchers have become aware of the importance and effectiveness of VR in treating anxiety-like behaviors when holding a talk in front of an audience, to the best of our knowledge this is the first study to directly address the scenario of giving a presentation in front of a virtual audience. Moreover, using non-invasive, state-of-the-art sensing technology capturing the presenters' nonverbal behavior patterns. Additionally, in this work we primarily focus on the quality of the performance itself rather than investigating possible treatment strategies of public speaking anxiety.

**Public Speaking Performance.** 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. In [17] 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. Further, manual annotations of disfluencies were identified to be negatively correlated with a positive rating.

In [13], the acoustic feature set, used in [17], 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.

The authors of [8] investigate more complex motion features, such as hand trajectories and identify correlates of gestures with ratings of personality. Again, they extract these motion features from videos of German politicians and present them as stick-figure representations to the raters. In all three above studies, non-experts assessed the performances of professional speakers (i.e. politicians), within this work we want to investigate features that are present in presentation performances by the general population and potentially untrained speakers. We further ask experts to rate their performances not only with respect to an overall assessment, but by utilizing a more fine grained questionnaire that disseminates the behaviors into multiple ratings.

## 3 Experimental Design and Dataset

In the following we provide details regarding the experimental setup in which the study took place and some details on how we setup the virtual audience for the experiments. Additionally, we detail the participant recruitment and the experimental procedure.

### 3.1 Experimental Setting

Figure 1 illustrates the room setup used in the study. As it can be seen, the lab was arranged to resemble a conference room. The experimenter initializes the virtual audience that is projected on the virtual audience screen. The characters approach life-size measures. The participant controls the presentation, which is projected on the presentation screen, with the help of a standard Logitech remote control. The nonverbal behaviors of the participant are captured using a Microsoft Kinect and 2 off-the-shelf webcams, mounted in front of the participant and to the side (cf. Figure 1). Acoustic information was collected using a lapel microphone.

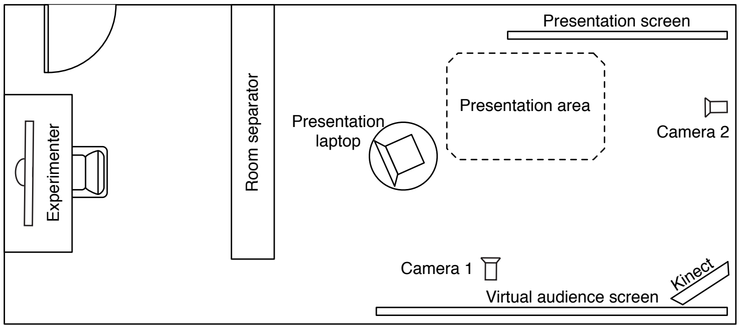

**Fig. 1.** Experimental setting of the Virtual Human Lab

### 3.2 The Virtual Audience

A high-performance desktop computer was used to project the virtual audience, animated using SmartBody character animation system [16] and VHToolkit [7]. In Figure  2 a snapshot of the virtual audience is provided. Each virtual human was able to change its posture (e.g. straight, relaxed, forward), head orientation (e.g. up, left, right, front) or eye-gaze. Eye-contact with the speaker was simulated by having the characters looking at the speaker. This enabled us to create an audience designed to give the impression of a real life audience. Within our study the virtual audience was modeled to display interest in the participant's presentation, which was accomplished by the proper combination of posture, head-orientation and eye-gaze directionality. It is important to mention that during the study, the attentive behavior of the virtual characters was not static and the audience remained attentive and lively.

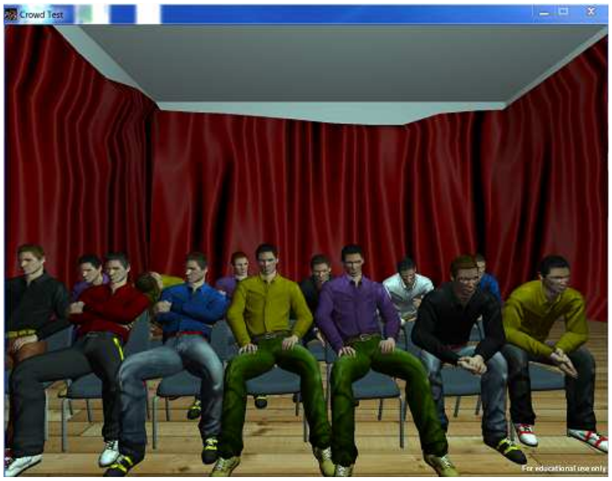

**Fig. 2.** Snapshot of the virtual audience

### 3.3 Participants and Experimental Procedure

In our study we recorded 14 participants, 13 of which were recruited from craigslist and one participant was recruited at a university. The data set consisted of 11 females and 3 males, with an average age of 39 and standard deviation of 14.34. The participants were provided with two different presentations to choose from, three days in advance of the scheduled presentation. They were instructed to rehearse, before giving the presentation at our lab, as they would normally prepare for an important presentation.

Prior to the presentation, the participants filled in a series of questionnaires including a brief demographic assessment, the Big-Five Inventory short form

| Source | Assessed behavior | Spearman's $\rho$ | p-value |
|---|---|---|---|
| Voice | Flow of speech | **.477** | .010 |
| | Clear intonation | **.436** | .021 |
| | Interrupted speech | .016 | .933 |
| | Speaks too quietly | -.363 | .057 |
| | Vocal variety | **.471** | .013 |
| Body | Paces too much | **.599** | $< .001$ |
| | Gestures to emphasize | .354 | .065 |
| | Gestures to much | -.062 | .764 |
| Gaze | Gazes at audience | .166 | .398 |
| | Avoids audience | -.358 | .062 |

**Table 1.** Summarizes the correlations between the expert assessed behaviors from three sources (i.e. voice, body, and gaze) with the experts' opinion of the presenters' overall performance. Spearman's $\rho$ values are reported, along with p-values of test if estimated correlation is significantly different from no correlation.

(BFI SF)[12], the Personal Report of Confidence as a Public Speaker (PRCS)[9], and the Self-Statements During Public Speaking (SSPS) [4]. Immediately before the start of the experiment, the participants were introduced to the experimental setting of the lab. They were instructed not to look directly into the cameras, but at the audience. Markers on the floor were provided to give a guidance to the presenters where to stand while giving the presentation, to ensure optimal viewing angles for the cameras.

Participants then filled in post-session questionnaires, including Positive and Negative Affect Schedule (PANAS) [19], and a modified version of [21] virtual audience presence questionnaire and performance self-assessment questionnaires. At the end the participants were debriefed and received 25 USD for their efforts.

## 4 Expert Assessment

In order to obtain independent expert opinions on the participants' performances, we invited two senior Toastmasters, i.e. members of a worldwide organization devoted to improve speaking skills through exercise and critique. Both experts assessed their own experience level with the highest possible value on a seven point Likert scale. Additionally, both feel very comfortable presenting themselves and performed 10 or more times in the last two years. Lastly, they estimate their own public speaking skill to be clearly above average.

The experts viewed the presentations using the frontal camera view with the audio from the lapel microphone. They viewed each presentation only once and assessed the performances of the participants using two sets of questions, all of which were answered on a seven point Likert scale. The first set consists of typical behaviors and observable characteristics of public speaking performances and comprise assessments of the flow of speech, the presenter's pacing behavior on the stage, the posture's stiffness, the presenter's nervousness, and the observed amount of eye contact with the virtual audience. In total we assessed 21 characteristics for each speaker, a subset of these are presented in Table 1. Additionally, we assess the experts' perception of the overall performance. All expert annotations are z-score normalized in order to remove perception biases. The inter-expert agreement on the overall assessment results in Krippendorff $\alpha$ = .715, which corresponds to considerable agreement. The correlation between the overall rating for the performances is with Spearman's $\rho = .648$ quite strong and significantly different from zero with p = .012.

Table 1 summarizes the estimated correlations of some of the assessed behaviors and characteristics with the overall estimated presenters' performance.

## 5 Automatic Behavior Descriptors

Similarly to the above evaluation of correlations between assessed behaviors and the overall performance, we investigate automatic behavior descriptor-correlations with the expert-assessed behaviors and characteristics listed in Table 1. The behavior descriptors are automatically extracted from three audiovisual sensory inputs using the multimodal sensor fusion framework called *MultiSense* [15, 14]. MultiSense is a flexible framework that is based on the Social Signal Interpretation framework (SSI) by [20] and it is created as a platform to integrate and fuse sensor technologies and develop probabilistic models for human behavior recognition. The modular setup of MultiSense allows us to integrate multiple sensing technologies for this analysis. We detail the extracted behavior descriptors in sections 5.1 and 5.2. The results of the correlation analysis are reported in Section 5.3 and Table 2. The automatic overall performance assessment evaluation is provided in Section 5.4.

### 5.1 Acoustic nonverbal behavior descriptors

Using the lapel microphone recordings, we extracted several basic acoustic and prosodic features. The features are extracted with a sample rate of 100 Hz. Hesitations and pause fillers were counted by one of the experts and noted on the evaluation sheet for each presenter. The following sections detail each acoustic feature.

**Energy in dB.** The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). This speech window $w(t)$ is filtered with a hamming window and the energy

$$e(t) = \sum_{i=1}^{|w(t)|} w_i(t)^2 \tag{1}$$

| Source | Assessed behavior | Behavior descriptor | Spearman's $\rho$ | p-value |
|---|---|---|---|---|
| Voice | Flow of speech | Num. pauses | -.469 | .09 |
| | Clear intonation | Avg. intensity | **.805** | .002 |
| | | Breathiness | **-.615** | .033 |
| | Interrupted speech | Num. pause fillers | **.612** | .034 |
| | Speaks too quietly | Avg. intensity | **-.842** | $<$ .001 |
| | Vocal variety | Std. $f_0$ | **.709** | .010 |
| | | Spectral Stationarity | **-.586** | .045 |
| Body | Paces too much | Leg movement | **.682** | .021 |
| | Gestures to emphasize | Arm movement | **.710** | .014 |
| | Gestures to much | Arm movement | .437 | .179 |
| Gaze | Gazes at audience | Face gaze towards | **.621** | .030 |
| | Avoids audience | Face gaze towards | -.548 | .065 |

**Table 2.** Summarizes the correlations between the expert assessed behaviors from three sources (i.e. voice, skeleton, and gaze) with automatic behavior descriptors extracted from the audiovisual data. Spearman's $\rho$ values are reported, along with p-values of test if estimated correlation is significantly different from no correlation.

is calculated and converted to the dB-scale

$$e_{dB}(t) = 10 \cdot \log_{10}(e(t)). \tag{2}$$

**Fundamental frequency $f_0$.** In [2], a method for $f_0$ tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process removes strong influences of noise and vocal tract resonances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, showing peaks for the harmonics of $f_0$, the fundamental frequency. Then, the summation of residual harmonics (SRH) is computed as follows [2]:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \tag{3}$$

for $f \in [f_{0,\mathrm{min}}, f_{0,\mathrm{max}}]$, with $f_{0,min} = 50$ and $f_{0,max} = 300$. The frequency $f$ for which $SRH(f)$ is maximal is considered the fundamental frequency of this frame. By using a simple threshold $\theta$, the unvoiced frames are discarded as in [2].

**Pause timings.** Pauses were considered as continuous segments of at least 300 ms in length with a signal strength of at least 25 dB below the $99^{th}$ percentile of the recording. This implementation follows the same parameter setting and recommendations as in the standard Praat pause detection algorithm [1].

**Spectral stationarity.** To characterize the range of the prosodic inventory used over utterances, we make use of the so called *spectral stationarity* measure $ss$. This measurement was previously used in [18] as a way of modulating the transition cost used in the dynamic programming method used for $f_0$ tracking. Spectral stationarity, $ss$ is measured with:

$$ss = \frac{0.2}{\text{itakura}(f_i, f_{i-k}) - 0.8} \quad \in [0,1], \tag{4}$$

where itakura(.) is the Itakura distortion measure [5] of the current speech frame $f_i$ and $f_{i-k}$ is the previous frame with $k = 1$. We use a relatively long frame length of 60 ms (with as shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function before measuring $ss$. The long frame length was used in the attempt to characterize relatively long periods of maintained vocal tract articulation. $ss$ is close to 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference.

**Voice tenseness measured by $\mathbf{OQ}_{NN}$.** In order to characterize the tenseness of the speaker's voice, we extract $\mathbf{OQ}_{NN}$ a novel parameter estimating the open quotient using standard Mel frequency cepstral coefficients and a trained neural network for open quotient approximation [6].

### 5.2 Visual nonverbal behavior descriptors

Visual behavior descriptors were extracted from the tracked skeleton and face using information provided by the Kinect sensor and the frontal web-camera. Measures were extracted using a sample rate of 30 Hz. The following sections detail each visual feature.

**Arm and Leg movement.** Based on the tracked skeletal information we calculate an overall intensity measure of the arm and leg movement respectively. We calculate movement by computing simple distances between consecutive frames of the tracked skeletal joints. These distances are summed up for the respective group (i.e. legs and arms) and normalized by the total length of the presentation.

**Face gaze towards.** We utilize the tracked face direction to assess the presenters' gaze. We track if the presenter looks towards the screen on which we present the virtual audience using the frontal webcam placed on a tripod (100 cm in height) facing the presenter. We consider a relatively wide range of degrees as facing towards the audience (i.e. +/- 45 degrees) as the audience is fairly large and close to the presenter. Additionally, we track the vertical gaze direction and consider angles above zero degrees as gazing towards the audience. Angles below zero are considered as looking at hand-held notes or the floor. We measure the gaze towards the audience as a ratio of the overall total duration of the presentation.

### 5.3 Automatic Behavior Descriptor Correlations

Here, we report results of the correlation analysis between automatic behavior descriptors and expert assessments of presenters' characteristics. We calculate Spearman's $\rho$ for each behavior descriptor with the associated expert-assessed behavior and report p-values indicating if the observed correlation is significantly different from zero.

As seen in Table 2, we could identify a large number of basic behavior descriptors that correlate significantly with expert assessments for all investigated modalities. Based on the voice descriptors we could identify that the average speech intensity is highly correlated with the "clear intonation" assessments ($\rho$ = .805; p = .002) and negatively correlated with the "speaks too quietly" assessment ($\rho$ = -.842; p < .001). Further, the breathiness as observed with higher values of $\mathrm{OQ}_{NN}$ is negatively correlated with clear intonation ($\rho$ = -.615; p = .033). Vocal variety is correlated with the standard deviation of $f_0$ ($\rho$ = .709; p = .010) and negatively correlated with the monotonicity measure spectral stationarity ($\rho$ = -.586; p = .045).

Based on the skeletal information, we can identify if the presenter is pacing too much on stage by using the leg movement descriptor ($\rho$ = .682; p = .021). Additionally, arm movement is correlated with the experts' assessment if the presenter uses gestures to emphasize points of the presentation appropriately ($\rho$ = .710; p = .014). Lastly, the "gazes at audience" assessment is correlated with the automatic behavior descriptor face gaze towards ($\rho$ = .621; p = .030).

We are aware of the fact that a statistical correction for multiple testing would be required at this point. However, with the relatively small sample size this would require extremely high correlations $|\rho| \geq .800$ for each individual behavior. In order to address this issue from another direction, we chose to conduct a sanity-check regression analysis with a leave-one-presenter-out testing paradigm to show the relevance of the observed nonverbal behaviors in the following section.

### 5.4 Automatic Performance Assessment

Based on the above findings and automatic behavior descriptors, we conduct a presenter-independent approximation experiment. We use simple support vector regression with a polynomial kernel of degree three and in total eight features as input (i.e. five voice features, two skeleton features, and one gaze feature). With a leave one presenter out testing paradigm we achieve an overall absolute mean error of .660 with a standard deviation of .540. The automatically approximated performance assessment corresponds with the experts' mean overall assessment with $\rho$ = .617 and p = .025.

## 6 Discussion

Based on expert assessments of a small number of presentations given to a virtual audience, we could identify several characteristic nonverbal behaviors that correlate positively or negatively with the overall perceived presenters' performance.

All three investigated modalities (i.e. voice, skeleton, and gaze) contribute to the assessment and Table 1 summarizes these findings. It is interesting to note, that some behaviors that anecdotally are associated with bad performances did not show any correlation with the overall assessment, such as interrupted speech or excessive gesturing. We believe that these behaviors might be outweighed by others and a more fine-grained overall performance estimation disseminating spoken, gestural, expressive, and structural quality might be required.

When approximating the expert-assessed nonverbal behaviors automatically, we could identify a number of basic behavior descriptors, such as the average intensity, overall leg movement, and gaze statistics, that are highly correlated with expert assessments (cf. Table 2). While, these basic descriptors achieved promising results using support vector regression in a speaker-independent experiment (cf. Section 5.4), they remain crude and on an abstract level. For example, the overall arm movement is correlated with appropriate gestural emphasis of arguments within a presentation, which would intuitively at least require knowledge about the arm gestures and the content of the spoken words. Hence, we plan to investigate multimodal information fusion to capture more meaningful and sophisticated measures of public speaking performances.

For future work, we additionally plan to investigate optimal ways of conveying the perceived information on the performance to the presenters. We will analyze ad-hoc visualizations, such as audience reactions or visual overlays, as well as post-hoc summaries and quantitative evaluations with typical statistical plots. We plan to base the audience's behavior on the presenter's automatically estimated performance to provide realtime feedback to the presenter. Here, we envision both subtle movements in the audience to create a more life-like and immersive experience for the presenter and more striking and interruptive behaviors to directly reflect the potential discontent or approval of the presenter's performance. The audience could for example show reduced interest in the presentation due to the lack of vocal variety in the presenter's voice. At present, we focused our analysis on nonverbal behaviors only and will expand the analysis to verbal contents in the future. Further, we will investigate usability and effectivity of different strategies, with respect to performance improvement and immersion.

## 7 Conclusions

This paper presents a proof-of-concept (and at present non-interactive) version of the research platform for public speaking training, called Cicero. Based on our research goals, stated in Section 1, we could identify the following main findings in this work: **R1** we reveal several expert estimates of nonverbal behaviors, such as flow of speech, vocal variety, or avoided eye contact with the audience, to be significantly correlated with an overall assessment of a presenter's performance; **R2** using multimodal information from three sensors we could identify automatic behavior descriptors that correlate strongly with expert estimates of nonverbal behaviors, comprising estimates for a clear intonation, vocal variety, pacing around, and eye contact with the audience. Lastly, **R3** we automatically approximate the experts' overall performance assessment with a mean error of .660 on a seven point scale. Further, the automatic approximation using support vector regression correlates significantly with the experts' opinion with Spearman's $\rho = .617$ (p = .025), which approaches the correlation between the experts' opinions (i.e. $\rho = .648$). Motivated by these promising results, we plan to expand the presented research platform Cicero in the near future to incorporate a more diverse and reactive virtual audience. Cicero will enable us to conduct a wide variety of experiments reaching from performance assessments to psychological experiments, which would not be possible with a real human audience.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army RDECOM. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We would like to thank our close collaborators Yuyu Xu, Teresa Dey, Margaux Lhommet, and Stacy Marsella for their help and advice while running this study. Additionally, we would like to thank the Toastmasters group Funny Bones of Culver City for their kind support in assessing the performance of our study participants.

## References

1. P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9):341–345, 2001.
2. T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976. ISCA, 2011.
3. S. R. Harris, R. L. Kemmerling, and M. M. North. Brief virtual reality therapy for public speaking anxiety. *Cyberpsychology and Behavior*, 5:543–550, 2002.
4. S. G. Hofmann and P. M. DiBartolo. An instrument to assess self-statements during public speaking: Scale development and preliminary psychometric properties. *Journal of Behavior Therapy*, pages 499–515, 2000.
5. F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:67–72, 1975.
6. J. Kane, S. Scherer, L.-P. Morency, and C. Gobl. A comparative study of glottal open quotient estimation techniques. In *to appear in Proceedings of Interspeech 2013*. ISCA, 2013.
7. P. Kenny, A. Hartholt, J. Gratch, W. Swartout, D. Traum, S. Marsella, and D. Piepol. Building interactive virtual humans for training environments. In *Proceedings of I/ITSEC*, 2007.
8. M. Koppensteiner and K. Grammer. Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, 44:374–379, 2010.

9. J. C. McCroskey. Measures of communication-bound anxiety. *Speech Monographs*, 37:269–277, 1970.
10. M. M. North, North S. M., and J. R. Coble. Virtual reality therapy: An effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, 3:2–6, 1998.
11. 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*, 11:68–78, 2002.
12. B. Rammstedt and O. P. John. 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*, 41:203–212, 2007.
13. 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 (LREC'12)*, pages 1114–1120. ELRA, 2012.
14. S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *Proceedings of Intelligent Virtual Agents (IVA'12)*, LNAI 7502, pages 455–463. Springer, 2012.
15. S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*. IEEE, 2013.
16. A Shapiro. Building a character animation system. In *Motion in Games*, pages 98–109. Springer, 2011.
17. 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. ISCA, 2008.
18. David Talkin. A Robust Algorithm for Pitch Tracking. In W. B. Kleijn and K. K. Paliwal, editors, *Speech coding and synthesis*, pages 495–517. Elsevier, 1995.
19. E. R. Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of Cross-Cultural Psychology*, 38(2):227–242, 2007.
20. J. Wagner, F. Lingenfelser, N. Bee, and E. André. Social signal interpretation (ssi). *KI - Kuenstliche Intelligenz*, 25:251–256, 2011. 10.1007/s13218-011-0115-x.
21. B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240, 1998.