Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach

Sunghyun Park, Han Suk Shim*, Moitreya Chatterjee*, Kenji Sagae, and Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
12015 Waterfront Dr., Los Angeles, CA 90094
{park, hshim, mchatterjee, sagae, morency}@ict.usc.edu

ABSTRACT
Our lives are heavily influenced by persuasive communication, and it is essential in almost any types of social interactions from business negotiation to conversation with our friends and family. With the rapid growth of social multimedia websites, it is becoming ever more important and useful to understand persuasiveness in the context of social multimedia content online. In this paper, we introduce our newly created multimedia corpus of 1,000 movie review videos obtained from a social multimedia website called ExpoTV.com, which will be made freely available to the research community. Our research results presented here revolve around the following 3 main research hypotheses. Firstly, we show that computational descriptors derived from verbal and nonverbal behavior can be predictive of persuasiveness. We further show that combining descriptors from multiple communication modalities (audio, text and visual) improve the prediction performance compared to using those from single modality alone. Secondly, we investigate if having prior knowledge of a speaker expressing a positive or negative opinion helps better predict the speaker's persuasiveness. Lastly, we show that it is possible to make comparable prediction of persuasiveness by only looking at thin slices (shorter time windows) of a speaker's behavior.

Categories and Subject Descriptors
H.4.3 **[Information Systems Applications]**: Communications Applications.
J.4 **[Computer Applications]**: Social and Behavioral Sciences.

General Terms
Algorithms, Performance, Experimentation, Human Factors.

Keywords
Persuasion; persuasiveness; multimodal; prediction; social multimedia; POM corpus; persuasive opinion multimedia corpus.

1. INTRODUCTION
Our daily lives are heavily influenced by persuasive communication. Making a convincing case in the courtroom [41], seeking patients' compliance to medical advice [29], advertising and selling products in business [24] and even interacting with our friends and family all have persuasion at the core of the interaction.

ICMI '14, November 12 - 16 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2663260

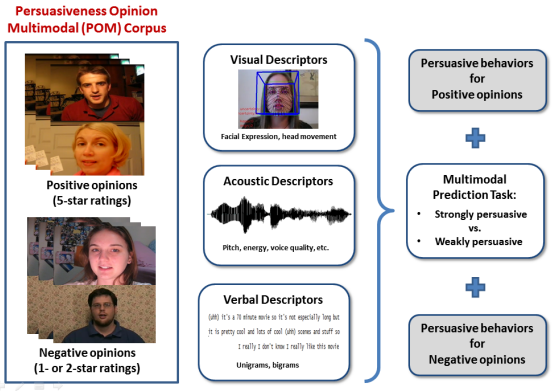

Figure 1. Overview of our multimodal approach in predicting persuasiveness in social multimedia.

With the advent of the Internet and the recent growth of social networking sites, more and more of our daily interaction is taking place in the online domain. Whereas the communication modality used online was predominantly text in the past, there is now an explosion of online content in the form of videos, making it more important and useful to understand persuasiveness in the context of online social multimedia content. What makes some people persuasive in online multimedia and influential in shaping other people's opinions and attitudes while others are ignored? This is the key question that we would like to start addressing with this paper.

While there has been a considerable amount of research on persuasion in the traditional sense, there has been very limited work investigating persuasion from the computational perspective and from the context of social multimedia. However, recent progress in computer vision and audio signal processing technologies [10, 13, 26, 27] is enabling automatic extraction of various visual and acoustic behavioral cues without having to depend on costly and time-consuming manual annotations, making it more feasible to tackle the problem from a more computational standpoint.

In this paper, we introduce our newly created Persuasive Opinion Multimedia (POM) corpus consisting of 1,000 movie review videos obtained from a social multimedia website called ExpoTV.com, which we plan to make freely available to the research community. Our experimental analysis revolves around the following 3 main research hypotheses. Firstly, we study if computational descriptors derived from verbal and nonverbal behavior can be predictive of persuasiveness. We further analyze the combination of descriptors from multiple communication modalities (audio, text and visual)

* The indicated authors made equal contribution to this work.

for predicting persuasiveness and compare with using single modality alone. Secondly, we investigate differences when speakers are expressing a positive or negative opinion and if having that prior knowledge helps in persuasiveness prediction. Lastly, we test if it is possible to make comparable prediction of persuasiveness by only looking at thin slices (i.e., shorter time windows) of a speaker's behavior.

In the next section, we give a brief overview of the literature that gave theoretical ground and motivation to our work, followed by our specific contributions. In Section 3, we outline our main research hypotheses, and Section 4 introduces our novel multimedia dataset designed for investigating persuasiveness in social multimedia. We give explanations on the design of our computational descriptors in Section 5 and experiments in Section 6. We report our results in Section 7 with discussions, and we conclude in Section 8.

2. THEORETICAL BACKGROUND

Persuasion in human communication has been a very hot topic for research over the past decades due to its wide applicability and substantial implications, and there is a plethora of sources in the literature that cover the topic in much breadth and depth. In this section, we give a brief review of past research that are only immediately relevant to our study. For an overview and history on persuasion research, interested readers are referred to other recent comprehensive texts such as [8, 28, 31].

In social psychology, the dual process models of persuasion [6, 32] have gained much attention and wide acceptance over the past decades. According to the models, there are two different routes we take when processing information that can influence our attitudes. One route is based on cognition that is more systematic and effortful while the other is based on peripheral or heuristic cues such as credibility or attractiveness of the message source. Our work in this paper can be seen in light of the dual process models with the focus on the peripheral route of information processing.

Modality Influence & Human Perception

Human communication is comprised of multiple modalities including verbal, acoustic and visual channels, and it is apparent that each modality has its own separate influence on human perception. Mehrabian [22] even goes as far to claim that our perception of an individual is determined 7% by his/her verbal content, 38% by his/her tone of voice, and 55% by his/her facial and bodily cues. Although his claim is arguable in our research context, it is obvious that multimodal analysis is an inevitable step to have a better understanding of human behavior and perception. In particular, Chaiken et al. [5] showed different influences on persuasion and comprehension when a message was delivered through the written, audiotaped or videotaped modality. Worchel et al. [42] also studied the effects on persuasion with different types of media, communicators, and positions of the message.

Acoustic Perspective

Showing the importance of acoustic cues in human speech, Stern et al. [40] reported that natural speech was more persuasive and taken more favorably than computer-synthesized speech. In addition, Mehrabian and Williams [23] reported that more intonation and higher speech volume contributed to perceived persuasiveness, Pittam [33] studied the relationship between nasality and perceived persuasiveness with a group of Australian speakers, Burgoon et al. [3] found a positive correlation between vocal pleasantness and perceived persuasiveness, and Pearce and Brommel [30] reported different effects of vocalic cues from conversational and dynamic speech styles on the perception of credibility and persuasiveness depending on the listener's preconceived notion of the speaker.

Verbal and Para-Verbal Perspective

Para-verbal cues are consistently found by many researchers to have a strong relationship with our perception of persuasiveness. For instance, Mehrabian and Williams [23] reported that higher speech rate and less halting speech contributed to perceived persuasiveness, Miller et al. [25] reported that a rapid speech rate positively influenced persuasion, and Pearce and Brommel [30] reported that dynamic and conversational styles (with varying characteristics in pitch, volume and use of pauses) had different effects on the perception of credibility and persuasiveness.

There are many components in the verbal domain that have strong relationship with persuasiveness [15, 44]. However, for the purpose of our work, we are not concerned with the validity or quality of argumentation in the textual data, and we are interested only at the level of finding key words that are informative in differentiating between strongly persuasive and weakly persuasive speakers.

Visual Perspective

Independent of text and voice, our facial expressions and bodily gestures convey much information as well. In relation to persuasion research, Mehrabian and Williams [23] found that more eye contact, smaller reclining angles, more head nodding, more gesticulation and more facial activity yielded significant effects for increasing perceived persuasiveness. LaCrosse [19] also found a similar set of nonverbal behavior related to persuasiveness that he calls affiliative nonverbal behavior. Moreover, Burgoon et al. [3] found that greater perceived persuasiveness correlated with kinesic / proxemic immediacy, facial expressiveness, and kinesic relaxation. Rosenfeld [36] found that the level of persuasiveness was positively correlated with positive head nods and negatively correlated with self-manipulations.

Thin Slice Prediction

Ambady and Rosenthal [1] showed that much inference is possible just by observing "thin slices" of nonverbal behavior, and Curhan and Pentland [9] applied the idea in a simulated employment negotiation scenario where they found that certain speech features within the first five minutes of negotiation were predictive of the overall negotiation outcome in the end. It is quite likely that the same idea can apply in the context of persuasiveness perception.

Contributions

To our knowledge, our new corpus is the first multimedia dataset created with the intention of studying persuasiveness in social multimedia. Furthermore, the main novelty of our work lies in investigating computational models of persuasiveness that take advantage of all 3 communicative modalities.

3. RESEARCH HYPOTHESES

Motivated by findings from past research outlined in the pervious section, our study presented in this paper was designed to specifically address the following three main hypotheses.

Computational Descriptors (Unimodal vs. Multimodal): As reviewed in the previous section, past research points to various cues in verbal and nonverbal behavior that influence human perception of persuasiveness. We hypothesize that we can capture such indicators of persuasiveness through computational descriptors to predict whether a speaker in social multimedia is

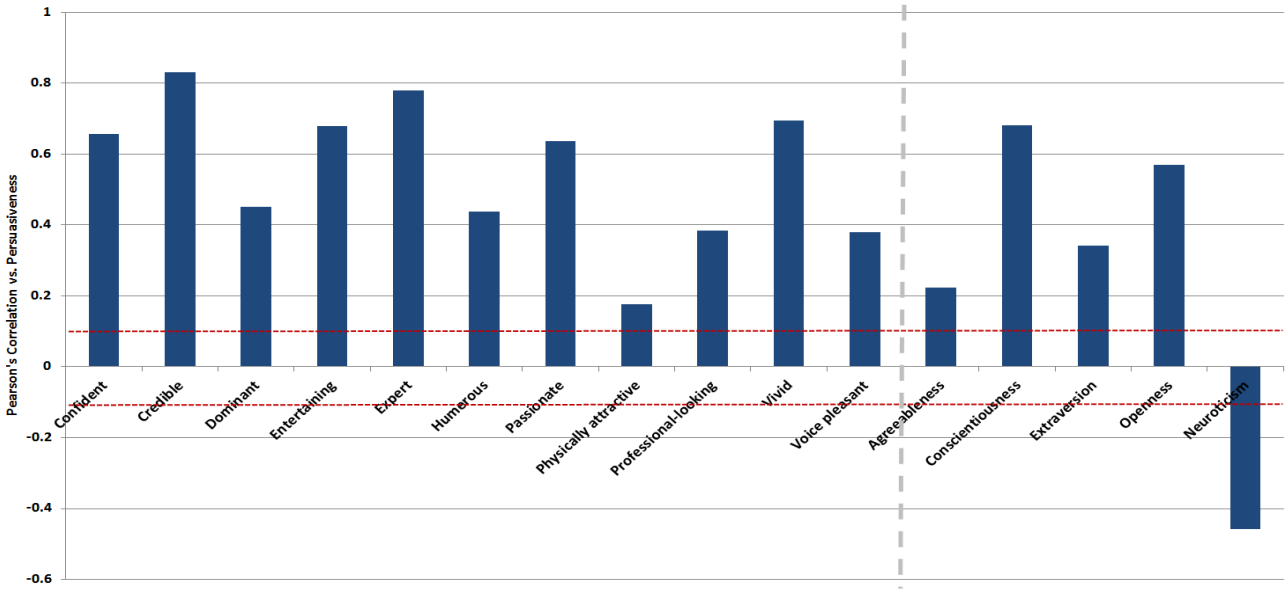

Figure 2. Pearson's correlation coefficients between persuasiveness and high-level and personality attributes (after taking the mean of 3 repeated annotations). The red line indicates statistical significance at $p < 0.001$, and the grey line visually divides the personality dimensions from other attributes.

strongly persuasive or weakly persuasive. In particular, we hypothesize that combining computational descriptors derived from multiple modalities of communication can make more accurate prediction than using those from a single modality alone from the acoustic, verbal or visual channel.

Hypothesis 1 (H1): Multimodal computational descriptors of verbal and nonverbal behavior perform better than unimodal descriptors in predicting a speaker's persuasiveness in social multimedia.

Prior Knowledge of Opinion Polarity: Persuasion can happen in a variety of context, and it is likely that we change our behavior depending on the context in our persuasion attempt. We hypothesize that if it is known in advance whether a speaker is trying to persuade one in favor of or against something, computational models can better capture the difference between persuasive and unpersuasive contents to make a more informed and better prediction.

Hypothesis 2 (H2): Persuasive behavior changes with opinion polarity and having prior knowledge of this sentiment polarity helps better predict a speaker's persuasiveness.

Thin Slice Prediction: In trying to persuade others, we may convey varying degrees of information in different stages of our persuasion attempt. For instance, we may tend to put more emphasis in the very beginning or we may typically want to close our speech with more impact close to the end. Combined with the idea of thin slices, we hypothesize that by looking at verbal and nonverbal behavior at specific shorter time periods, we can still make comparable prediction of persuasiveness of a speaker in social multimedia compared to making prediction based on the entire length of the speaker's behavior in video.

Hypothesis 3 (H3): Computational descriptors derived from a thin slice time period can make comparable prediction of a speaker's persuasiveness compared to those derived from the entire length of his/her video.

4. PERSUASIVE OPINION MULTIMEDIA CORPUS

Since there is currently no suitable corpus in the research community to study persuasiveness in the context of online social multimedia (currently the most relevant one to our knowledge is a dataset of online conversational videos by Biel et al. [2]), we found ExpoTV.com to be a good source to create a new corpus for our research topic. We plan to make our new Persuasive Opinion Multimedia (POM) corpus freely available to the research community. ExpoTV.com is a popular website housing videos of product reviews. Each product review has a video of a speaker talking about a particular product, as well as the speaker's direct rating of the product on an integral scale from 1 star (for most negative review) to 5 stars (for most positive review). This direct rating is useful for the purpose of our study because the star rating has close relationship with the direction of persuasion. For instance, the speaker in a 5-star movie review video would most likely try to persuade the audience in favor of the movie while the speaker in a 1-star movie review video would argue against watching the movie. Our corpus includes only movie review videos for the consistency of context. Since we were interested in exploring the difference in behavior between the cases when a speaker is trying to persuade the audience positively and negatively, we collected a total of 1,000 movie review videos as follows:

- **Positive Reviews:** 500 movie review videos with 5-star rating (315 males and 185 females).
- **Negative Reviews:** 500 movie review videos with 1 or 2-star rating, consisting of 216 1-star videos (151 males and 65 females) and 284 2-star videos (212 males and 72 females). We included 2-star videos due to a lack of 1-star videos on the website.

Each video in the corpus has a frontal view of one person talking about a particular movie, and the average length of the videos is about 94 seconds with the standard deviation of about 32 seconds. The corpus contains 372 unique speakers and 600 unique movie titles, including all types of common movie genres.

Table 1. Krippendorff's alpha agreement for our attributes.

Attribute	Kripp. alpha	Attribute	Kripp. alpha
Confident	0.73	Passionate	0.75
Credible	0.69	Professional-looking	0.70
Dominant	0.67	Vivid	0.68
Entertaining	0.68	Voice pleasant	0.67
Expert	0.70	Phys. Attractive	0.76
Humorous	0.74	**Persuasive**	**0.68**
Agreeableness	0.68	Openness	0.66
Conscientiousness	0.70	Neuroticism	0.64
Extraversion	0.73		

4.1 Subjective Annotations

Amazon Mechanical Turk (AMT) [21], which is a popular online crowdsourcing platform, was used to obtain subjective evaluations of the speaker in each video. A total of 50 native English-speaking workers based in the United States participated in the evaluation process online, and the task was evenly distributed among the 50 workers. To minimize gender influence, the task was distributed such that the workers only evaluated the speakers of the same gender. We note that although our primary focus was in investigating persuasiveness, various other high-level attributes including personality were also evaluated, making the corpus more widely applicable for other related research topics.

4.1.1 Persuasiveness & High-Level Attributes

For each video in the corpus, we obtained 3 repeated annotations on the level of persuasiveness of the speaker by asking the workers to give direct rating on the speaker's persuasiveness on a Likert scale from 1 (very unpersuasive) to 7 (very persuasive). In addition to persuasiveness, we also obtained evaluations on various high-level attributes, many of which past research suggests for having close relationship with our perception of persuasiveness. The high-level attributes were evaluated similarly as persuasiveness on a 7-point Likert scale with 1 being the least descriptive of the attribute and 7 being the most descriptive. For evaluating personality, a 10-item version of the Big Five Inventory [35] was used to assess the personality of the speaker in each video. We note that we also have self-assessed personality of the workers who performed the evaluations so that a deeper analysis can be possible by investigating the relationship between the personality of the perceiver and the perceived.

- **High-Level Attributes:** confident, credible, dominant, entertaining, expert, humorous, passionate, physically attractive, professional-looking, vivid, and voice pleasant.
- **Personality Dimensions (Big Five Model):** agreeableness, conscientiousness, extraversion, openness, and neuroticism.

4.1.2 Analysis

Due to variability in human perception and judgment, taking the mean or majority vote of repeated evaluations would be a sensible method of obtaining final labels. For our study, we used the mean score of 3 repeated Likert-scale evaluations as the final measure. Table 1 summarizes the mean agreement measured with Krippendorff's alpha between our final measure and each coder. The agreement is generally high around 0.70. Figure 2 shows the correlations between persuasiveness and other attributes when using our final measures, and many of the high-level attributes show a strong correlation with persuasiveness, which is consistent with past research in the literature [8, 28, 31]. It is particularly interesting to see which traits are not correlated or inversely correlated. The fact that physical attractiveness is only weakly correlated is most likely due to our design of the same-gender evaluation. Neuroticism is inversely correlated. Some of the most strongly correlated traits are credibility, confidence, and expertise.

To validate the persuasiveness measure, we included in the annotation tasks two questions related to the annotators' interest in watching the reviewed movies. For the first question, annotators were shown the synopsis of the movie and were asked, "How interested are you in watching this movie?" This first question was before watching the review. Then after watching the review, the annotators were asked the following second question, "After seeing this movie review, how interested are you in watching this movie?", with a scale ranging from -3 (much less interested than before) to +3 (much more interested than before). The annotators were also asked another question to note whether they have seen the movies before. Out of 3000 annotation tasks (1000 movies multiplied by 3 for repeated annotations), 1089 were marked by annotators who have seen the movies, and excluding those, our validity analysis shows a strong correlation between the persuasiveness score rating and the annotators' interest after watching the movie reviews, 0.69 and 0.54 for positive and negative reviews respectively.

4.1.3 Transcriptions

Using AMT and 18 participants from the same worker pool for the subjective evaluations, we obtained verbatim transcriptions, including pause-fillers and stutters. Each transcription was reviewed and edited by in-house experienced transcribers for accuracy.

Table 2. Overview of our computational multimodal descriptors

Acoustic
• Formants: F1 ~ F5 • Mel frequency cepstral coefficients: MFCC 1 ~ 24 • Pitch / Fundamental frequency (F0) • Voice qualities: normalized amplitude quotient (NAQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), quasi-open quotient (QOQ), difference between the first two harmonics (H1-H2), and peak-slope
Verbal and Para-Verbal
• Unigrams • Bigrams • Verbal fluency qualities: articulation rate, pause, pause-filler, speech disturbance ratio, and stutter
Visual
• Emotions: anger, contempt, disgust, fear, joy, sadness, and surprise • Valence: negative, neutral, and positive • Facial Action Units: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, and AU28 • Eye gaze movements: displacement in x and y axes • Head movements: displacement and rotation in x, y and z axes • Approximated posture: displacement in the z-axis
Statistical Functionals (acoustic and visual descriptors only)
mean, median, percentiles (10^{th}, 25^{th}, 75^{th}, and 90^{th}), ranges (between min and max, 10^{th} and 90^{th} percentiles, and 25^{th} and 75^{th} percentiles) skewness, standard deviation

5. COMPUTATIONAL DESCRIPTORS

In this section, we give details on the extraction and computational encoding of multimodal descriptors as potential candidates for capturing persuasiveness. Except for most of the verbal and para-verbal descriptors, which depended on manual transcriptions for feature extraction, all other descriptors were extracted and encoded automatically using various acoustic and visual tracking tools. All the computational descriptors that we used are summarized in Table 2.

5.1 Acoustic Descriptors

Following common approaches for conducting automatic speech analysis [39], we extracted various speech features related to pitch, formants, voice qualities and mel-frequency cepstral coefficients (MFCCs) using a publicly available software called Covarep [10]. The raw feature values were then used to compute common statistical descriptors including mean, median, percentiles, ranges, skewness, and standard deviation. The encoded features were then used to explore their feasibility in capturing persuasiveness in acoustic signals of speech.

- Formants: The information of acoustic resonance of the human vocal track, called formant, is commonly used for speech recognition and emotion recognition. We explored formants F1 through F5.
- Mel frequency cepstral coefficients (MFCC): Also widely used for speech and emotion recognition are MFCCs, and we explored MFCC 1~24.
- Pitch (F0), also referred to as the fundamental frequency, is closely tied to the affective aspect of speech [4].
- Voice Qualities: Many studies show a strong relation between voice quality features and perceived emotion [14], and it is widely used for emotion recognition in speech. We used various voice quality descriptors including normalized amplitude quotient (NAQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), quasi-open quotient (QOQ), difference between the first two harmonics (H1-H2), and peak-slope. For more details, readers are referred to [16, 17, 37].

5.2 Verbal and Para-Verbal Descriptors

From the verbatim transcriptions of the dataset, we extracted all standard unigram and bigram features commonly used in natural language processing, with the only difference in that the term frequencies were normalized by the video length. In addition, we observed a set of frequent para-verbal cues that could be associated with the level of persuasiveness.

- Articulation rate: Articulation rate is the rate of speaking in which all pauses are excluded from calculation and was computed by taking the ratio of the number of spoken words in each video to the actual time spent speaking.
- Pause: We computed this descriptor by counting all instances of silence during speech that are greater than 0.5 seconds in length, normalized by the total length of the video. FaceFX software [12] was used to automatically extract and encode this descriptor.
- Pause-filler: Pause-fillers are sounds that are used to fill the pause in speech, such as “um” or “uh.” This descriptor was computed by counting all instances of pause-fillers, normalized by the total number of words spoken in each video.
- Speech disturbance ratio: Pause-fillers and stuttering can be considered as the same category of speech disturbance [20]. We computed speech disturbance ratio by counting the number of speech disturbance instances (pause-fillers and stutter), normalized by the total number of words spoken in each video.
- Stutter: For this descriptor, we counted all instances of stuttering in each video, normalized by the number of words spoken in the video.

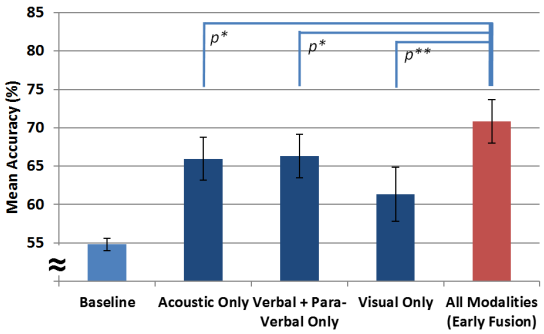

Figure 3. Persuasiveness prediction results for unimodal and multimodal models when using both positive and negative reviews combined ($p^* < 0.05$ and $p^{**} < 0.01$).

5.3 Visual Descriptors

Using readily available visual tracking technologies [13, 26, 27], we extracted various raw features from the face and the head movement of each speaker in the video. Similarly as the acoustic descriptors, we computed the same statistical descriptors to explore their usefulness in indicating persuasiveness.

- Discrete Emotions: The level of anger, contempt, disgust, fear, joy, sadness, and surprise.
- Valence: The level of high-level valence including negative, neutral and positive valence.
- Facial Action Units: The level of movements in various facial areas as codified by Facial Action Coding System (FACS) [11] including AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, and AU28.
- Eye gaze movements: The gaze movement in the x and y axes
- Head movements: The head displacement and rotation in the x, y and z axes.
- Approximated posture: The movement in the z axis (toward or away from the camera).

6. EXPERIMENTS

This section gives details on the experimental methodology, particularly on our prediction models and the experimental conditions we designed to test our research hypotheses.

6.1 Experimental Conditions

In order to address the first hypothesis (H1), we trained and tested the prediction models and performed feature selection in the following groups depending on the communication modality:

- Acoustic descriptors only.
- Verbal and para-verbal descriptors only.
- Visual descriptors only.

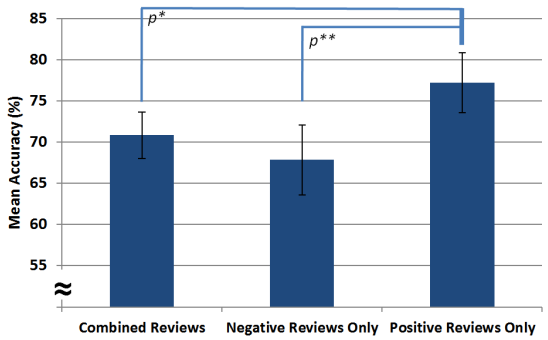

Figure 4. Persuasiveness prediction results for sentiment-independent and sentiment-dependent conditions using all modalities (p* < 0.05 and p < 0.01).**

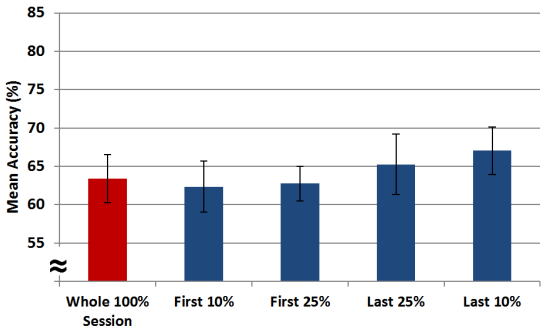

Figure 5. Persuasiveness prediction results using acoustic and visual descriptors in thin slices (shorter time windows).

- Early fusion: All types of descriptors together.

To address the second hypothesis (H2), we also performed our experiments in 3 different groups depending on the sentiment (opinion polarity) of the reviews:

- Positive reviews only (sentiment-dependent).
- Negative reviews only (sentiment-dependent).
- Both review types combined (sentiment-independent).

To address the third hypothesis (H3), we performed additional experiments with the acoustic and visual descriptor groups by computing the descriptors separately within different thin slices. More specifically, we divided each review video into 4 different thin slices of first 10%, first 25%, last 25% and last 10% of the video and repeated the same experiments. We did not use verbal and para-verbal descriptors for this because most of them were derived from manual transcriptions that did not have timestamp information.

6.2 Persuasiveness Labels

The ground-truth persuasiveness score of equal to or greater than 5.5 was taken as persuasive speakers and the score of equal to or less than 2.5 weakly persuasive speakers. After taking this step, we ended up with a total of 300 videos, specifically 157 videos of positive reviews (75 persuasive and 82 unpersuasive) and 143 videos of negative reviews (62 persuasive and 81 unpersuasive). We note that this selection process was done so that we could primarily focus on investigating behavioral differences between strongly persuasive and weakly persuasive videos. We keep as future work the regression analysis.

6.3 Methodology

For all experiments, we used the support vector machine (SVM) classifier with the radial basis function kernel as the prediction models [7]. The experiments were performed with 10-fold cross-validation (CV). Each CV experiment had 1 fold testing and 3-fold validation (among 9 training folds) for automatic selection of hyper parameters (i.e., gamma and C parameters for SVM). We repeated the experiments 3 times with randomly and newly separated folds. It is worth emphasizing that our folds were created such that no 2 folds contained samples from the same speaker or the same movie title. These restrictions assure user-independent experiments for better generalizability of our prediction models and results. Our evaluation metric is the averaged accuracy over all folds.

For feature selection, we mainly used the Information Gain (IG) metric [43]. We note that we performed feature selection only using the training samples from each cross-validation experiment. None of the test samples were used for feature selection. That is, we performed 10 separate feature selections using only the training samples for each 10 iterations during cross-validation testing. In order to keep the feature space to roughly $1/10^{th}$ of the sample size, we selected top 30 features as determined by the IG score.

7. RESULTS AND DISCUSSIONS

In this section, we report and discuss our results testing our 3 main research hypotheses described in Section 3, followed by descriptor analysis.

7.1 Multimodal vs. Unimodal (H1)

Figure 3 shows the prediction accuracy results with the combined movie review condition, the persuasiveness predictors trained by fusing the descriptors from all 3 modalities together performed better with a statistical significance compared to each type of unimodal predictors, confirming our first research hypothesis. Specifically, our multimodal predictors predicted persuasiveness with a mean accuracy of 70.84%, and paired-sample t-tests showed that the performance was statistically significance compared to acoustic only predictors that performed at 65.97% ($p < 0.05$), verbal and para-verbal only predictors that performed at 66.29% ($p < 0.05$), and visual only predictors at 61.35% ($p < 0.01$). The baseline or the majority vote accuracy when always siding with the class label with more samples was at 54.83%.

Our results suggest that visual, acoustic and verbal descriptors are complementing each other during the persuasiveness assessment of online multimedia content. Our multimodal predictor is able to identify this complementarity as shown by other statistically significant results.

7.2 Prior Knowledge of Sentiment (H2)

Figure 4 shows the prediction accuracy results for the multimodal predictors with all 3 modalities across different conditions of positive reviews only, negative reviews only, and all reviews combined. The predictors trained and tested using only the positive reviews performed best at 77.24%, and paired-sample t-tests show statistical significance compared to those trained and tested using only the negative reviews which performed at 67.84% ($p < 0.01$) and those using the combined reviews which performed at 70.84% ($p < 0.05$).

This result partially confirms our second hypothesis and suggests that our computational descriptors can capture indicators of persuasiveness far better when the speakers are trying to persuade to watch a movie compared to when they are trying to persuade you against it, with the prediction performance falling somewhere between the two when using both kinds of reviews together. This

Table 3. Top computational descriptors for positive and negative reviews which were automatically selected using Information Gain.

Positive Reviews	Info Gain	Negative Reviews	Info Gain
		Acoustic	
MFCC4: stddev	0.12	MFCC3: mean	0.12
MFCC4: range (min ~ max)	0.11	MFCC3: median	0.12
MFCC2: range (10th ~ 90th perc.)	0.11	MFCC3: 75th perc.	0.10
MFCC4: range (10th ~ 90th perc.)	0.10	F3: 10th perc.	0.10
MFCC8: skew	0.10	F4: median	0.09
		Verbal and Para-Verbal	
pause	0.13	pause	0.11
articulation rate	0.08	articulation rate	0.09
n-grams: in, things, there are, very, one, well, is the, you, say, movie that, even though, will be, movie its, deleted, a pretty, a lot, and they're, movie in, good movie, don't know, you should, character, that I'm		n-grams: make, to make, it but, not even, even, just, the film, night, do not, character, premise, out, movie this, end, terrible, real, kids, but if, at all, writing, feels, poor, yourself	
		Visual	
posture: 10th perc.	0.13	gaze movement x-axis (left/right): range	0.14
posture: 25th perc.	0.12	gaze movement x-axis: range (25th ~ 75th perc.)	0.13
posture: mean	0.12	head rotation in all x, y, z axes: range (25th ~ 75th perc.)	0.13
head movement z-axis: 90th perc.	0.12	head rotation in all x, y, z axes: range (10th ~ 90th perc.)	0.12
posture: median	0.11	AU12: stddev.	0.12

may be explained by the imbalance of sentiment or opinion polarity in our dataset. Due to the lack of 1-star movie reviews on ExpoTv.com, we included 2-star reviews to match the number of positive reviews in our dataset. It is quite possible that the 2-star review videos convey weaker verbal and nonverbal cues than the two extremes of 1-star and 5-star review videos, making the prediction with negative reviews more difficult than the positive reviews.

7.3 Thin Slice Prediction (H3)

Figure 5 shows the accuracy results of the predictors using both the acoustic and visual modalities across different thin slices. Using the whole length of each review yielded the mean prediction accuracy of 63.41%, and the figure shows that the performance is comparable across all thin slices. This result is a typical demonstration of the idea of thin slices, and it suggests that we can still make much inference on a speaker's persuasiveness just by looking at the very beginning or end of the movie review. Although not conclusive, another interesting finding is that there seems to be mild tendency that more verbal and nonverbal indicators of persuasiveness are present or stronger toward the end of the video.

7.4 Descriptor Analysis

Table 3 summarizes a short analysis highlighting top descriptors that have been discriminative in separating strongly persuasive and weakly persuasive speakers.

From the acoustic modality, MFCC descriptors in the low frequency regions stand out for predicting persuasive speakers in both positive and negative reviews, which are expected to perform better than high frequency regions due to denser resolutions and being more robust to noise. Although MFCC features are heavily used in speech recognition and analysis, they unfortunately do not give room for much interpretation.

Consistent with the literature described in Section 2, para-verbal descriptors of pause and articulate (speech) rate proved to show much discriminative power in separating speakers who are perceived as persuasive and those who are not. The unigram and bigram descriptors also look reasonable in that they tend to be more positive with words such as "will be," "a pretty," "a lot," "good movie" and "you should" for positive reviews. On the other hand, for negative reviews, the words tend to be a little more negative such as "it but," "not even," "do not," "terrible," and "poor."

From the visual modality, the descriptors from the approximated posture were predominant followed by those from head displacement movement for positive reviews. For negative reviews, the descriptors from gaze movement and head rotation showed most predictive power. Although not shown as top descriptors, AU1 (inner brow raiser) was a notable facial descriptor for indicating persuasiveness in the positive reviews, and AU12 (lip corner puller), AU14 (dimpler) and AU20 (lip stretcher) were among the next top descriptors in the negative review condition.

8. CONCLUSIONS & FUTURE WORK

We introduced a novel multimedia corpus specifically designed to study persuasiveness in the context of social multimedia. We presented our computational approach in using verbal and nonverbal behavior from multiple modalities of communication to predict a speaker's persuasiveness in online social multimedia content and showed that having prior knowledge of the speaker's sentiment has partial influence in better predicting the level of persuasiveness. Furthermore, we demonstrated that the idea of thin slices can be used to observer a short window of a speaker's behavior in the beginning and toward the end to achieve comparable prediction compared to observing the entire length of the video.

Interesting future directions include investigating more ways of computationally capturing various indicators of persuasiveness and better algorithmic methods of fusing information from multiple modalities. Our results will provide a baseline for all future studies using this new corpus for carrying out deeper analysis of understanding the relationship between persuasiveness and relevant high-level attributes including personality.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1118018 and the U.S. Army. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

We thank the USC Annenberg Graduate Fellowship Program for supporting the first author's graduate studies.

10. REFERENCES

[1] Ambady, N. and Rosenthal, R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin* 111, 2 (Mar 1992), 256-274.

[2] Biel, J., Teijeiro-Mosquera, L., and Gatica-Perez, D. 2012. FaceTube: Predicting personality from facial expression of emotion in online conversational video. In *Proc. of the 14th ACM Int'l Conf. on Multimodal Interaction*. ICMI '12, 53-56.

[3] Burgoon, J., Birk, T., and Pfau, M. 1990. Nonverbal behaviors, persuasion, and credibility. *Human Communication Research* 17, 1 (Sept. 1990), 140-169.

[4] Busso, C., Lee, S. and Narayanan, S. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17, 4 (May 2009), 582-596.

[5] Chaiken, S. and Eagly, A. 1983. Communication modality as a determinant of message persuasiveness and message comprehensibility. *Journal of Personality and Social Psychology* 34, 4 (Oct. 1976), 605-614.

[6] Chaiken, S., Liberman, A., and Eagly, A. Heuristic and systematic information processing within and beyond the persuasion context. In J. Uleman and J. Bargh (Eds.), *Unintended Thought: Limits of Awareness, Intention, and Control* (pp. 212-252). New York: Guildford Press.

[7] Chang, C.-C. and Lin, C.-J. 2011. LIBSVM: A library for support vector machine. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (Apr. 2011), 27:1-27:25.

[8] Crano, W. and Prislin, R. 2006. Attitudes and persuasion. *Annual Review of Psychology* 57 (Jan. 2006), 345-374.

[9] Curhan, J. and Pentland, A. 2007. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology* 92, 3 (May 2007), 802-811.

[10] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proc. of the 39th Int'l Conf. on Acoustics, Speech, and Signal Processing*. ICASSP '14.

[11] Ekman, P. and Rosenberg, E. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, New York, NY.

[12] FaceFX. http://www.facefx.com/

[13] FACET. http://www.emotient.com/cert

[14] Gobl, C. and Chasaide, A. 2003. The role of voice quality in communication emotion, mood and attitude. *Speech Communication* 40, 1-2 (Apr. 2003), 189-212.

[15] Hosman, L. 2002. Language and persuasion. In J. Dillard and M. Pfau (Eds.), *The Persuasion Handbook: Developments in Theory and Practice* (pp. 371-390). New York: Sage.

[16] Kane, J., Scherer, S., Aylett, M., Morency, L.-P., and Gobl, C. 2013. Speaker and language independent voice quality classification applied to unlabeled corpora of expressive speech. In *Proc. of the 38th Int'l Conf. on Acoustics, Speech, and Signal Processing*. ICASSP '13, 7982-7986.

[17] Kane, J., Scherer, S., Morency, L.-P., and Gobl, C. 2013. A comparative study of glottal open quotient estimation techniques. In *Proc. of the 14th Annual Conf. of the Int'l Speech Communication Association*. Interspeech '13, 1658-1662.

[18] Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology.* Sage, Beverly Hills, CA.

[19] LaCrosse, M. 1975. Nonverbal behavior and perceived counselor attractiveness and persuasiveness. *Journal of Counseling Psychology* 22, 6 (Nov. 1975), 563-566.

[20] Mahl, G. 1956. Disturbances and silences in the patient's speech in psychotherapy. *The Journal of Abnormal and Social Psychology* 53, 1 (Jul. 1956), 1-15.

[21] Mason, W. and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods 44*, 1 (Mar. 2012), 1-23.

[22] Mehrabian, A. 1971. *Silent messages*. Wadsworth Publishing Company, Inc., Belmont, CA.

[23] Mehrabian, A. and Williams, M. 1969. Nonverbal concomitants of perceived and intended persuasiveness. *Journal of Personality and Social Psychology* 13, 1 (Sept. 1969), 37-58.

[24] Meyers-Levy, J. and Malaviya, P. 1999. Consumers' processing of persuasive advertisements: An integrative framework of persuasion theories. *Journal of marketing* 63, 4 (Oct. 1999), 45-60.

[25] Miller, N., Maruyama, G., Beaber, R., and Valone, K. 1976. Speed of speech and persuasion. *Journal of Personality and Social Psychology* 34, 4 (Oct. 1976), 615-624.

[26] Morency, L.-P., Whitehill, J., and Movellan, J. 2008. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Proc. of the 8th Int'l Conf. on Automatic Face and Gesture Recognition*. FG '08, 1-8.

[27] OKAO Vision. http://www.omron.com/r_d/coretech/vision/okao.html

[28] O'Keefe. 2002. *Persuasion: Theory and Research*. Sage, Thousand Oaks, CA.

[29] O'Keefe, D. and Jensen, J. 2007. The relative persuasiveness of gain-framed loss-framed messages for encouraging disease prevention behaviors: A meta-analytic review. *Journal of Health Communication* 12, 7 (Oct. 2007), 623-644.

[30] Pearce, W. and Brommel, B. 1972. Vocalic communication in persuasion. *Quarterly Journal of Speech* 58, 3 (1972), 298-306.

[31] Perloff, R. 2010. *The Dynamics of Persuasion: Communication and Attitudes in the Twenty-First Century*. Routledge, New York, NY.

[32] Petty, R. and Cacioppo, J. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

[33] Pittam, J. 1990. The relationship between perceived persuasiveness of nasality and source characteristics for Australian and American listeners. *The Journal of Social Psychology* 130, 1 (1990), 81-87.

[34] Pornpitakpan, C. 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology* 34, 2 (Feb. 2004), 243-281.

[35] Rammstedt, B. and John, O. 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 (Feb. 2007), 203-212.

[36] Rosenfeld, H. 1966. Approval-seeking and approval-inducing functions of verbal and nonverbal responses in the dyad. *Journal of Personality and Social Psychology* 4, 6 (Dec. 1966), 597-605.

[37] Scherer, S., Kane, J., Gobl, C., and Schwenker, F. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language* 27, 1 (Jan. 2013), 263-287.

[38] Scherer, S., London, H., and Wolf, J. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7, 1 (June 1973), 31-44.

[39] Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. 2011. The Interspeech 2011 speaker state challenge. In *Proc. of the 12th Annual Conf. of the Int'l Speech Communication Association*. Interspeech '11, 3201-3204.

[40] Stern, S., Mullennix, J., and Wilson, S. 2002. Effects of perceived disability on persuasiveness of computer-synthesized speech. *Journal of Applied Psychology* 87, 2 (Apr. 2002), 411-417.

[41] Voss, J. 2005. The science of persuasion: An exploration of advocacy and the science behind the art of persuasion in the courtroom, *Law and Psychology Review* 29 (2005), 301-327.

[42] Worchel, S., Andreoli, V., and Eason, J. 1975. Is the medium the message? A study of the effects of media, communicator, and message characteristics on attitude change. *Journal of Applied Social Psychology* 5, 2 (June 1975), 157-172.

[43] Yang, Y. and Pedersen, J. 1997. A comparative study on feature selection in text categorization. In *Proc. of the 14th Int'l Conf. on Machine Learning*. ICML '97, 412-420.

[44] Young, J., Martell, C., Anand, P., Ortiz, P., and Gilbert, H. 2011. A microtext corpus for persuasion detection in dialog. In *Proc. of the AAAI-11 Workshop on Analyzing Microtext*, 80-85.