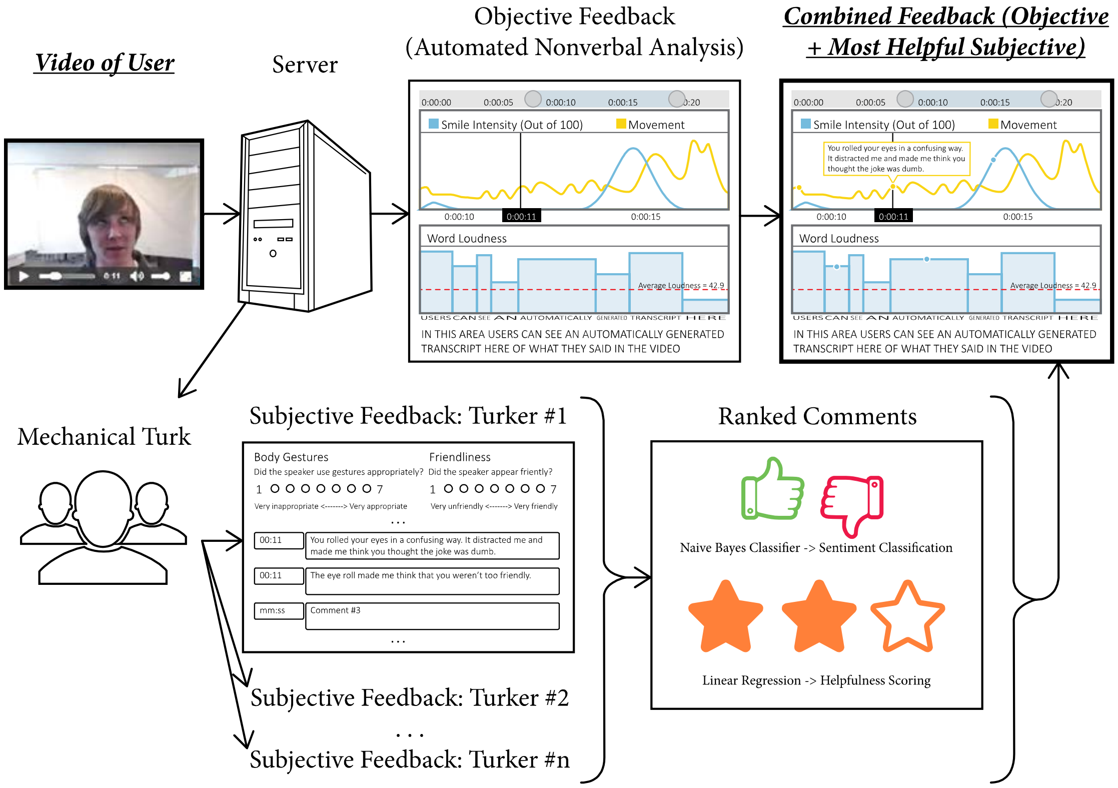
ROC Speak: Semi-Automated Personalized Feedback on Nonverbal Behavior from Recorded Videos

Michelle Fung, Yina Jin, RuJie Zhao, Mohammed (Ehsan) Hoque
Rochester Human-Computer Interaction (ROC HCI), University of Rochester, NY
{mfung, yjin18, rzhao2, mehoque}@cs.rochester.edu


Video of User
Server
Objective Feedback
(Automated Nonverbal Analysis)
Combined Feedback (Objective
+ Most Helpful Subjective)
0:00:00
0:00:05
0:00:10
0:00:15
0:20
Smile Intensity (Out of 100)
Movement
0:00:10
0:00:11
0:00:15
Word Loudness
Average Loudness = 42.9
USERS CAN SEE AN AUTOMATICALLY GENERATED TRANSCRIPT HERE
Mechanical Turk
Subjective Feedback: Turker #1
Subjective Feedback: Turker #2
. . .
Subjective Feedback: Turker #n
Ranked Comments
Naive Bayes Classifier -> Sentiment Classification
Linear Regression -> Helpfulness Scoring

IN THIS AREA USERS CAN SEE AN AUTOMATICALLY GENERATED TRANSCRIPT HERE OF WHAT THEY SAID IN THE VIDEO

You rolled your eyes in a confusing way. It distracted me and made me think you thought the joke was dumb.

Body Gestures
Did the speaker use gestures appropriately?
1 ○ ○ ○ ○ ○ ○ ○ 7
Very inappropriate <-------> Very appropriate

Friendliness
Did the speaker appear frientl y?
1 ○ ○ ○ ○ ○ ○ ○ 7
Very unfriendly <-------> Very friendly

. . .

00:11 You rolled your eyes in a confusing way. It distracted me and made me think you thought the joke was dumb.

00:11 The eye roll made me think that you weren't too friendly.

mm:ss Comment #3

. . .

Figure 1. An overview of our system. Once the user finishes recording, the video is analyzed on the server for objective feedback and sent to Mechanical Turk for subjective feedback. The objective feedback is then combined with subjective feedback that is scored based on helpfulness, under which the sentiment is then classified.

ABSTRACT

We present a framework that couples computer algorithms with human intelligence in order to automatically sense and interpret nonverbal behavior. The framework is cloud-enabled and ubiquitously available via a web browser, and has been validated in the context of public speaking. The system automatically captures audio and video data in-browser through the user's webcam, and then analyzes the data for smiles, movement, and volume modulation. Our framework allows users to opt in and receive subjective feedback from Mechanical Turk workers ("Turkers"). Our system synthesizes the Turkers' interpretations, ratings, and comment rankings with the machine-sensed data and enables users to interact with, explore, and visualize personalized and presentational feedback. Our results provide quantitative and qualitative evidence in support of our proposed synthesized feedback, relative to video-only playback with impersonal tips. Our interface can be seen here: **http://tinyurl.com/feedback-ui** (Supported in Google Chrome.)

Author Keywords

Nonverbal behavior interpretation, automated multimodal affect sensing, automated feedback, public speaking.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces – *input devices and strategies, evaluation/*

UbiComp '15, September 7-11, 2015, Osaka, Japan.
Copyright 2015 © ACM 978-1-4503-3574-4/15/09…$15.00.
http://dx.doi.org/10.1145/2750858.2804265

methodology, user-centered design. H.1.2 Models and Principles: User/Machine Systems – *human factors, software psychology*

INTRODUCTION

Automated modeling of the full range of human nonverbal behavior remains a challenging endeavor. Solely by using the 43 muscles in our face, we can produce over 10,000 unique combinations of facial expressions. Modalities like vocal tone, body language, and physiological elements add to the complexity. While research has progressed to recognize basic expressions like smiling and frowning [1], the automated interpretation of human expressions remains an active area of exploration. For example, a smiling customer does not necessarily indicate that he or she is satisfied [2]. This demonstrates a current limitation on the utility of technology built to detect human expressions.

In this paper, we introduce a new technique of harnessing human insights to semi-automate the process of interpreting raw, machine-sensed nonverbal behavior. Our primary motivation is the observation that, while computer algorithms are more reliable at consistently and objectively sensing subtle human behavior, human intelligence is superior at interpreting contextual behavior. Therefore, we see an opportunity to allow computer algorithms to perform the sensing portion while outsourcing the interpretation process to humans. This allows us to develop a semi-automated behavior analysis system. To instantiate our approach, we developed an online framework that can automatically record and analyze videos, and later provide data-driven feedback to the users. In addition, our framework allows users to share their data with Turkers for subjective interpretation of their nonverbal behavior. Our system automatically prioritizes the Turkers' comments and presents users with those that are most helpful and constructive.

The following example outlines how a user would interact with our system. An Internet user opens their web browser and navigates to our application. Before starting, the user is given the choice to practice in normal mode or private mode. (Private mode allows users to practice without storing any audio or visual data beyond the duration of the session.) Once the user chooses to proceed, the system asks for the user's permission to enable his or her webcam and microphone. The user then begins practicing his or her speech, clicking on the "Stop Recording" button to initiate the upload and analysis of the recording. Our sensing framework in the server looks for features like smile intensity, body movement, loudness, pitch, speaking rate, volume modulation, and word prosody. With the user's consent, the framework can create a task on the Amazon Mechanical Turk website, where anonymous Turkers can view the video and rate it based on various behaviors. Soon after, the user is provided with synthesized behavioral data that includes machine-sensed nonverbal data coupled with the Turkers' most constructive interpretations and recommendations. We summarize our contributions in this paper below:

1. A semi-automated framework that is able to interpret and personalize human nonverbal behavior and provide appropriate and helpful recommendations in the various categories—overall impression, friendliness, body gestures, and volume modulation—using online workers available through Amazon's Mechanical Turk.
2. A streamlined browser-based platform that can, with user consent, automatically sense and analyze human nonverbal behavior—smiles, movement, loudness, pitch, speaking rate, volume modulation, and word prosody—in the cloud.
3. An online interface that allows users to visualize, explore, and understand their video, the machine-sensed nonverbal data, and resultant subjective interpretations.

BACKGROUND AND RELATED WORK

The design of a hybrid system that combines automated sensing with human insights to analyze, interpret, and personalize human nonverbal behavior unites several disparate bodies of knowledge, including social signal processing, affective computing, and crowdsourcing. The paragraphs below outline work done in these areas.

Social Signal Processing

The desire to recognize social meaning by interpreting nonverbal cues gave rise to a new field called Social Signal Processing (SSP) [3] [4]. The SSP research community has progressed greatly in automatically sensing gesture and postures [5], inferring basic emotions from facial and eye behavior [6] [7] [8], and vocal characteristics [9]. While sensing a set of nonverbal behaviors is now a tractable problem, accurate interpretation of the nonverbal behavior requires contextual understanding. Context chiefly refers to *where* the interactions take place, *what* the activity is of the individuals involved, *when* the interactions take place, *who* is involved in the interaction, *why* the interaction is taking place, and *how* the interaction occurs. These questions can explain communicative intention, including the affective and cognitive states of the observed person(s) [10]. However, context-sensing is an extremely difficult problem and is still largely unexplored in affective computing research [11]. Therefore, we believe it may be useful to explore ideas related to crowdsourcing as a possible solution to this otherwise intractable problem.

Feasibility and Reliability of Outsourcing Tasks to Crowdsourced Workers

How feasible and reliable is it to outsource tasks to crowdsourced workers? Recent studies have demonstrated that it is possible for non-expert workers to achieve a high level of agreement and interpretive convergence both by being prescreened for requisite cognitive aptitudes and by obtaining basic training [12]. Examples of crowd power in interactive computing include editing written work [13], answering questions about photographs (nearly in real-time) to aid individuals with visual impairment(s) [14], and providing real-time captions by converting speech to text to help individuals who are hard-of-hearing [15]. Crowdsourcing techniques have also been used in behavioral video annotations. For example, Lasecki et al. [16] developed a tool called Glance that allows researchers to rapidly annotate video data with online workers, who annotate the data in parallel. While coding a single type of behavioral event, Glance produced judgments on a 12-minute-long video in two minutes (6x speedup) and coded 48 minutes of video in only five minutes (nearly 10x speedup), achieving an accuracy of 99% at identifying event occurrences. This further motivates our approach to use Turkers to perform more complicated behavioral labeling tasks.

Recent work by Cheng et al. [17] provides compelling evidence for the benefit of hybrid crowd-machine learning, over pure human or pure machine learning classification. This illustrates the power of synthesizing computation with crowd to solve problems that humans or machines cannot perform alone.

FRAMEWORK FOR GENERATING, STORING, AND DISPLAYING FEEDBACK ON NONVERBAL BEHAVIOR

Developing a framework that can extract, analyze, and interpret human nonverbal behavior in the cloud introduces a set of technical challenges: (1) building an accessible system that can capture and analyze audio-visual data across many different platforms; (2) extracting relevant nonverbal features from recorded audio and video; (3) automating crowdsourcing to generate contextualized ratings and commentary; (4) ranking the most constructive crowdsourced feedback; and (5) synthesizing the machine-sensed data with subjective feedback as part of an intuitive and interactive interface for users to explore their behavior.

Recording Platform

In order to enhance the accessibility of our framework, we designed it to be self-contained and to require minimal effort from the user. Users only require a webcam, microphone, and an Internet connection to use our system.

The first step in using our platform is to record a video. We use a JavaScript-based video recording library—RecordRTC [18]—for users to record and view their videos on our site with native support from the Google Chrome browser without the need for additional extensions. Audio and video data are recorded separately in the browser and merged in the server. Video is saved at a resolution of 320px by 240px. The audio is recorded in stereo, but the channels are merged to mono prior to upload to reduce file size.

The system automatically uploads the recorded audio to the server once the user stops recording. Users are directed to the feedback page once video processing completes. A two-minute video may take around five minutes to upload and undergo feature extraction.

Extraction of Nonverbal Features from Recordings

To select features to be integrated into our system, we reviewed existing literature on nonverbal communication and attended a public speaking workshop organized by Own the Room! [19]. We incorporated features based on whether they are effective in public speaking scenarios, and whether state-of-the-art sensing technology would be mature enough to capture and analyze them reliably, accounting for the hardware that users likely have. Given those considerations, we selected smile intensity, body movement, pitch, loudness, and word prosody.

Smile Intensity

Our system extracts smile intensity from each video frame. This shows how the user's smile intensity changes over time. The system integrates the state-of-the-art Shore Framework [20] to detect faces and facial features in each frame. We trained a binary classifier with sample images using the Adaboost algorithm. Each image represented two possible cases: "smiling" or "neutral." Facial features were used for the boosting. Smile intensity is gauged for each frame as a normalized number ranging from zero to 100, with 100 representing greatest smile intensity. Evaluation of the smile classifier using the Cohn-Kanade dataset [21] resulted in precision, recall, and F-measure values of 0.90, 0.97, and 0.93, respectively.

Movement

Gesturing while speaking can help convey a person's point. For instance, exaggerated hand movements may accompany important or interesting statements. While additional devices like Kinect provide a comprehensive set of body movement measurements, we do not assume users of our program, ROC Speak, possess them. To ensure the ubiquity of our system, we use a background-subtraction technique to gauge how much a human participant moves.

Our algorithm first calculates the absolute pixel-wise differences between every adjacent pair of video frames. The average of the pixel differences represents the raw movement value between two frames. We smooth these values by averaging the raw movement values within a window of length n.

Though background subtraction provides one measure of the movement, using this method for movement analysis can be limiting. The algorithm assumes that the background is constant and that only the user moves. Decoupling overall body movement into specific bodily gestures will be part of our future work.

Loudness and Pitch

Appropriate loudness while speaking is important in order to be heard by an audience. Changing pitch is one way to break monotony and emphasize a point. *Praat* [22], an open source speech processing toolkit, is used to extract loudness in decibels (dB) and pitch in Hz over time. We use a pitch floor of 50 Hz and a pitch ceiling of 350 Hz as constraints. This range encompasses the normal range of pitch in the human voice. The accuracy of the movement, pitch, and volume data are affected by the noise of the data generated by the users' environments and by the quality of the captured audio and video.

Word Prosody

Changes in the duration and loudness can affect perception of emphasis on words. To illustrate how the user modulates his or her speech, our system determines what words are spoken, how long it takes the user to say each word, and the average loudness of the speech.

Our framework uses both the Google Web Speech API [23] and Nuance Speech Recognition SDK to generate transcripts of speeches. The Penn Phonetics Lab Forced Aligner (P2FA) [24] matches the words of the transcript with the sounds in the recorded audio. The forced alignment provides the exact start and end times of each transcribed word. These times are used to find the average loudness and duration of each spoken word.

Gathering Human Feedback

Along with automated feedback, we implemented a functionality that allows users to automatically seek crowd opinions by sharing his or her videos. The following sections outline the challenges of relying on crowdsourcing subjective tasks and describe our contributions to the process.

Event Sequence

To initiate the process of receiving feedback, the user clicks a button that automatically generates tasks to gather input from a set of number of Turkers. As the Turkers complete the tasks, the feedback interface—which continuously polls for new results—updates with the newly-acquired results.

Measures

Turkers are asked to provide: (1) numerical ratings from 1-7—a score of 7 being the best—for overall performance, bodily gestures, friendliness, and volume modulation; and (2) at least one comment in one of the previous categories. The comment is associated with a specific timestamp.

Quality Assurance

Overseeing the quality of a subjective task completed by Turkers remains a challenging problem. To address this, we designed an intelligent interface to ensure that proper time and attention are devoted to each task. For instance, we added scripts to the online surveys that revealed the questionnaire and the option to submit the task *only* after a Turker had watched the entire video at least once, without skipping ahead. We check this by tracking interactions with the video's seek bar. An example of the interface is available at http://tinyurl.com/ratingui.

Response Time

The response time from Turkers varied between monetary incentive amounts. Depending on the time of day (and with possible outliers), we found that it usually took between 20 minutes and an hour to gather feedback from 10 workers, when the workers were rewarded 50 cents. ROC Speak shows a time estimate so users may decide on an optimal strategy for gathering feedback.

Automated Ranking of Human Feedback

Users may find it distracting to view a high volume of feedback. To address this issue, we have implemented two different learners that can rank the comments based on helpfulness and sentiment (e.g., positive and negative). This, by default, allows the users to view a combination of the most helpful and most (or least) positive comments. The sections below highlight the process of collecting data, and training the *helpfulness* and *sentiment* learners.

Data Collection

We realize that helpfulness and sentiment are subjective notions. To minimize subjective variance as we quantify these, we considered using a crowd of non-experts to produce an aggregate score. This inspired us to collect helpfulness ratings from Amazon Mechanical Turk. We recruited 25 Turkers to give a speech—on a personal hobby, finding cheap plane tickets, or commencement—and gathered 53 unique videos. We then recruited additional Turkers—at least 10 per video, resulting in at least 30 comments per video—who remarked on a speaker's friendliness, body language, and volume modulation.

Ground Truth Labels for Helpfulness Scorer

Each comment was rated by 10 different Turkers with a number ranging from 1 to 4, with a score of 4 being the most helpful. For the ground truth label, we simply sum the ratings from the crowd to obtain a helpfulness score. For instance, a highly-ranked comment could be "You don't use many hand gestures, and you look off the screen often. Focus on the camera." (An unhelpful comment: "Good speech.") For our training and testing dataset, we obtained a total of 1,649 time-stamped comments with corresponding helpfulness scores.

Ground Truth Labels for Sentiment Classifier

As the sentiment classifier aims to provide insights about helpful criticisms and praises, we excluded comments that were labeled most unhelpful—that is, those in the lowest quartile of the helpfulness ratings. A research assistant manually labeled the remaining comments as either positive or negative. Thus, we obtained a dataset with 742 positive comments and 285 negative comments.

Ranking Comment Helpfulness

We obtained the helpfulness ranking by modeling and predicting the helpfulness score.

Features Extracted: We extracted two types of features, text and video. (Described below.)

We observed that comments rated with high helpfulness usually contain complete sentences. These comments tend to be longer, contain punctuation, and are capitalized at the

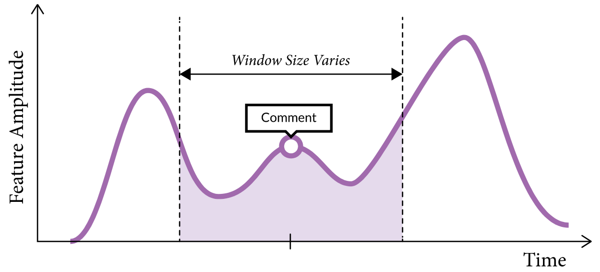


Figure 2. Visual and audio features are extracted as the amplitude within windows of different sizes.

beginning. We extracted the following text features to capture these qualities: total number of characters, presence of punctuation in the comment (e.g., commas, periods, dashes, and question marks), and presence of capital letters. We also extracted the number of pronouns, nouns, verbs, and adjectives present within the comment. We used the Natural Language Toolkit (NLTK) [25] parser to determine the part of speech of each word.

We also integrate the audio and visual information from the recorded video as features. For each of our four measurements M, where M is either smile intensity, movement, loudness and pitch, or prosody, we extract the range of M within one-, two-, and four-second windows of the timestamp, as well as the value of M at the timestamp.

Algorithm: We considered helpfulness to be the response variable, gathered from the crowd on Amazon Mechanical Turk. In fitting the model, we evaluated the predictions using two metrics: (1) mean absolute error (MAE) of the helpfulness score; and (2) coefficient of determination (R^2) between the actual and predicted score.

We used linear regression to predict the helpfulness, denoted y, based on features x extracted from the textual, visual, and audio features. Given an input feature vector $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{p}}$, the linear model predicts $\hat{\boldsymbol{y}} \in \mathbb{R}^{\boldsymbol{p}}$ as $\hat{\boldsymbol{y}} = \boldsymbol{\beta_0} + \boldsymbol{x^T}\boldsymbol{\beta}$. To learn the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta})$ from the training set of n pairs $(\boldsymbol{x_i}, \boldsymbol{y_i})$, we minimize the sum-of-square error

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - (\beta_0 + \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}))^2 + \lambda R(\boldsymbol{\beta})$$

We used elastic net [26] formulation for the regularization term

$$R(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|)$$

where $\alpha \in (0, 1)$ is the tradeoff between l_1 and l_2 norms. We believe the mixture between l_1 and l_2 regularization is more robust in selecting the key features of our predictive task. We trained three regression models in body movement, volume, and friendliness by tuning the parameter set (α, λ) using tenfold cross-validation.

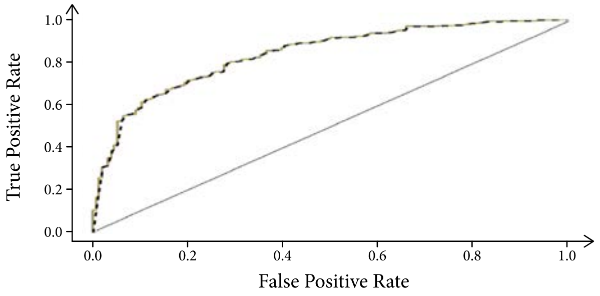


Figure 3. The receiver operating characteristic (ROC) curve.

Sentiment Classification

We extracted unigrams and bigrams as features for sentiment classification. Each comment is then transformed to a feature vector in its tf-idf [27] representation. We applied a Naive Bayes classifier over the training data to predict the sentiment of the comments.

Evaluation

Helpfulness: Table 1 shows our results for the three predictive tasks. While body movement has the largest MAE, it is the most representative model among the three categories. This result is consistent with our observations of the data; while Turkers tend to comment on overall volume and friendliness, they usually time-tag their comments to precise points where particular body movements are detected.

Table 1. Test set performance for the regression models, measured in mean absolute error (MAE) and coefficient of determination ($\mathbf{R^2}$).

Regression	Mean Absolute Error (MAE)	Coefficient of Determination (R^2)
Body Movement	3.343	0.304
Volume	3.037	0.178
Friendliness	3.340	0.238

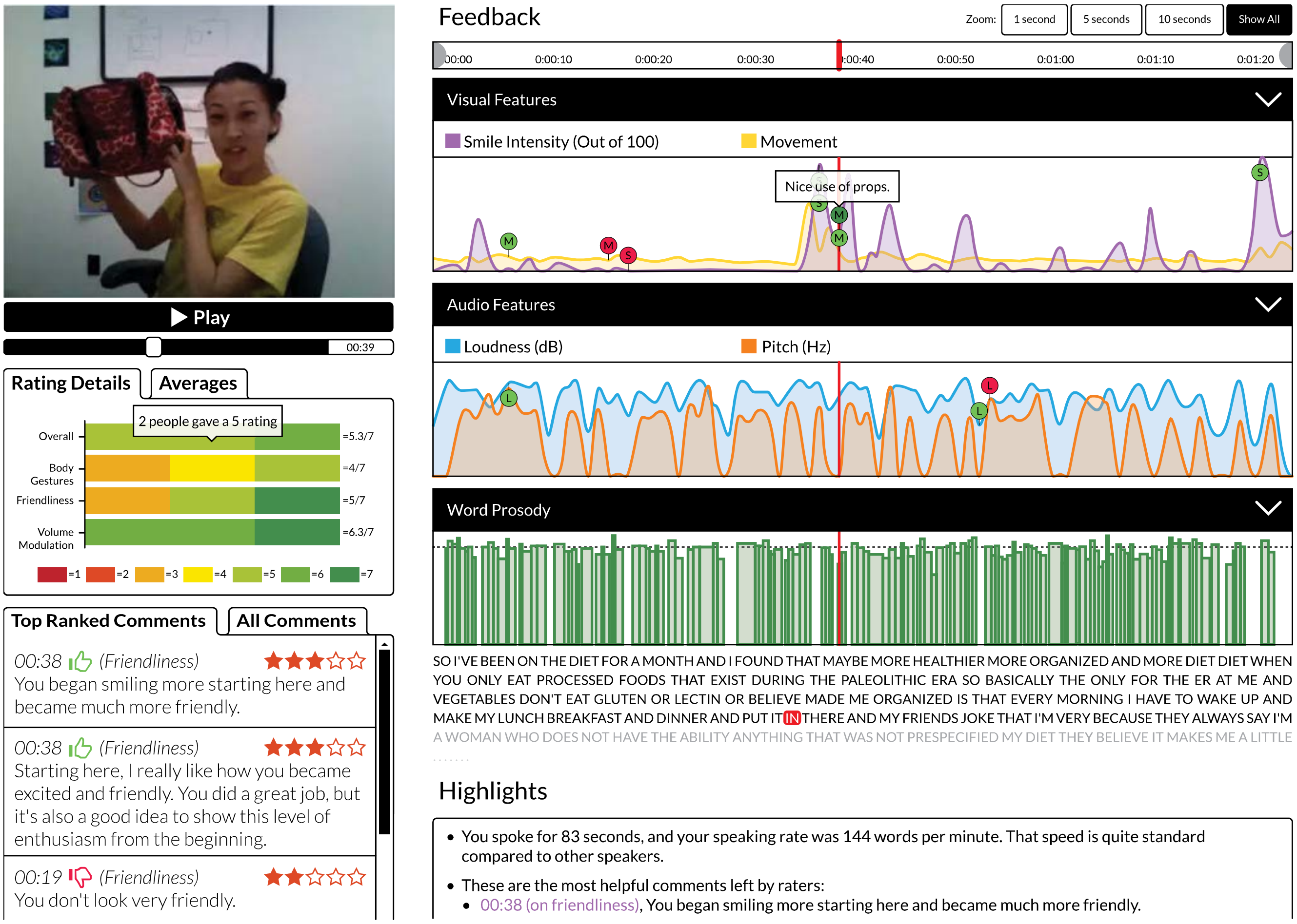
Sentiment Analysis: Figure 3 shows the ROC curve for the sentiment analysis classifier. When we used 70% of all data to train the sentiment classifier and 30% for testing, the classifier achieves an accuracy of 82.03%.

Feedback Interface Design and Synthesis of Automated and Human Feedback

Our feedback interface design (Figure 4) was motivated by a desire to be relevant, intuitive, and interactive.

Default View

When the user loads the feedback page for the first time, the visual features, audio features, and word prosody graph panels are collapsed by default to avoid overwhelming the user with information. When the user is ready to explore the graphs, he or she may open them.



00:38 (Friendliness) ★★★☆☆
You began smiling more starting here and became much more friendly.

00:38 (Friendliness) ★★★☆☆
Starting here, I really like how you became excited and friendly. You did a great job, but it's also a good idea to show this level of enthusiasm from the beginning.

00:19 (Friendliness) ★★☆☆☆
You don't look very friendly.

SO I'VE BEEN ON THE DIET FOR A MONTH AND I FOUND THAT MAYBE MORE HEALTHIER MORE ORGANIZED AND MORE DIET DIET WHEN YOU ONLY EAT PROCESSED FOODS THAT EXIST DURING THE PALEOLITHIC ERA SO BASICALLY THE ONLY FOR THE ER AT ME AND VEGETABLES DON'T EAT GLUTEN OR LECTIN OR BELIEVE MADE ME ORGANIZED IS THAT EVERY MORNING I HAVE TO WAKE UP AND MAKE MY LUNCH BREAKFAST AND DINNER AND PUT IT IN THERE AND MY FRIENDS JOKE THAT I'M VERY BECAUSE THEY ALWAYS SAY I'M A WOMAN WHO DOES NOT HAVE THE ABILITY ANYTHING THAT WAS NOT PRESPECIFIED MY DIET THEY BELIEVE IT MAKES ME A LITTLE ....

- You spoke for 83 seconds, and your speaking rate was 144 words per minute. That speed is quite standard compared to other speakers.
- These are the most helpful comments left by raters:
  - 00:38 (on friendliness), You began smiling more starting here and became much more friendly.

Figure 4. Crowdsourced comments and ratings and machine-generated automated feedback are displayed on the same interface. Users can hover over the color-coded markers on the graphs to see the time-stamped comments, and zoom in on the graphs to see their machine-sensed nonverbal behavior in detail. The example feedback page shown here can be seen live at http://tinyurl.com/feedback-ui. (Currently only supported in Google Chrome.)

Embedded Comments in Graphs

The graphs of audio and visual features are annotated with markers, color-coded by classified sentiment, where red indicates negative sentiment and green indicates positive sentiment. The markers are placed at the original time that the rater commented. When the user hovers over the marker, the comments appear. This allows users to explore comments in the graph as the video is playing.

Human Ratings

The subjective ratings for the four categories—overall, bodily gestures, friendliness, and volume modulation—are shown on the left using two different representational styles. One view shows only the averages in a simple radar chart, while the other shows a detailed, color-coded distribution of ratings. In the latter representation, green indicates better performance.

Ranked Comments

The most helpful comments for each category are featured in the highlights box. On the left, users can toggle between viewing all comments and viewing only the top-ranked, most helpful comments. Our linear regression model determines the comment helpfulness score by looking at both textual features and the audio-visual features that are extracted from the recorded video. This helpfulness score is expressed to the user on a five-star rating scale. Comment sentiment is indicated with a colored thumbs up/down image, and users can click on the comment time to go to the corresponding time in the video and graphs.

EVALUATION

Evaluation of our system sought to answer the following three questions through three separate studies, in the context of public speaking.

1. How helpful and accurate are machine-generated automated features when presented as feedback to the participants?
2. Are non-expert Turkers able to provide personalized, meaningful, and accurate feedback to the participants, in the context of public speaking?
3. How valuable is our feedback when automated feedback is combined with human interpretation, relative to watching one's own video and receiving generic tips?

Study #1: Determining Helpfulness and Accuracy of Automated Feedback

Experimental Design

To determine the helpfulness and accuracy of the machine-generated automated feedback, we set up a user study in which participants were randomly divided into two groups and counterbalanced. All of the participants were told to narrate two jokes that we provide to them. Telling a joke requires many of the same skills needed for public speaking, including appropriate volume modulation, timely body language, and suspenseful build-up, in order to engage the audience.

We designed two kinds of feedback: video-only and automated. In the video-only feedback, participants viewed only their own videos. In the automated feedback, participants saw graphs on smiles, body movement, and volume modulation, along with their own videos, but with no subjective interpretation. Each participant received a $10 gift certificate for successful completion of the study.

Participants

For this study, we recruited 23 university students (17 male, six female) to test our system. The participants used our laptop in an empty room in our laboratory. 12 participants received the automated feedback first, while 11 participants received only their video.

Study Procedure

The participants were first briefed on the procedure of the study. For the first round, participants were given a joke script to memorize. After narrating the first joke from memory in front of the camera, one group saw their own video recording after the first round and the automated feedback after the second. The other group saw the automated feedback after the first round and their own video recording after the second. The procedure for the two groups was identical except for the order of the feedback conditions.

For each round, the researchers stepped out of the room as the participants recorded themselves, interacted with the feedback, and completed the surveys, in order to reduce any influence on the participants. After the participants completed both rounds, we debriefed the participants and allowed them to share any impressions that the questionnaire did not capture.

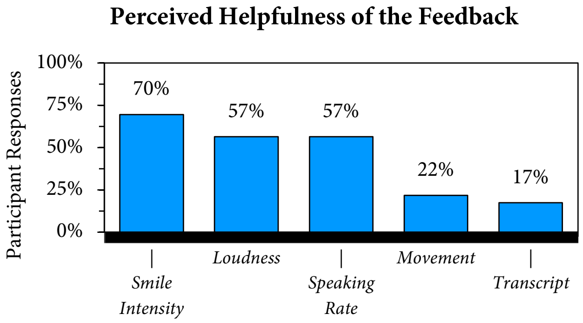


Figure 5. Participant responses to the question, "Which of the following nonverbal features were most useful to you?"

Measures

The questionnaire participants completed after receiving the video-only feedback can be found at http://tinyurl.com/feedback-video-only. The questionnaire participants completed after viewing the automated feedback can be found at http://tinyurl.com/feedback-rocspeak.

Results

When asked in the questionnaires about which nonverbal feedback features they preferred, participants responded with smile intensity, loudness, speaking rate, movement, and content of speech, respectively (Figure 5). The participants also rated their own performance and the helpfulness of the system at the end of each round. There was no significant difference between the ratings participants gave themselves after receiving the automated feedback and after receiving the video-only feedback.

Qualitative Feedback from the Participants

After the experiment, researchers interviewed each participant about their experience with the system and recorded the conversations. Even though participant self-ratings on performance did not change between sessions, they did find certain aspects of the automated feedback helpful. By analyzing their feedback, the following trends emerged.

Feedback with video is more helpful than video alone.

Most of the participants thought that the automated feedback with smile intensity graphs, body movement, and volume modulation was significantly more helpful than watching one's own video. One participant commented:

"It is lot harder to get a judgment based on watching the video alone; I guess I would prefer something in addition to the video to help supplement it."

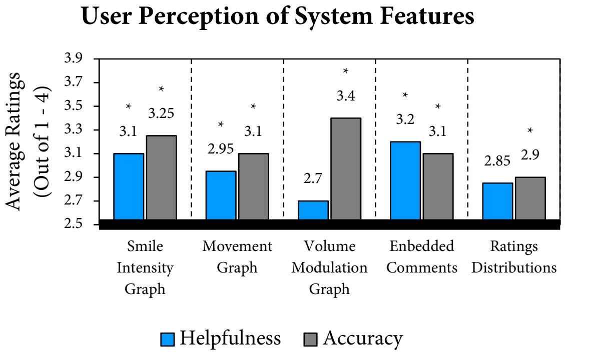


Figure 6. Average ratings from all participants. Ratings range from 1 to 4, with 1 being "Very Unhelpful/Very Inaccurate" and 4 being "Very Helpful/Very Accurate". A star indicates that the value was significantly above neutral (2.5), according to two-tailed t-tests with alpha 0.05.

<u>*Quantified nonverbal features are interesting and valuable.*</u>

Being able to quantify the subtle nonverbal behavior that we could interact with, interpret, and understand was valuable to the participants. One participant commented:

"I like the way [the system] gave me the intensity of the smile; it's something that you can't see. You can hear how loud you are, you can tell when you are taking a pause, but you can't look at yourself to check whether you are smiling or not."

Some participants also desired more personalized guidelines on their nonverbal behavior to understand the quality of their own performance and how they appeared to others. One participant said:

"I kind-of found the video with the feedback and the video-only feature almost equally helpful. Because I didn't find any metric to base the charts, I don't know what is a good joke, what makes a good presenter, other than my own little biases."

Findings

The results indicated that, while the automated feedback was adding more value than solely watching one's own video, context and personalization added value to the automated feedback.

Study #2: Determining Helpfulness and Accuracy of Non-Expert Human Commentary

In our second study, we explored whether non-expert Turkers would be a viable source of helpful and accurate personalization to the automated feedback.

Experimental Design

Participants were told to come to our lab and narrate a specified joke from memory as they are recorded by the system. After the experiment, participants received a feedback page via email. The page contained both

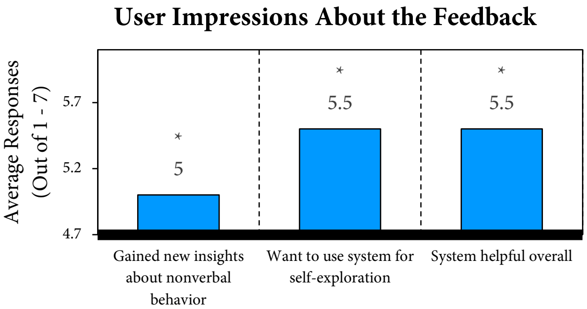


Figure 7. Participants level of agreement with statements, with 1 being most negative and 7 being most positive. Each bar shows the average response among all participants. A star indicates that the average value was significantly above the mean (4), according to two-tailed t-tests with alpha 0.05.

subjective ratings and comments generated by Turkers and automated feedback on their performance. Each participant received a $10 gift certificate for successful completion of the study.

Participants

We recruited 20 university students (nine females, 11 males). In addition, we recruited two sets of Turkers (10 per set) through a pre-screening test. The first set of Turkers would view the videos of our participants, and the second set would rate the helpfulness of the comments generated by the first set.

Study Procedure

Each participant was given enough time to memorize the provided joke—the first of which is available at http://goo.gl/sdl0d6—before they retold it in front of our system. After recording ended, the videos were sent—with participant consent—to Mechanical Turk. 10 Turkers per video provided comments and numerical ratings on overall performance, friendliness, body movement, and volume modulation. 10 different Turkers then rated the helpfulness of the preceding comments, allowing the system to prioritize the most helpful. Participants were emailed the link to the feedback on their nonverbal behavior after all of the Mechanical Turk tasks were complete. They were instructed to interact with the feedback for 15 minutes and complete a questionnaire to evaluate our system. The questionnaire can be found at http://tinyurl.com/f14-rocspeak.

Results

On average, the students rated most of the feedback features to be significantly helpful and accurate with regard to the interpretation of their nonverbal behavior (Figure 6). However, the feedback on volume modulation displayed the interesting condition of receiving the highest accuracy score, while also receiving the lowest helpfulness score, making it the only feature with significantly different accuracy and helpfulness ratings.

In addition to rating highly the automated feedback and Turker responses, a handful of participants noted that the commentary personalized the automated graphs:

"The graphs helped visualize these, but the comments made them personal. People noticed what I was doing and their personalized feedback affected me more than a chart of smile intensity."

Findings

The second study revealed that while graphs are helpful for the users, incorporating more context-specific feedback aids users' understanding of their own nonverbal behavior. The ratings from the participants show that it is possible for non-expert Turkers to generate helpful commentary and personalized feedback (Figure 7). We were also able to objectively validate the accuracy and helpfulness of the subjective comments, automatically-generated smile intensity, movement, and volume modulation graphs.

Study #3: Helpfulness and Accuracy of Combined Human and Automated Feedback

Our third study was a continuation of our second study with a key change where, instead of relying on Turkers to rate and rank the user comments for helpfulness and sentiment, we trained machine learning algorithms to automate the classification. We also made iterative improvements to the user interface based on the user input from the first two studies. More specifically, we wanted to quantify whether our proposed synthesized feedback, automated algorithms, and improved user interface would add value to the participants, in the context of public speaking.

Experimental Design

In order to validate the effectiveness of our feedback, we set up a study with Turkers. The Turkers were first asked to give a speech on a hobby or interest, using our system. We hired a second set of Turkers to generate comments on those videos. We then split the participants into two groups based on their self-reported confidence in their own public speaking skills, overall scores from the other Turkers for the speech given in the first round, and gender. One group of participants received the proposed ROC Speak feedback (Figure 4) while the other group received their own video, along with a list of general tips on public speaking.

Participants

We recruited 16 study participants from Mechanical Turk. 11 of the participants were male, and five were female. All were native English speakers.

Procedure

All recruited participants completed a pre-study survey, viewable at http://tinyurl.com/rocspeak-feb15-prestudy. They then brainstormed and delivered a two-minute speech on a personal hobby or interest, while being recorded by our system. The researchers initiated a request on behalf of participants for raters (other Turkers) to score and leave comments on the videos. As participants were themselves Turkers, they were told not to rate their own videos. This was verified by checking their unique worker IDs. Next, we split the participants into two groups—ROC Speak feedback (two females, six males) and video (three females, five males)—that had similar means and standard deviations of self-reported confidence in public speaking, and overall scores provided by other Turkers.

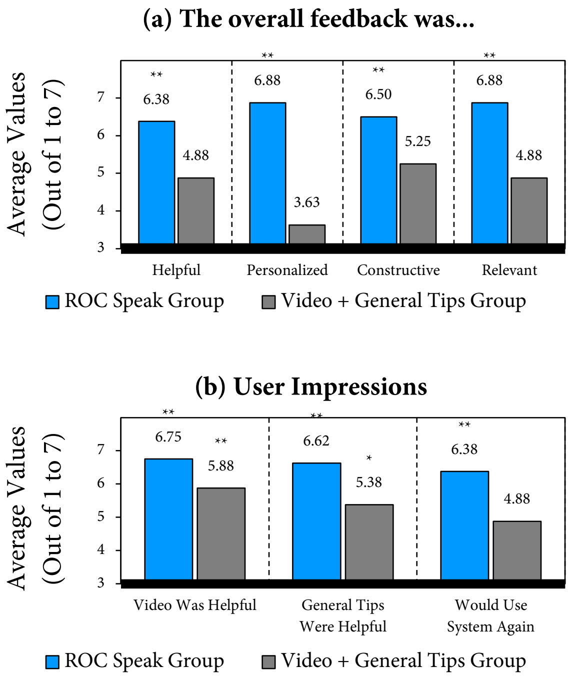


Figure 8. Participant responses to the post-study survey. Scores range from 1 to 7, with 1 meaning "Strongly Disagree" and 7 meaning "Strongly Agree." One star indicates that the average value was significantly above neutral (=4), according to a one-tailed t-test, and two stars indicate that the value was above neutral, according to a two-tailed t-test with an alpha of 0.05.

The group that received video and general tips saw only their own video and the list of tips on their page. Users who received the ROC Speak feedback were able to see the interface shown in Figure 4, which included the same general tips as in the video feedback case.

The users were instructed to interact with their feedback for 10-15 minutes. Mouse click and hover events in the feedback page were logged so that we could verify whether the participants interacted with the feedback for the required amount of time.

Measures

After reviewing the feedback for 15 minutes, the participants were asked to complete a post-study survey. The post-study survey for the video and general tips group can be seen at http://tinyurl.com/rocspeak-feb15-videopost. The survey for the ROC Speak feedback group can be seen at http://tinyurl.com/rocspeak-feb15-rocpost.

Results

According to the logged events, both groups interacted with the feedback for approximately 20 minutes, which is longer than the minimum requirement of 15 minutes.

As shown in Figure 8 (a), the ROC Speak group felt the feedback they received was helpful, personalized, constructive, and relevant, while the video group did not. In Figure 8 (b), we see that both groups thought the video and general tips were helpful. However, the general tips were perceived to be significantly more helpful when paired with ROC Speak feedback. The ROC Speak group showed a much higher interest in using the system again.

DISCUSSION

After our first study, where many participants liked to see quantitative results of nonverbal behavior analysis more than exclusively watching their own video, we wonder how the ROC Speak system compares to generic, non-personalized feedback. We consider the possibility that some participants like the automated graphs because they serve as reminders of what types of nonverbal behavior (e.g., smiles and bodily movement) to look for. That gives us intuition to investigate whether providing general tips on different aspects of nonverbal behavior would have a similar effect. Additionally, some participants desired alternative perspectives in their feedback. While most trust their own interpretations of the automated feedback, they would like to receive feedback that captures "audience opinion or interpretation."

Findings from our first study motivated us to introduce human insights to enhance feedback with personalization. We accomplished this by using the power of crowdsourcing. However, feedback generated by Turkers may only capture "audience opinion," and should not be considered "expert," like opinions from professionally-trained coaches.

Our second study shows that users found feedback from non-experts to be helpful and accurate. More specifically, the participants evaluated the comments and ratings of the quantified feedback as accurate and constructive. Positive comment ratings from participants indicate that Turkers can collectively offer helpful and accurate feedback, despite varying quality of individual comments. Participants also noted that, while graphs helped visualize nonverbal behavior, comments personalized the feedback.

In our third study, conducted outside of the lab, we evaluated our current system against a control case where users only receive generic feedback. Our current version of the feedback, consisting of human insights, automated rankings, and automatically-extracted features of nonverbal behavior, was rated to be helpful, personalized, and relevant, significantly more so than the video with general tips only. All participants expressed interest in using the ROC Speak system again for public speaking practice.

FUTURE WORK

We do not envision our system replacing expert help on improving nonverbal communication skills. Instead, our system serves as a supplement to existing methods, encouraging superior individual practice, especially if interaction with experts is inconvenient or impossible.

Some features can be refined in the current version of the ROC Speak system. Our future work will include support of Microsoft Kinect to add more fine grain analysis on body movement. For this experiment, we used the readily-available Turkers to comment on videos in exchange for a monetary incentive. However, our framework has the functionality for users to share their data with trusted circles in social media, allowing them to receive respectful feedback privately. Many individuals may not feel comfortable sharing their data, yet still want to receive personalized feedback. Given the latest advances in deep learning on generating freeform language descriptions of image regions [28], it may be possible to automatically generate captions for a video, addressing these privacy concerns.

Over time, ROC Speak could potentially amass the largest set of naturally-occurring nonverbal data ever collected, opening up new ideas and algorithms for behavior modeling. We look forward to sharing the framework and data with the research community.

CONCLUSION

In this paper, we presented ROC Speak, a ubiquitously-available framework for receiving personalized feedback on nonverbal behavior. We have developed a system that uses semi-automated human insight to interpret and contextualize human nonverbal behavior in the setting of public speaking. Our system automates the process of requesting feedback on nonverbal behavior from humans, compiles and ranks the comments and ratings, and visualizes them alongside fully-automated, quantitative interpretations.

Though we have validated our prototype in the context of public speaking, there are other areas in which our system could be used. As an automated sensing platform that can generate social cues and record fine-grained continuous measurement autonomously, this system can be deployed outside of the lab or school, increasing both the quantity and quality of data inexpensively and unobtrusively impacting areas like mental health and behavioral assessment. The most exciting part of this technology is that it places users in the driver's seat, providing an opportunity to improve their interaction skills.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation Award #IIS-1464162, Google Faculty Research Award, and Microsoft Azure Award. Kazi Islam and Anis Kallel helped conduct some experiments.

REFERENCES

[1] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, “Towards Practical Smile Detection,” *Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2106– 2111, 2009.

[2] M. E. Hoque, D. J. Mcduff, and R. W. Picard, “Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles,” *IEEE Trans. Affect. Comput.*, vol. PP, no. 99, pp. 1–13, 2012.

[3] A. Pentland, “Social Signal Processing [Exploratory DSP],” *IEEE Signal Process. Mag.*, vol. 24, no. 4, 2007.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” *Stud. Comput. Intell.*, vol. 411, pp. 119–135, 2013.

[6] F. D. La Torre, W. Chu, X. Xiong, F. Vicente, and J. F. Cohn, “IntraFace,” in *IEEE International Conference on Face and Gesture Recognition*, 2015.

[7] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, “The computer expression recognition toolbox (CERT),” in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops FG*, 2011, pp. 298–305.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[9] S. Ananthakrishnan and S. S. Narayanan, “Automatic prosodic event detection using acoustic, lexical, and syntactic evidence,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, pp. 216–228, 2008.

[10] A. Vinciarelli, H. Salamin, and M. Pantic, “Social Signal Processing: Understanding social interactions through nonverbal behavior analysis,” *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, 2009.

[11] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huanag, “Human-Centred Intelligent Human Computer Interaction (HCI^2): how far are we from attaining it?,” *International Journal of Autonomous and Adaptive Communications Systems*, vol. 1. p. 168, 2008.

[12] T. Mitra, C. J. Hutti, and E. Gilbert, “Comparing Person - and Process - centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk,” in *CHI 2015*, 2015.

[13] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, “Soylent,” in *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST ’10*, 2010, p. 313.

[14] J. P. Bigham, S. S. White, T. Yeh, C. Jayant, H. Ji, G. Little, A. Miller, R. C. R. Miller, A. Tatarowicz, and B. White, “VizWiz: Nearly real-time answers to visual questions.,” in *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST ’10*, 2010, pp. 333–342.

[15] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham, “Real-time captioning by groups of non-experts,” in *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST ’12*, 2012, p. 23.

[16] W. S. Lasecki, M. Gordon, D. Koutra, M. F. Jung, S. P. Dow, and J. P. Bigham, “Glance: Rapidly Coding Behavioral Video with the Crowd,” in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2014)*, 2014.

[17] J. Cheng and M. S. Bernstein, “Flock : Hybrid Crowd-Machine Learning Classifiers,” in *CSCW: ACM Conference on Computer-Supported Cooperative Work*, 2015.

[18] M. Khan, “RecordRTC to PHP,” *WebRTC Experiments*, 2013. [Online]. Available: https://www.webrtc-experiment.com/RecordRTC/PHP/.

[19] “OwnTheRoom.” [Online]. Available: owntheroom.com.

[20] B. Froba and A. Ernst, “Face detection with the modified census transform,” in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 91–96.

[21] T. Kanade, J. F. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” *Proc. Fourth IEEE Int. Conf. Autom. Face Gesture Recognit. (Cat. No. PR00580)*, pp. 46–53, 2000.

[22] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program].” [Online]. Available: www.praat.org.

[23] “Google Speech Recognition API.” [Online]. Available: https://www.google.com/intl/en/chrome/demos/speech.html.

[24] “Penn Phonetics Lab Force Aligner Documentation.” [Online]. Available: https://www.ling.upenn.edu/phonetics/p2fa/readme.txt.

[25] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, vol. 43. 2009, p. 479.

[26] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, pp. 301–320, 2005.

[27] C. D. Manning, P. Raghavan, and H. Schütze, “Scoring, term weighting, and the vector space model,” in *Introduction to Information Retrieval*, 2008, p. 100.

[28] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *CVPR*, 2015.