Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia

SUNGHYUN PARK, HAN SUK SHIM, MOITREYA CHATTERJEE, and KENJI SAGAE,
University of Southern California
LOUIS-PHILIPPE MORENCY, Carnegie Mellon University

Our lives are heavily influenced by persuasive communication, and it is essential in almost any type of social interaction 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 article, we introduce a newly created multimedia corpus of 1,000 movie review videos with subjective annotations of persuasiveness and related high-level characteristics or attributes (e.g., confidence). This dataset will be made freely available to the research community. We designed our experiments around the following five main research hypotheses. First, we study if computational descriptors derived from verbal and nonverbal behavior can be predictive of persuasiveness. We further explore combining descriptors from multiple communication modalities (acoustic, verbal, para-verbal, and visual) for predicting persuasiveness and compare with using a single modality alone. Second, we investigate how certain high-level attributes, such as credibility or expertise, are related to persuasiveness and how the information can be used in modeling and predicting persuasiveness. Third, we investigate differences when speakers are expressing a positive or negative opinion and if the opinion polarity has any influence in the persuasiveness prediction. Fourth, we further study if gender has any influence in the prediction performance. Last, we test if it is possible to make comparable predictions of persuasiveness by only looking at thin slices (i.e., shorter time windows) of a speaker's behavior.

Categories and Subject Descriptors: **H.1.2** [**Models and Principles**]: User/Machine Systems—*Human information processing*; **J.4** [**Computer Applications**]: Social and Behavioral Sciences—*Psychology, sociology*; **I.2.1** [**Artificial Intelligence**]: Applications and Expert Systems; **I.5.4** [**Pattern Recognition**]: Applications—*Signal processing*

General Terms: Algorithms, Experimentation, Human Factors, Theory

Additional Key Words and Phrases: Persuasion, persuasiveness, multimodal prediction, multimodal behavior, multimodal analysis, social multimedia, POM corpus, POM dataset, persuasive opinion multimedia corpus

The reviewing of this article was managed by associate editor Ed Chi.
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.
This article belongs to Part 2 of the Special Issue on Highlights of ICMI 2014, edited by Ali Salah, Björn Schuller, Jeffrey F. Cohn, Oya Aran, and Louis-Philippe Morency. Its reviewing was managed by regular TiiS associate editor Ed Chi, because one of the authors is also an editor of the special issue.
Authors' addresses: S. Park, H. S. Shim, M. Chatterjee, and K. Sagae, Institute for Creative Technologies / Computer Science Department, University of Southern California, 12015 Waterfront Dr, Playa Vista, CA 90094; emails: sunghyup@usc.edu, {hshim, mchatterjee, sagae}@ict.usc.edu; L.-P. Morency, Language Technology Institute, School of Computer Science, Carnegie Mellon University, Gates-Hillman Center (GHC) Office 5411, 5000 Forbes Avenue, Pittsburgh, PA 15213; email: morency@cs.cmu.edu.

DOI: http://dx.doi.org/10.1145/2897739

ACM Reference Format:
Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. ACM Trans. Interact. Intell. Syst. 6, 3, Article 25 (October 2016), 25 pages.
DOI: http://dx.doi.org/10.1145/2897739

1. INTRODUCTION

Our daily lives are heavily influenced by persuasive communication. Making a convincing case in the courtroom [Voss 2005], seeking patients' compliance to medical advice [O'Keefe and Jensen 2007], advertising and selling products in business [Meyers-Levy and Malaviya 1999], and even interacting with our friends and family all have persuasion at the core of the interaction.

With the advent of the Internet and a 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 article.

This research has many practical implications from the human-computer interaction perspective. For one, an automatic technology that can analyze multimodal signals from a human user in real-time and predict his/her level of persuasiveness from behavioral and verbal indications can be useful as a training system. Such a system can help a speaker to behave as a more persuasive speaker and a better negotiator in daily interactions. Furthermore, such a system can be used as a filtering tool and aid a person with real-time analysis of online video and audio content.

While there has been a considerable amount of research on persuasion from the standpoints of psychology and social science, there has been very limited work investigating persuasion from the computational perspective and from the context of social multimedia. Fortunately, recent progress in computer vision and audio signal processing technologies [Degottex et al. 2014; Lao and Kawade 2005; Littlewort et al. 2011; Morency et al. 2008] enables automatic extractions 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 article, we introduce our newly created Persuasive Opinion Multimedia (POM) corpus consisting of 1,000 movie review videos with subjective annotations of persuasiveness (see Figure 1) as well as high-level related characteristics or attributes (e.g., confidence).[1] Our experimental analysis revolves around the following five main research hypotheses. First, we study if computational descriptors derived from verbal and nonverbal behavior can be predictive of persuasiveness. We further explore combining descriptors from multiple communication modalities (acoustic, verbal, para-verbal, and visual) for predicting persuasiveness and compare with using a single modality alone. Second, we investigate how certain high-level attributes, such as credibility or expertise, are related to persuasiveness and how the information can be used in modeling and predicting persuasiveness. Third, we investigate differences when speakers are expressing a positive or negative opinion and if the opinion polarity has any influence in the persuasiveness prediction. Fourth, we further study if gender has any influence

[1]Researchers interested in the dataset can contact Sunghyun Park (park@ict.usc.edu) or Prof. Louis-Philippe Morency (morency@cs.cmu.edu, http://www.cs.cmu.edu/∼morency).

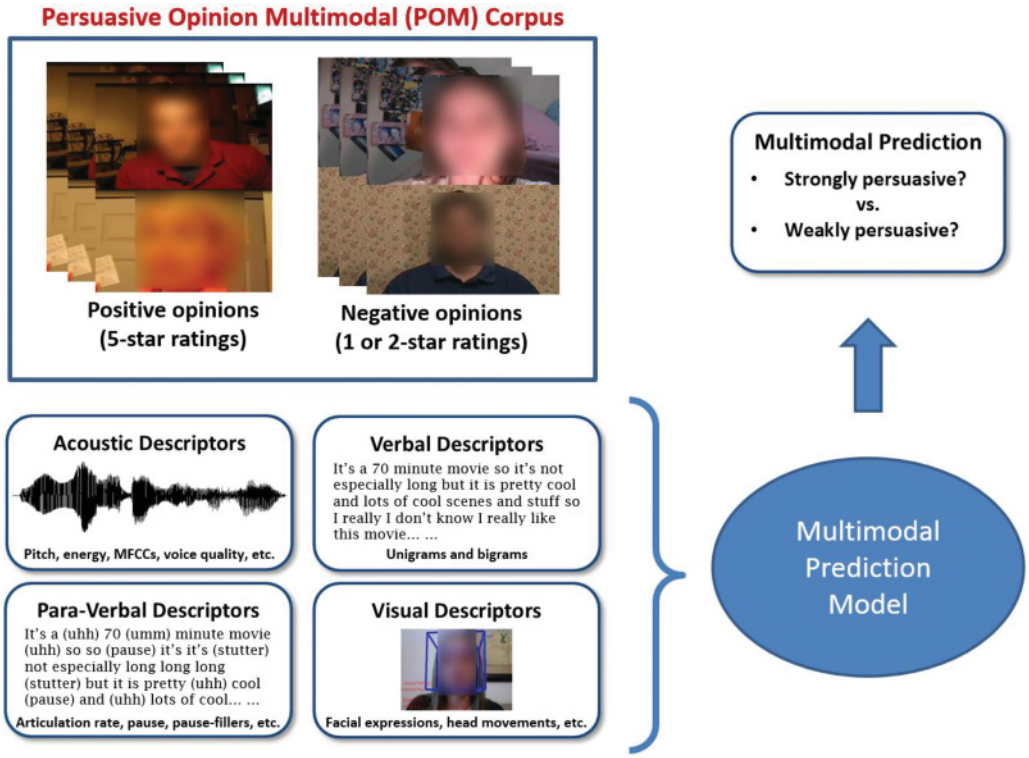

Fig. 1. Overview of the paper with a newly created multimedia corpus and our multimodal approach for predicting persuasiveness with acoustic, verbal, para-verbal, and visual computational descriptors.

in the prediction performance. Last, we test if it is possible to make comparable predictions 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 grounds and motivations to our work. In Section 3, we outline our research hypotheses, and Section 4 introduces our novel multimedia dataset designed for investigating persuasiveness in online social multimedia. We explain the design of our computational descriptors in Section 5 and experiments in Section 6. We report our results and discussions in Sections 7 and 8, and we conclude in Section 9.

2. THEORETICAL BACKGROUND

Persuasion in human communication has been a 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 findings that are only immediately relevant to our research problem. For an overview and history on persuasion research, interested readers are referred to other recent comprehensive texts [Crano and Prislin 2006; O'Keefe 2002; Perloff 2010].

In social psychology, dual process models of persuasion [Chaiken et al. 1989; Petty and Cacioppo 1986] 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 article can be seen in light of the dual process models with the focus on the peripheral route of information processing.

2.1. Tele-Mediated Video Interaction vs. Face-to-Face Interaction

Since tele-mediated video interaction (e.g., online videos with webcams) very closely simulates face-to-face interaction, both types of interaction can be thought of as having more or less the same interactional influence. Past research in the literature also supports it to some extent [Campbell 1998; Williams 1977]. However, there are also many studies that indicate their differences. Chen [2003] argues that visual conversational cues such as lip movements and eye contact can be subtly distorted in videoconferencing systems, causing negative attributes to be associated with the interacting partners. Furthermore, Storck and Sproull [1995] showed that people form less positive impressions of their interactional partners in video conferences, and Fullwood [2007] also found that video-mediated systems cause people to be perceived as less likable and intelligent. We note that the scope of this article is in online persuasion using tele-mediated online videos. Although online persuasion would share many similarities with face-to-face persuasion, potential differences should be kept in mind when applying the experimental results in this article in a face-to-face setting.

2.2. Modality Influence and Human Perception

Human communication is comprised of multiple modalities including acoustic, verbal, and visual channels, and it is apparent that each modality has its own separate influence on human perception. Mehrabian [1971] 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 and Eagly [1983] showed different influences on persuasion and comprehension when a message was delivered through the written, audiotaped, or videotaped modality. Worchel et al. [1975] also studied effects on persuasion when a message was delivered with different types of media, communicators, and positions.

2.3. Acoustic Perspective

Showing the importance of acoustic cues in human speech, Stern et al. [2002] reported that natural speech was more persuasive and taken more favorably than computer-synthesized speech. In addition, Mehrabian and Williams [1969] reported that more intonation and higher speech volume contributed to perceived persuasiveness, Pittam [1990] studied the relationship between nasality and perceived persuasiveness with a group of Australian speakers, Burgoon et al. [1990] found a positive correlation between vocal pleasantness and perceived persuasiveness, and Pearce and Brommel [1972] 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.

2.4. Verbal Perspective

There are many components in the verbal domain that have strong relationship with persuasiveness [Hosman 2002; Young et al. 2011]. 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.

2.5. 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 [1969] reported that higher speech rate and less halting speech contributed to perceived persuasiveness, Miller et al. [1976] reported that a rapid speech rate positively influenced persuasion, and Pearce and Brommel [1972] 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.

2.6. 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 [1969] 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 [1975] also found a similar set of nonverbal behavior related to persuasiveness that he calls affiliative nonverbal behavior. Moreover, Burgoon et al. [1990] found that greater perceived persuasiveness correlated with kinesic/proxemic immediacy, facial expressiveness, and kinesic relaxation. Rosenfeld [1966] found that the level of persuasiveness was positively correlated with positive head nods and negatively correlated with self-manipulations.

2.7. High-Level Attributes Related to Persuasion

Researchers investigating persuasion long knew that it was a complex phenomenon involving multiple dimensions, or high-level characteristics or attributes of a speaker, such as his/her level of credibility or confidence. For instance, many researchers identified that a message's persuasiveness partially depended on its source, which comprised of multiple dimensions such as credibility, a high-level attribute that is particularly known to be similar across cultures in its relationship with persuasiveness. More interested readers can find a review of persuasiveness and source credibility by Pornpitakpan [2006]. Similarly, there are multiple attributes that have been under study in relation to persuasiveness, such as attractiveness, likableness, confidence, expertise, message vividness, etc. [Chaiken 1979; LaCrosse 1975; Carli et al. 1995; Maslow et al. 2011; Inglis and Mejia-Ramos 2009; Maddux and Rogers 1980; Frey and Eagly 1993].

2.8. Thin Slice Prediction

Ambady and Rosenthal [1992] showed that much inference is possible just by observing "thin slices" of nonverbal behavior, and Curhan and Pentland [2007] applied the idea in a simulated employment negotiation scenario. 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.

2.9. Contributions

To our knowledge, our new corpus is the first multimedia dataset created with annotations for studying persuasiveness in online social multimedia. Furthermore, another main novelty of our work lies in investigating computational models of persuasiveness that take advantage of several natural multimodal communicative modalities encompassing acoustic, verbal, para-verbal, and visual channels.[2] In addition to providing

[2]In our previous works [Chatterjee et al. 2014; Shim et al. 2015], we focused only on a subset of these modalities, and our more comprehensive multimodal work [Park et al. 2014] was also limited in terms of the dataset and the scope of experiments.

an extensive set of experiments for computationally modeling persuasiveness, we also introduce a novel attribute-based multimodal fusion approach in which we use various high-level attributes related to persuasion in the middle layer for predicting persuasiveness.

3. RESEARCH HYPOTHESES

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

3.1. Computational Descriptors (Multimodal vs. Unimodal Prediction)

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 strongly persuasive or weakly persuasive. In particular, we hypothesize that combining computational descriptors derived from multiple communication modalities can make more accurate predictions compared to using those from a single modality alone from the acoustic, verbal, para-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.

3.2. Attribute-Based Multimodal Approach

Past research findings and intuition both tell us that several high-level attributes, such as credibility and expertise, are very likely to have close relevance to persuasiveness. And there can be a handful of key high-level attributes, each of which is a critical and distinct component in shaping a speaker's persuasiveness. We hypothesize that we can achieve better performance in predicting the level of persuasiveness by first using multimodal computational descriptors to predict the levels of such high-level attributes in the middle layer and subsequently predicting the level of persuasiveness from the refined, higher-level information.

Hypothesis 2 (H2). Using multimodal computational descriptors of verbal and nonverbal behavior to predict the levels of key high-level attributes related to persuasiveness and then subsequently using the intermediate information to predict a speaker's persuasiveness yield better performance compared to directly predicting persuasiveness from the computational descriptors.

3.3. Effect of Opinion Polarity

Persuasion can happen in a variety of contexts, and it is likely that we change our behavior depending on the context in our persuasion attempt. For instance, we might nod our head more when we try to persuade someone to go watch a particular movie, while we shake our head more in the opposite case. 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 3 (H3). Opinion polarity (sentiment) dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are polarity independent.

3.4. Effect of Gender

Gender can have an influence on how a speaker behaves in his/her persuasion endeavor. For instance, female speakers might be more verbally descriptive while male speakers are less expressive overall. We hypothesize that same gender speakers have more similarity in their behavior, allowing gender-dependent computational models to better capture the difference between strongly persuasive and weakly persuasive speakers.

Hypothesis 4 (H4). Gender-dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are gender independent.

3.5. 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 (see Section 2.7), we hypothesize that by looking at verbal and nonverbal behavior at specific shorter time periods, we can still make comparable predictions of persuasiveness of a speaker in social multimedia compared to making predictions based on the entire length of the speaker's behavior.

Hypothesis 5 (H5). Computational descriptors derived from a thin slice time period can make comparable predictions of a speaker's persuasiveness compared to those derived from the entire length of his/her video.

4. PERSUASIVE OPINION MULTIMEDIA (POM) CORPUS

Since there is currently no suitable corpus in the research community to study persuasiveness in the context of online social multimedia,[3] we found ExpoTV.com to be a good source to create a new corpus for our research topic. 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 a 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 are interested in exploring the difference in behavior between the cases when a speaker is trying to persuade the audience positively and negatively (see Section 3.3), we collected a total of 1,000 movie review videos as follows:

—*Positive Reviews*. 500 movie review videos with 5-star rating (306 males and 194 females).
—*Negative Reviews*. 500 movie review videos with 1- or 2-star rating, consisting of 208 1-star videos (145 males and 63 females) and 292 2-star videos (218 males and 74 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 93 seconds with the standard

[3]To our knowledge, currently the most relevant dataset to ours is a dataset of online conversational videos of vloggers by Biel et al. [2012]. It was not completely suitable for our research purpose because it was created for studying personality and the topics were too broad. Researchers interested in our new Perrsuasive Opinion Multimedia (POM) corpus can contact Sunghyun Park (park@ict.usc.edu) or Prof. Louis-Philippe Morency (morency@cs.cmu.edu, http://www.cs.cmu.edu/∼morency).

deviation of about 31 seconds. The corpus contains 352 unique speakers and 610 unique movie titles, including all types of common movie genres.

4.1. Subjective Annotations and Quality Assurance

Amazon Mechanical Turk (AMT) [Mason and Suri 2012], which is a popular online crowd-sourcing platform, was used to obtain subjective evaluations of the speaker in each video. Each video received three repeated annotations from three different workers, making the total number of our complete annotations 3,000 instances (or HITs) for 1,000 videos in the corpus. To minimize gender influence, all the annotations were distributed such that the workers only evaluated the speakers of the same gender.

4.1.1. Worker Demographics. A total of 87 workers participated in the annotation process with each worker annotating about 35 videos ($M = 34.5$, $SD = 13.4$). All the annotations were obtained from native English-speaking workers based in the United States. Out of 87 workers, 56 were male and 31 female workers. The workers predominantly identified themselves as White/Caucasian and were from their 20s and 30s. Specifically, 74 workers identified themselves as White/Caucasian, 5 workers as Hispanic, 4 workers as African-American, and 4 workers as Asian. In addition, 2 workers were between the age of 10∼19, 44 workers between 20∼29, 19 workers between 30∼39, 10 workers between 40∼49, 11 workers between 50∼59, and 1 worker between 60∼69.

4.1.2. Minimum Prior Bias. To ensure that no prior knowledge of the movies biased how the annotators rated each speaker's level of persuasiveness and other high-level characteristics, the annotations were obtained in two separate phases. In the first phase, a total of 49 workers participated in the evaluation process online, and the task was evenly distributed among them. During this first phase, the workers were asked if they had previously seen the movie being reviewed for each video. This information was then used to filter out all such annotations that were made with prior knowledge of the movie under review (about one-third of all the annotations). In the second phase, a total of 38 workers participated, first indicating which movies they had seen or not seen from a list of movies that we needed to re-annotate. Then, the annotation task was distributed as evenly as possible among the workers such that each worker only annotated those videos that discuss movies that he/she had not seen before.

4.1.3. Worker Quality Control and Assurance. The workers were screened and selected with very rigorous criteria. Beside 7 workers whose exact working statistics were unknown (who all nevertheless had a history of at least 99% approval rating and more than 1,000 HITs approved due to our default worker requirements), the workers on average had 99.9% approval rating and more than 44,000 approved HITs submitted, suggesting their track records of faithful and attentive works on AMT. Furthermore, there were many devices in our annotation HITs to ensure workers' attention and faithfulness. When working on a HIT, the workers were not able to skip any part of the movie review video except for rewinding several seconds. Once the video completed, the workers could not go back to see the video and were asked several questions throughout the HIT to make sure they paid attention to detail. For instance, the workers were asked to identify the speaker's age and recommendation level in the video. Furthermore, the workers were asked several open-ended questions to briefly write the speaker's reasons for recommending or not recommending a movie and why the workers thought that he/she was persuasive or unpersuasive. Each HIT webpage also had hidden features flagging those workers who tried to skip the video or submit the HIT without making sure that they have answered all the questions.

How interested are you in watching this movie? (asked before watching the video and the speaker)	After seeing this movie review, how interested are you in watching this movie?
O 1: Not interested	O -3: Much less interested than before
O 2	O -2: Less interested than before
O 3: A little interested	O -1: A little less interested than before
O 4	O 0: The same as before
O 5: Interested	O +1: A little more interested than before
O 6	O +2: More interested than before
O 7: Strongly interested	O +3: Much more interested than before
How persuasive was the reviewer?	How confident was the reviewer?
O 1: Not persuasive	O 1: Not confident
O 2	O 2
O 3: A little persuasive	O 3: A little confident
O 4	O 4
O 5: Persuasive	O 5: Confident
O 6	O 6
O 7: Strongly persuasive	O 7: Very confident

Fig. 2. Snippets of the questions used for annotating the POM dataset on each speaker's level of persuasiveness and other high-level attributes.

4.2. Persuasiveness and High-Level Attributes

In addition to investigating persuasiveness, another goal of our dataset was to better understand other high-level attributes that could be related to persuasiveness (e.g., personality traits). We believe that the extra annotations on the high-level attributes will make the corpus more widely applicable for other related research topics (e.g., personality trait modeling).

For each video in the corpus, a worker's complete annotation set comprised of watching a movie review video followed by 26 short questions, mostly multiple-choice with several open-ended text-input questions. The workers on average took 10.2 minutes to complete an annotation set or HIT on a video. As shown in Figure 2, we obtained annotations on the level of persuasiveness of the speaker by asking the workers to give a 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 a 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.

Asking the workers several open-ended questions, as mentioned in the previous subsection, served another purpose other than quality control. For instance, requiring the workers to write about the speaker's reasons for recommending or not recommending a movie prompts the workers to turn on their cognitive and logical evaluation. The same holds for another question asking why the workers thought that the speaker was persuasive or unpersuasive in the movie review video. In our annotations, such open-ended questions were another step to ensure that the workers' evaluations were similar to a real-life scenario of seeking and judging others' advice on movies, not just trying to hastily finish their tasks in the experimental setting. In other words, the questions were meant to prompt the workers not to completely and blindly depend on peripheral or heuristic cues when judging the speaker's level of persuasiveness and other high-level attributes.

Table I. Krippendorff's Alpha Agreement for the Annotations of Persuasiveness, Other Related High-Level Attributes, and the Big Five Personality Dimensions

Attribute	Kripp. alpha	Attribute	Kripp. alpha
Confident	0.74 (0.36)	Passionate	0.76 (0.40)
Credible	0.69 (0.26)	Professional-looking	0.71 (0.32)
Dominant	0.69 (0.28)	Vivid	0.68 (0.25)
Entertaining	0.68 (0.25)	Voice pleasant	0.69 (0.27)
Expert	0.69 (0.27)	Physically attractive	0.73 (0.35)
Humorous	0.67 (0.32)	**Persuasive**	**0.69** (0.26)
Agreeableness	0.65 (0.20)	Openness	0.67 (0.23)
Conscientiousness	0.71 (0.32)	Neuroticism	0.64 (0.18)
Extraversion	0.75 (0.38)		

Each alpha shows the mean agreement between the ground-truth used (measured as the mean of three ratings) and each of the three raters across 1,000 videos in the dataset. The alphas in the parentheses show agreement among the raw ratings.

For evaluating personality, a 10-item version of the Big Five Inventory [Rammstedt and John 2007] was used to assess the personality of the speaker in each video. We note that this instrument of evaluating personality is originally designed for self-reported data. In our experiments, the intended use would be for the speaker in each video to answer the 10-item questionnaire himself/herself. However, we also note that more relevant information for our research purpose is the speakers' perceived personality as seen by others rather than the speakers' actual personality. Our work in this article does not use the personality data other than introducing the information as part of the dataset, but future work using our dataset should take it into account. Other than the speakers' perceived personality, we also obtained self-assessed personality of the workers who performed the evaluations so that a future analysis is possible by investigating the relationship between the personality of the perceiver and that of 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.

4.3. 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 three repeated Likert-scale evaluations as the final measure for each video. Table I summarizes the mean agreement measured with Krippendorff's alpha [Krippendorff 2012] between our final measure and each coder. The agreement is generally high around 0.70. The agreement measured among raw ratings ranges from 0.20 to 0.40, suggesting variability in human perception and the challenge of the research topic.

Figure 3 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 [Crano and Prislin 2006; O'Keefe 2002; Perloff 2010]. 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.

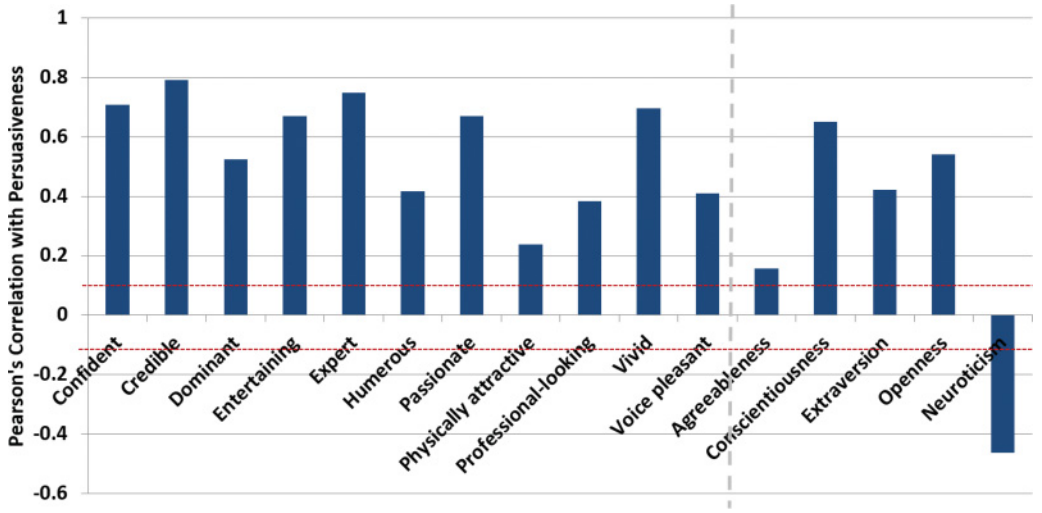

Fig. 3. Pearson's correlation coefficients between persuasiveness and high-level and personality attributes (after taking the mean of three repeated annotations). The two horizontal dotted lines indicate critical values at $p^{***} < 0.001$ for two-tailed probabilities, and the vertical dotted line visually divides the personality dimensions from other attributes.

To validate the persuasiveness measure, we included in the annotation tasks two questions related to the annotators' interest in watching the reviewed movies (see Figure 2). For the first question, the annotators were shown general information on the movie, including the synopsis and cast. Then, they were asked, "How interested are you in watching this movie?" This first question was answered before watching each review video. 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). Out of 3,000 annotation instances (1,000 movies multiplied by 3 for repeated annotations), our validity analysis shows a strong correlation between the persuasiveness score rating and the annotators' interest after watching the movie reviews, 0.71 for positive reviews and −0.56 for negative reviews (since viewers are discouraged to watch the movies for negative reviews).

4.4. Transcriptions

Using AMT and 17 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.

5. COMPUTATIONAL DESCRIPTORS

In this section, we give details on the extractions and computational encodings of multimodal descriptors as potential candidates for capturing persuasiveness. The verbal and para-verbal descriptors were computed automatically from the manual transcriptions. The acoustic and visual descriptors were also extracted automatically, directly from the audio and video streams. Table II summarizes all of our computational descriptors.

5.1. Acoustic Descriptors

Following common approaches for conducting automatic speech analysis [Schuller et al. 2011], we extracted various speech features related to pitch, formants, voice

Table II. 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
• Unigrams
• Bigrams
Para-Verbal
• 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

qualities and mel-frequency cepstral coefficients (MFCCs) using a publicly available software called Covarep [Degottex et al. 2014]. The raw feature values were then used to compute common statistical descriptors including the mean, median, percentiles, ranges, skewness, and standard deviation. The encoded descriptors were then used to explore their feasibility in capturing persuasiveness in acoustic signals of speech. All the descriptors had numerical values.

—*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*. Closely tied to the affective aspect of speech [Busso et al. 2009].
—*Voice Qualities*. Many studies show a strong relation between voice quality features and perceived emotion [Gobl and Chasaide 2003], 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 other works more focused on acoustic analysis [Kane et al. 2013a, 2013b; Scherer et al. 2013].

5.2. Verbal Descriptors

From the verbatim transcriptions of the dataset, we extracted all standard unigram and bigram features commonly used in natural language processing [Rosenfeld 2000], with the only difference in that the term frequencies were normalized by the video length.

5.3. Para-Verbal Descriptors

In addition, we observed a set of frequent para-verbal cues that could be associated with the level of persuasiveness. All the descriptors had numerical values.

—*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 [FaceFX] 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 [Mahl 1956]. 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.

5.4. Visual Descriptors

Using readily available visual tracking technologies [Littlewort et al. 2011; Lao and Kawade 2005; Morency et al. 2008], we extracted frame-by-frame 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. All the descriptors had numerical values.

—*Discrete Emotions*. The level of anger, contempt, disgust, fear, joy, sadness, and surprise (mostly between −10 and 10).
—*Valence*. The level of high-level valence including negative, neutral, and positive valence (mostly between −10 and 10).
—*Facial Action Units*. The level of movement in various facial areas as codified by Facial Action Coding System (FACS) [Ekman 1997] including AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, and AU28 (mostly between −10 and 10).
—*Eye Gaze Movements*. Horizontal and vertical angles (mostly between −180 to 180).
—*Head Movements*. Horizontal, vertical, and rotational angles and displacements (mostly between −90 to 90).
—*Approximated Posture*. The movement in the z axis (toward or away from the camera, estimated by calculating face size in pixel).

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 (see Section 3).

6.1. Persuasiveness Labels

For our experiments, we explored two types of labels – discrete and continuous persuasiveness ratings. For the discrete labels, we tested with classifiers and we tested with regressors for continuous labels. For the regression experiments, we computed the ground-truth scores on all 1,000 videos by averaging the 3 repeated annotations

(see Section 4 for more detail about the annotation process). For the classification experiments, the ground-truth persuasiveness scores of equal to or greater than 5.5 were taken as strongly persuasive speakers and the scores of equal to or less than 2.5 weakly persuasive speakers. We note that this dataset trimming or selection process was done for the classification experiments so that we could primarily focus on investigating behavioral differences between strongly persuasive and weakly persuasive videos. After taking this discretization step, we ended up with a total of 253 videos. In terms of the opinion polarity, the final sample set comprised of 137 videos of positive reviews (63 strongly persuasive and 74 weakly persuasive) and 116 videos of negative reviews (61 strongly persuasive and 55 weakly persuasive). In terms of gender, the final sample set comprised of 152 videos of male reviewers (75 strongly persuasive and 77 weakly persuasive) and 101 videos of female reviewers (49 strongly persuasive and 52 weakly persuasive).

6.2. Experimental Conditions

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.

For the first hypothesis (H1), we explored both types of discrete and continuous persuasiveness labels. For both kinds of labels, we compared the performance of our multimodal approach of combining all descriptors at the feature level with the performance of using descriptors only from a single modality. In addition to investigating whether the multimodal models perform better than any unimodal ones, we have also tried all possible combinations of the modality groups for further analysis. The text below summarizes the experimental conditions of our prediction models designed to test H1:

—Acoustic descriptors only (see Section 5.1).
—Verbal descriptors only (see Section 5.2).
—Para-verbal descriptors only (see Section 5.3).
—Visual descriptors only (see Section 5.4).
—Multimodal descriptors. All computational descriptors concatenated together at the feature level.

Hypothesis 2 (H2). Using multimodal computational descriptors of verbal and nonverbal behavior to predict the levels of key high-level attributes related to persuasiveness and then subsequently using the intermediate information to predict a speaker's persuasiveness yield better performance compared to directly predicting persuasiveness from the computational descriptors.

For the second hypothesis (H2), we designed and investigated a new approach in fusing multimodal information in relation to several high-level attributes and persuasiveness (see Figure 4). We specifically selected those attributes that showed an absolute correlation of at least 0.5 with persuasiveness (see Figure 3), which came out to be seven speaker attributes: credible, expert, confident, vivid, passionate, entertaining, and dominant (we leave personality traits for future work). We first trained a regressor for each attribute, and the predicted regression level was then subsequently used to finally classify samples into strongly and weakly persuasive speakers. The performance of this new attribute-based approach was compared with that of the multimodal predictor just described above. The motivation behind this model design was that we suspected that it could be easier to predict the levels of high-level attributes compared to trying to predict persuasiveness directly. If so, we suspected that we could predict a speaker's persuasiveness more accurately by adding an intermediate abstract layer of first measuring high-level attributes that are closely related to persuasiveness.

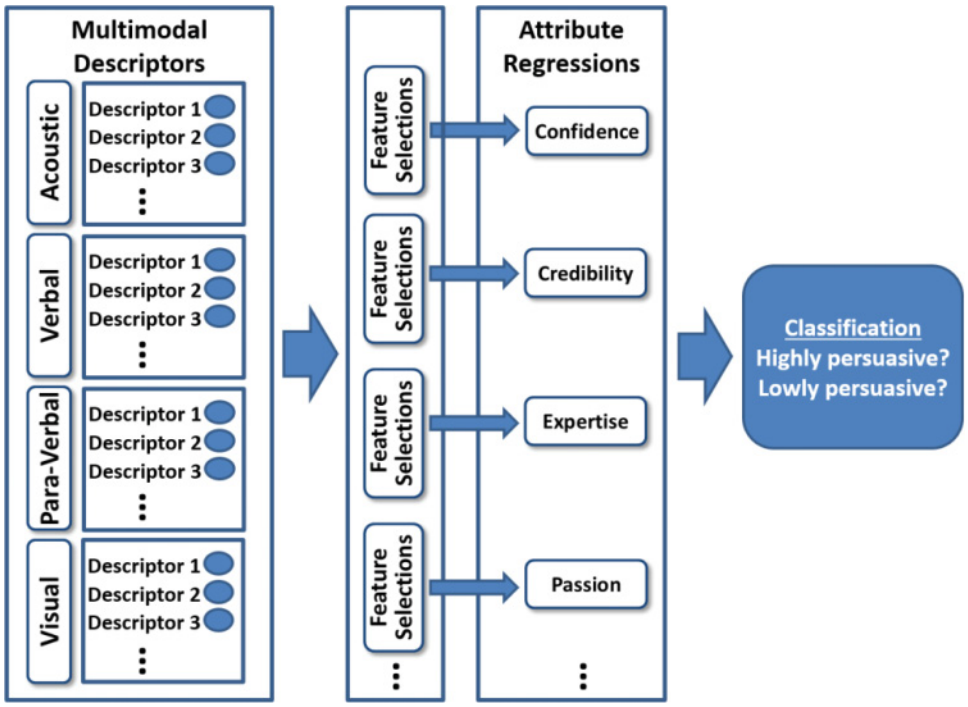

Fig. 4. Overview of our attribute-based multimodal prediction approach in which we use high-level attributes in the middle layer before predicting a speaker's level of persuasiveness.

Hypothesis 3 (H3). Opinion-polarity- (sentiment)-dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are polarity independent.

To address the third hypothesis (H3), we performed new classification experiments using all multimodal descriptors and grouping the dataset in three different ways depending on the opinion polarity expressed in the videos (as just explained in Section 6.1):

—*Positive Reviews Only (Sentiment-Dependent)*. Multimodal models were explored using only the 137 positive reviews from the trimmed samples of 253 videos.
—*Negative Reviews Only (Sentiment-Dependent)*. Multimodal models were explored using only the 116 negative reviews from the trimmed samples of 253 videos.
—*Both Review Types Combined (Sentiment-Independent)*. The same classification models as the multimodal models in H1, using all 253 trimmed samples.

Hypothesis 4 (H4). Gender-dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are gender independent.

To address the fourth hypothesis (H4), we performed new multimodal classification experiments using three different groups depending on gender of the reviewers:

—*Male Reviewers Only (Gender-Dependent)*. Multimodal models were explored using only the 152 reviews by male speakers from the trimmed samples of 253 videos.
—*Female Reviewers Only (Gender-Dependent)*. Multimodal models were explored using only the 101 reviews by female speakers from the trimmed samples of 253 videos.
—*Both Gender Reviewers Combined (Gender-Independent)*. The same classification models as the multimodal models in H1, using all 253 trimmed samples.

In regards to H3 and H4, we note that there are certainly many ways of using the gender and sentiment polarity information in the prediction models. For instance, one could simply use it in a combined model and encode gender and sentiment information into the feature space. In our work, there were several reasons for training separate models for each gender and sentiment type. First, gender and sentiment have the potential to completely alter human behavior. For instance, female population is essentially different from male population and their behavior is also likely to be quite different. People in a positive mood talking about something they like are also a population that is very different from those talking about something they dislike. By training separate models for each gender and sentiment type, we can potentially have a more accurate model for each specific population that we want to target. Second, having separate models can also give us more insights into understanding which behavioral cues are important for persuasion in different gender and sentiment contexts. For instance, in our same-gender evaluation design, male-gender models could allow us to understand which behavioral cues are particularly important for male speakers to be perceived as persuasive to male reviewers. The same holds for female speakers to female reviewers and also for the context of positive and negative sentiments. Third, from the perspective of building real-time systems, accurately identifying sentiment polarity and gender are themselves challenging research problems, and it is likely that such information may not be available as reliable features. However, we note that using external state-of-the-art predictors for encoding the gender and sentiment information as features in a single model and also in real-time (as similarly done with our para-verbal, visual, and acoustic descriptors) would be a future study that can strongly complement our work in this article.

Hypothesis 5 (H5). Computational descriptors derived from a thin-slice time period can make comparable predictions of a speaker's persuasiveness compared to those derived from the entire length of his/her video.

To address the last hypothesis (H5), we performed additional classification experiments using all the multimodal descriptors computed separately within different thin slices. More specifically, we divided each review video into 10 equal-length thin slices of first 10%, 10% to 20%, 20% to 30%, and so forth, and repeated the same classification experiments within each thin slice window. Furthermore, we also looked at the performance in progressive cumulative thin slices of first 5%, first 10%, first 15%, etc., to find out how soon the performance reaches that of using the whole 100% sessions. For the verbal descriptors, we estimated time using word count.

6.3. Methodology

For all the experiments, we used the support vector machines (SVMs for classification and SVRs for regression experiments) with the radial basis function kernel as the prediction models [Chang and Lin 2011]. The experiments were performed with 20-fold cross-validation (CV). Each CV experiment had 1-fold testing and 3-fold validation (among 19 training folds) for automatic selection of hyper parameters using a grid-search method as recommended by Chang and Lin [2011]. We emphasize that our folds were created such that no 2-folds-contained samples from the same speaker. This restriction assures speaker-independent experiments for better generalizability of our prediction models and results. Our evaluation metric was the averaged Pearson's correlations for regression experiments and averaged accuracies for classification experiments over all 20 testing folds.

For feature selection, we used top features as suggested by the absolute correlations for regressions and Information Gain (IG) metric for classifications [Yang and Pedersen 1997]. We limited the feature space to be always around or less than 1/10th of the sample size. For instance, in the classification experiments for H1 involving all the 253

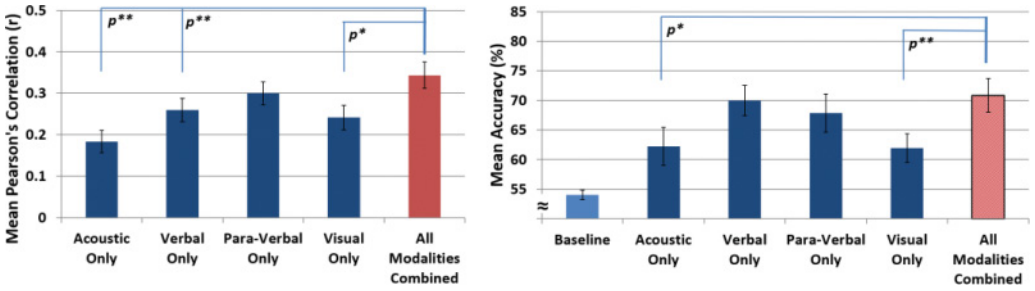

Fig. 5. Persuasiveness prediction results for the multimodal and unimodal models with the regression results on the left and the classification results on the right ($p^* < 0.05$ and $p^{**} < 0.01$). The error bars show ± 1 standard error.

video samples, the top 25 features were selected with highest IG scores. We emphasize that we performed feature selections only using the training samples for each CV experiment. None of the test samples were used for feature selection in each CV experiment. That is, we performed 20 separate vocabulary buildings of n-grams and 20 separate feature selections using only the training samples for the 20 CV experiments and testing. Since our CV folds were made speaker-independent, feature selections done strictly only on training samples and systematically using correlation and IG scores, and feature space always limited to 1/10th of the sample size, the chance of overfitting would be very low for our models.

7. RESULTS

In this section, we report our experimental results centered on our five main research hypotheses described in Section 3.

7.1. Multimodal Vs. Unimodal (H1)

The left graph in Figure 5 shows the regression results of predicting the continuous persuasiveness labels by each unimodal and the multimodal models. The multimodal models predicted the level of persuasiveness with a mean Pearson's correlation of 0.34, the acoustic descriptors only models with 0.18, the verbal descriptors only models with 0.26, the para-verbal descriptors only models with 0.30, and the visual descriptors only models with 0.24. The paired-sample t-tests showed that the performance of the multimodal models was better with a statistical significance compared with that of the acoustic descriptors only models ($p^{**} < 0.01$), the verbal descriptors only models ($p^{**} < 0.01$), and the visual descriptors only models ($p^* < 0.05$).

The right graph in Figure 5 shows the classification results of predicting between the strongly and weakly persuasive speakers by each unimodal and the multimodal models. The multimodal models predicted between the strongly and weakly persuasive speakers with a mean accuracy of 70.34%, the acoustic descriptors only models with 62.21%, the verbal descriptors only models with 69.98%, the para-verbal descriptors only models with 67.85%, and the visual descriptors only models with 61.94%. The paired-sample t-tests showed that the performance of the multimodal models was better with a statistical significance compared to that of the acoustic descriptors only models ($p^* < 0.05$) and the visual descriptors only models ($p^{**} < 0.01$). The majority baseline for the classification experiments was 55.02%.

7.2. Attribute-Based Multimodal Approach (H2)

Figure 6 shows the classification results of the attribute-based multimodal models, which performed at 76.03%. A paired-sample t-test didn't show a statistical significance

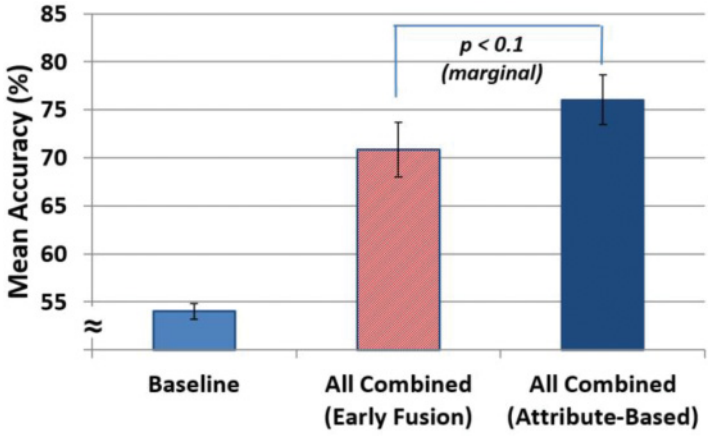

Fig. 6. Persuasiveness prediction results for two different multimodal approaches, the one combining all the descriptors at the feature level and the other using attribute-based fusion. The error bars show ±1 standard error.

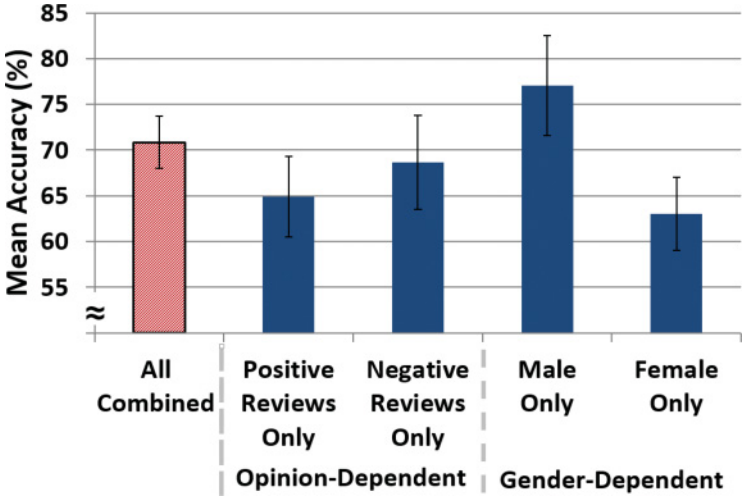

Fig. 7. Persuasiveness prediction results for the multimodal models when made opinion polarity-dependent and gender-dependent. The error bars show ±1 standard error.

(marginal at $p < 0.10$) between the performance of the attribute-based approach and that of the early-fusion approach at 70.85%.

7.3. Effect of Opinion Polarity (H3)

Figure 7 shows the classification results of the multimodal predictors across different conditions of positive reviews only, negative reviews only, and all reviews combined (the leftmost bar labeled "all combined"). Compared to the predictors using all the reviews at 70.85%, the predictors trained and tested using only the positive reviews performed at 64.91% and those trained and tested using only the negative reviews performed at 68.65%.

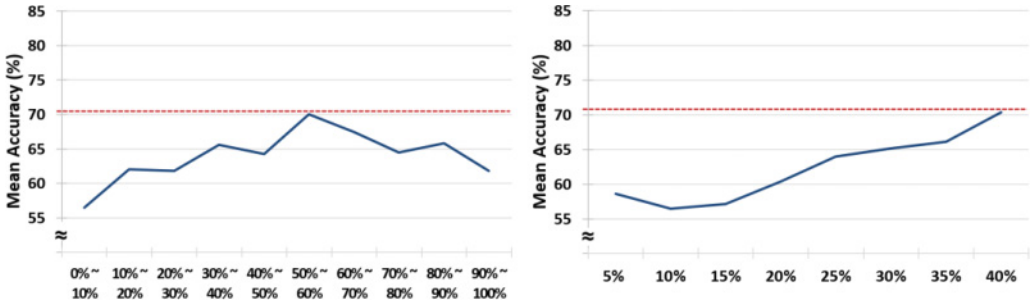

Fig. 8. Persuasiveness prediction results for various thin slices. The left graph shows the thin-slice results of using computational descriptors encoded from the length of only 1/10th of each review session, and the right graph shows the results for cumulative thin-slice windows (i.e., first 5% of the session, first 10%, first 15%, and so on). The dotted line in each graph indicates the prediction level for the multimodal approach in H1 using computational descriptors from all modalities and the whole 100% session.

7.4. Effect of Gender (H4)

Figure 7 shows the classification results of the multimodal predictors across different gender conditions of male reviewers only, female reviewers only, and all reviewers combined (the same leftmost bar labeled "all combined"). We note that our current work is limited to the same-gender design in evaluating a speaker's persuasiveness and other high-level attributes. Male speakers were rated only by male annotators and female speakers only by female annotators. Compared to the predictors using all the reviewers at 70.85%, the predictors trained and tested using only the male reviewers performed at 77.05% and those trained and tested using only the female reviewers performed at 63.01%.

7.5. Thin Slice Prediction (H5)

Figure 8 shows the classification results of the all-modalities predictors across different thin slices. Compared with the prediction accuracy of 70.85% when using the whole length of each review, using 1/10th of the session mostly yielded between 60% and 70% prediction accuracy, with the highest prediction in the 50%∼60% session thin slice that performed at 70.02% prediction accuracy. The cumulative thin slice results show that the prediction performance reached that of using the whole session when the cumulative thin slice was taken up to the 40% of the session from the beginning, performing at 70.33% prediction accuracy.

8. DISCUSSIONS

In this section, we discuss and interpret our experimental results centered on our five main research hypotheses described in Section 3. These discussions are followed by an analysis of the multimodal descriptors.

8.1. Limitations

One limitation of our current research work is in its same-gender design in evaluating a speaker's persuasiveness and other high-level attributes. Male speakers were rated only by male annotators and female speakers only by female annotators. There may be interesting effects due to gender influence, and we note that a cross-gender study as future work would strongly complement our findings in this article.

In light of the dual process models of persuasion [Chaiken et al. 1989; Petty and Cacioppo 1986], we note that the context of movie review videos might trigger not only the central route of information processing related to cognition and logical reasoning, but also the peripheral or heuristic route such as looking at the message

source's credibility. Our experimental design in annotating the dataset for the level of persuasiveness and other high-level attributes ensured the annotators' faithfulness and careful attention, at the same time prompting them to also engage in cognitive thoughts. Although it cannot be sure which route of information processing was mainly used by the annotators, the context and our design most likely triggered both routes in combination, simulating a real-life scenario of how a person would perceive a speaker's persuasiveness and get influenced by an online movie review video.

Another limitation of our work involves the inherent variability in human perception and judgment. Our final persuasiveness measure using the mean score of three repeated Likert-scale evaluations shows Krippendorff's alpha of 0.69 compared with each individual coder's evaluations (see Table I). Although this final ground-truth measure we use shows relatively high agreement with each coder, the agreement measured among raw ratings themselves is at 0.26, which suggests much variability in average human perception of persuasiveness. Future studies could also obtain evaluations from the coders trained in a specific way, but it needs careful attention since trained evaluations could be different from average human perception of persuasiveness.

8.2. Multimodal vs. Unimodal (H1)

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.

For both the regression and classification results, our first hypothesis partially confirmed that the multimodal information improve the prediction performance compared to that of using unimodal information, especially with statistical significance for the acoustic only or visual only information. However, for the regression results, there was no statistical significance between the multimodal models and the para-verbal only models. For the classification results, the multimodal models also similarly performed better but did not show any statistical significance compared to that of the verbal only models and the para-verbal only models. For the performance difference of the verbal only models in regression and classification, one possible explanation is in the values of the *n*-gram features having very limited ranges and non-continuous numerical values depending on term frequency. Such feature representation might have imposed restrictions on the verbal only models in the regression experiments.

Our results suggest that especially the para-verbal behavioral cues, captured in the form of computational descriptors, are powerful in predicting persuasiveness. Table III summarizes both the regression and the classification results in all possible combinations of the modality groups for more detailed analysis of multimodal information fusion. We observed that combining all four modalities was not necessarily better than using a subset of them, especially since the para-verbal descriptors were very powerful.

8.3. Attribute-Based Multimodal Approach (H2)

Hypothesis 2 (H2). Using multimodal computational descriptors of verbal and nonverbal behavior to predict the levels of key high-level attributes related to persuasiveness and then subsequently using the intermediate information to predict a speaker's persuasiveness yield better performance compared to directly predicting persuasiveness from the computational descriptors.

The motivation behind the attribute-based approach was to use more information by breaking down a speaker's persuasiveness into several dimensions. For instance, a speaker may be persuasive particularly based on his/her level of credibility or passion, and such information also has the potential benefit of providing a deeper understanding of why he/she is more or less persuasive.

Table III. Multimodal Prediction Results Using Computational Descriptors in All Combinations of Modalities

Early fusion sources (• signifies inclusion)				Regression	Classification
Acoustic	Verbal	Para-verbal	Visual	(Pearson's correlation r)	Accuracy (%)
•	•	•	•	0.34	70.85
•	•	•		0.31	70.45
•	•		•	0.32	69.77
•		•	•	0.31	71.27
	•	•	•	0.34	70.34
•	•			0.27	67.26
•		•		0.26	67.49
•			•	0.25	65.40
	•	•		0.32	71.05
	•		•	0.28	66.83
		•	•	0.31	68.56
•				0.18	62.21
	•			0.26	69.98
		•		0.30	67.85
			•	0.24	61.94

A paired-sample t-test showed that the difference of performance between the attribute-based approach (N = 20, M = 76.03, SD = 11.63) and the early-fusion approach using all modalities (N = 20, M = 70.85, SD = 12.69) was not statistically significant, $t(19) = -1.89$, $p = 0.07$, 95% CI [−0.57, 10.94]. The hypothesis was not confirmed and no further conclusions could be drawn from the results. More analysis, model improvements, and experiments as future work would provide more conclusive insights on how to best use the attribute-based approach.

8.4. Effect of Opinion Polarity (H3)

Hypothesis 3 (H3). Opinion polarity (sentiment)-dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are polarity independent.

Our experiments did not support the third hypothesis, and opinion polarity-dependent classifiers did not show any improvement in the performance. None of the results showed statistical significance however, and no conclusions could be drawn from the results. We suspect that the reduced sample sizes for training the opinion-dependent models could have been the cause of relatively reduced performance. We also cannot rule out the possibility that the behavior change is not significant enough to give an advantage of opinion-dependent modeling.

8.5. Effect of Gender (H4)

Hypothesis 4 (H4). Gender-dependent models perform better in predicting a speaker's level of persuasiveness compared to those that are gender-independent.

Our experiments did not conclusively support the fourth hypothesis, and gender-dependent classifiers did not necessarily show any improvement in the performance. Although the male-only classifiers did show some improved performance compared to the all-reviewers classifiers, the results did not show any statistical significance and no conclusions could be drawn from the results. One possible cause of the female-reviewers classifiers performing relatively poorly compared to the male-reviewers classifiers could be due to the difference in the sample sizes. The male-reviewers classifiers had a sample size that was 50% greater than that for training female-reviewers classifiers, and such small sample size could have resulted in models that were not generalized enough.

Table IV. Top Computational Descriptors in Each Modality for Predicting between Strongly and Weakly Persuasive Speakers

Descriptors	Info Gain
Acoustic	
F2: range (min ∼ max)	0.09
Peak Slope: range (25th ∼ 75th percentile)	0.08
MFCC4: 25th percentile	0.07
MFCC2: range (10th ∼ 90th percentile)	0.07
MFCC4: mean	0.06
Para-Verbal	
Pause	0.20
Visual	
Gaze movement (up / down): range (25th ∼ 75th percentile)	0.11
Gaze movement (up / down): range (10th ∼ 90th percentile)	0.09
Gaze movement (up / down): 25th percentile	0.08
Surprise: range (min ∼ max)	0.06
AU20: 75^{th} percentile	0.06

8.6. Thin Slice Prediction (H5)

Hypothesis 5 (H5). Computational descriptors derived from a thin slice time period can make comparable predictions of a speaker's persuasiveness compared to those derived from the entire length of his/her video.

The results are a typical demonstration of the idea of thin slices and suggest that we can still make much inference on a speaker's persuasiveness just by looking at a smaller window of behavior. It is particularly interesting that only looking at 1/10th of a movie review, especially toward the middle, is enough to reasonably predict the speaker's level of persuasiveness.

8.7. Descriptor Analysis

Table IV highlights several top descriptors that have been particularly discriminative in separating strongly persuasive and weakly persuasive speakers. The verbal modality was not included in the analysis due to the nature of the bag-of-words descriptors that they are useful collectively.

From the acoustic modality, the ranges in the second formant and the peak slope voice quality were particularly useful in the classification experiments. MFCC descriptors in the low-frequency regions also stood out for predicting persuasive speakers, which were expected to perform better than high-frequency regions due to denser resolutions and being more robust to noise. Consistent with the literature described in Section 2, the para-verbal descriptor of pause proved to show much discriminative power in separating speakers who are perceived as strongly persuasive and weakly persuasive. From all the descriptors and from all the modalities combined, this descriptor was the single most predictive cue. From the visual modality, the descriptors from the gaze were predominant followed by those from discrete emotion of surprise and AU20 (lip stretcher).

9. CONCLUSIONS AND FUTURE WORK

We introduced a novel multimedia corpus specifically designed to study persuasiveness in the context of social multimedia. We presented our computational approaches in using verbal and nonverbal behavior from multiple channels of communication to predict a speaker's persuasiveness in online social multimedia content and showed a novel approach of using high-level attributes related to persuasion in predicting the level of persuasiveness. Furthermore, we demonstrated that the idea of thin slices can be used to observe a short window of a speaker's behavior 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 to understand relationship between persuasiveness and relevant high-level attributes including personality.

ACKNOWLEDGMENTS

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

REFERENCES

Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psych. Bull.* 111, 2 (Mar. 1992), 256–274. DOI:http://dx.doi.org/10.1037/0033-2909.111.2.256

Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: Predicting personality from facial expression of emotion in online conversational video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI'12)*. ACM Press, New York, 53–56. DOI:http://dx.doi.org/ 10.1145/2388676.2388689

Judee Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. *Hum. Commun. Res.* 17, 1 (Sept. 1990), 140–169. DOI:http://dx.doi.org/10.1111/j.1468-2958.1990.tb00229.x

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Trans. Audio, Speech Lang. Process.* 17, 4 (May 2009), 582–596. DOI:http://dx.doi.org/ 10.1109/TASL.2008.2009578

John Campbell. 1998. Participation in videoconferenced meetings: User disposition and meeting context. *Inf. Manage.* 34, 6 (Dec. 2006), 329–338. DOI:http://dx.doi.org/10.1016/S0378-7206(98)00073-1

Linda Carli, Suzanne LaFleur, and Christopher Loeber. 1995. Nonverbal behavior, gender, and influence. *J. Personal. Social Psych.* 68, 6 (Jun. 1995), 1030–1041. DOI:http://dx.doi.org/10.1037/0022-3514.68.6.1030

Shelly Chaiken. 1979. Communicator physical attractiveness and persuasion. *J. Personal. Social Psych.* 37, 8 (Aug. 1979), 1387–1397. DOI:http://dx.doi.org/10.1037/0022-3514.37.8.1387

Shelly Chaiken and Alice Eagly. 1976. Communication modality as a determinant of message persuasiveness and message comprehensibility. *Journal of Personality and Social Psychology* 34, 4 (Oct. 1976), 605–614. DOI:http://dx.doi.org/10.1037/0022-3514.34.4.605

Shelly Chaiken, Akiva Liberman, and Alice Eagly. 1989. Heuristic and systematic information processing within and beyond the persuasion context. In *Unintended Thought: Limits of Awareness, Intention, and Control*, James Uleman and John Bargh (Eds.). Guildford Press, New York, 212–252.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machine. *ACM Trans. Intell Syst Tech.* 2, 3 (Apr. 2011), 27, 1–27, 25. DOI:http://dx.doi.org/10.1145/1961189.1961199

Moitreya Chatterjee, Sunghyun Park, Han Suk Shim, Kenji Sagae, and Louis-Philippe Morency. 2014. Verbal behaviors and persuasiveness in online multimedia content. In *Proceedings of the 2nd Social NLP Workshop (SocialNLP'14)*, 50–58.

Milton Chen. 2003. *Conveying Conversational Cues through Video*. Ph.D. Dissertation. Stanford University, Palo Alto, CA.

William Crano and Radmila Prislin. 2006. Attitudes and persuasion. *Ann. Rev. Psych.* 57 (Jan. 2006), 345–374. DOI:http://dx.doi.org/10.1146/annurev.psych.57.102904.190034

Jared Curhan and Alex Pentland. 2007. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *J. Appl. Psych.* 92, 3 (May 2007), 802–811. DOI:http://dx.doi.org/10.1037/0021-9010.92.3.802

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*. IEEE, 960–964. DOI:http://dx.doi.org/10.1109/ICASSP.2014.6853739

Paul Ekman. 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.

FaceFX. http://www.facefx.com/.

Kurt Frey and Alice Eagly. 1993. Vividness can undermine the persuasiveness of messages. *J. Personal. Social Psych.* 65, 1 (July 1993), 32–44. DOI:http://dx.doi.org/10.1037/0022-3514.65.1.32

Chris Fullwood. 2007. The effect of mediation on impression formation: A comparison of face-to-face and video-mediated conditions. *Appl. Ergon.* 38, 3 (May 2007), 267–273. DOI:http://dx.doi.org/10.1016/j.apergo.2006.06.002

Christer Gobl and Ailbhe Chasaide. 2003. The role of voice quality in communication emotion, mood and attitude. *Speech Commun.* 40, 1–2 (Apr. 2003), 189–212. DOI:http://dx.doi.org/10.1016/S0167-6393(02)00082-1

Lawrence Hosman. 2002. Language and persuasion. In *The Persuasion Handbook: Developments in Theory and Practice*, James Dillard and Michael Pfau (Eds.). New York: Sage, 371–390.

Matthew Inglis and Juan Mejia-Ramos. 2009. The effect of authority on the persuasiveness of mathematical arguments. *Cognit. Instruct.* 27, 1 (2009), 25–50. DOI:http://dx.doi.org/10.1080/07370000802584513

John Kane, Stefan Scherer, Matthew Aylett, Louis-Philippe Morency, and Christer Gobl. 2013a. Speaker and language independent voice quality classification applied to unlabeled corpora of expressive speech. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13)*, IEEE, 7982–7986. DOI:http://dx.doi.org/ 10.1109/ICASSP.2013.6639219

John Kane, Christer Gobl, Stefan Scherer, and Louis-Philippe Morency. 2013b. A comparative study of glottal open quotient estimation techniques. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech'13)*. 1658–1662.

Klaus Krippendorff. 2012. *Content Analysis: An Introduction to Its Methodology* (3rd ed.). Sage, Beverly Hills, CA.

Michael LaCrosse. 1975. Nonverbal behavior and perceived counselor attractiveness and persuasiveness. *Journal of Counseling Psychology* 22, 6 (Nov. 1975), 563–566. DOI:http://dx.doi.org/10.1037/0022-0167.22.6.563

Shihong Lao and Masato Kawade. 2005. Vision-based face understanding technologies and their applications. In *Advances in Biometric Person Authentication*, Stan Li, Jianhuang Lai, Tieniu Tan, Guocan Feng, and Yunhong Wang (Eds.). Lecture Notes in Computer Science, Vol. 3338, Springer Berlin Heidelberg, 339–348. DOI:http://dx.doi.org/10.1007/978-3-540-30548-4_39

Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Barlett. 2011. The computer expression recognition toolbox (CERT). In *Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*. IEEE, 298–305. DOI:http://dx.doi.org/10.1109/FG.2011.5771414

James Maddux and Ronald Rogers. 1980. Effects of source expertness, physical attractiveness, and supporting arguments on persuasion: A case of brains over beauty. *J. Personal. Social Psych.* 39, 2 (Aug. 1980), 235–244. DOI:http://dx.doi.org/10.1037/0022-3514.39.2.235

George Mahl. 1956. Disturbances and silences in the patient's speech in psychotherapy. *J. Abnor. Social Psych.* 53, 1 (July1956), 1–15. DOI:http://dx.doi.org/10.1037/h0047552

Catha Maslow, Kathryn Yoselson, and Harvey London. 2011. Persuasiveness of confidence expressed via language and body language. *Brit. J. Social Clin. Psych.* 10, 3 (Sept. 1971), 234–240. DOI:http://dx.doi.org/10.1111/j.2044-8260.1971.tb00742.x

Winter Mason and Siddharth Suri. 2011. Conducting behavioral research on Amazon's mechanical turk. *Behav. Res. Meth.* 44, 1 (Mar. 2012), 1–23. DOI:http://dx.doi.org/10.3758/s13428-011-0124-6

Albert Mehrabian. 1971. *Silent messages*. Wadsworth Publishing Company, Inc., Belmont, CA.

Albert Mehrabian and Martin Williams. 1969. Nonverbal concomitants of perceived and intended persuasiveness. *J. Personality Social Psych.* 13, 1 (Sept. 1969), 37–58. DOI:http://dx.doi.org/10.1037/h0027993

Joan Meyers-Levy and Prashant Malaviya. 1999. Consumers' processing of persuasive advertisements: An integrative framework of persuasion theories. *Journal of Marketing* 63, Special Issue (1999), 45–60. DOI:http://dx.doi.org/10.2307/1252100

Norman Miller, Geoffrey Maruyama, Rex Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *J. Personal. Social Psych.* 34, 4 (Oct. 1976), 615–624. DOI:http://dx.doi.org/10.1037/0022-3514.34.4.615

Louis-Philippe Morency, Jacob Whitehill, and Javier Movellan. 2008. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG'08)*. IEEE, 1–8. DOI:http://dx.doi.org/10.1109/AFGR.2008.4813429

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

Daniel O'Keefe and Jakob Jensen. 2007. The relative persuasiveness of gain-framed loss-framed messages for encouraging disease prevention behaviors: A meta-analytic review. *J. Health Commun.: Int. Perspect.* 12, 7 (Oct. 2007), 623–644. DOI:http://dx.doi.org/10.1080/10810730701615198

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI'14)*. ACM, New York, 50–57. DOI: http://dx.doi.org/10.1145/2663204.2663260

W. Barnett Pearce and Bernard Brommel. 1972. Vocalic communication in persuasion. *Quart. J. Speech* 58, 3 (1972), 298–306. DOI: http://dx.doi.org/10.1080/00335637209383126

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

Richard Petty and John Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*.: Springer-Verlag, New York.

Jeffery Pittam. 1990. The relationship between perceived persuasiveness of nasality and source characteristics for Australian and American listeners. *J. Social Psych.* 130, 1 (1990), 81–87. DOI: http://dx.doi.org/10.1080/00224545.1990.9922937

Chanthika Pornpitakpan. 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. *J. Appl. Social Psych.* 34, 2 (Feb. 2004), 243–281. DOI: http://dx.doi.org/10.1111/j.1559-1816.2004.tb02547.x

Beatrice Rammstedt and Oliver John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *J. Res. Personal.* 41, 1 (Feb. 2007), 203–212. DOI: http://dx.doi.org/10.1016/j.jrp.2006.02.001

Howard Rosenfeld. 1966. Approval-seeking and approval-inducing functions of verbal and nonverbal responses in the dyad. *J. Personal. Social Psych.* 4, 6 (Dec. 1966), 597–605. DOI: http://dx.doi.org/10.1037/h0023996

Roni Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. IEEE* 88, 8 (Aug. 2000), 1270–1278. DOI: http://dx.doi.org/10.1109/5.880083

Stefan Scherer, John Kane, Christer Gobl, and Fiedhelm Schwenker. 2013. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Comput. Speech Lang.* 27, 1 (Jan. 2013), 263–287. DOI: http://dx.doi.org/10.1016/j.csl.2012.06.001

Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. 2011. The Interspeech 2011 speaker state challenge. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech'11)*. 3201–3204.

Han Suk Shim, Sunghyun Park, Moitreya Chatterjee, Stefan Scherer, Kenji Sagae, and Louis-Philippe Morency. 2015. Acoustic and para-verbal indicators of persuasiveness in social multimedia. In *Proceedings of the 40th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'15)*.

Steven Stern, John Mullennix, and Stephen Wilson. 2002. Effects of perceived disability on persuasiveness of computer-synthesized speech. *J. Appl. Psych.* 87, 2 (Apr. 2002), 411–417. DOI: http://dx.doi.org/10.1037/0021-9010.87.2.411

John Storck and Lee Sproull. 1995. Through a glass darkly: What do people learn in videoconferences? *Hum. Commun. Res.* 22, 2 (Dec. 1995), 197–219. DOI: http://dx.doi.org/10.1111/j.1468-2958.1995.tb00366.x

Jansen Voss. 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.

Ederyn Williams. 1977. Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin* 84, 5 (Sept. 1977), 963–976. DOI: http://dx.doi.org.libproxy1.usc.edu/10.1037/0033-2909.84.5.963

Stephen Worchel, Virginia Andreoli, and Joe Eason. 1975. Is the medium the message? A study of the effects of media, communicator, and message characteristics on attitude change. *J. Appl. Social Psych.* 5, 2 (Jun. 1975), 157–172. DOI: http://dx.doi.org/10.1111/j.1559-1816.1975.tb01305.x

Yiming Yang and Jan Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*. 412–420.

Joel Young, Craig Martell, Pranav Anand, Pedro Ortiz, and Henry Gilbert. 2011. A microtext corpus for persuasion detection in dialog. In *Proceedings of the AAAI-11 Workshop on Analyzing Microtext*. 80–85.

Received June 2015; revised August 2016; accepted August 2016