A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes

Maarten Brilman
University of Twente
Drienerlolaan 5, 7522 NB
Enschede, The Netherlands
m.f.brilman@alumnus.utwente.nl

Stefan Scherer
USC Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista, CA 90094, United States
scherer@ict.usc.edu

ABSTRACT

Interpersonal skills such as public speaking are essential assets for a large variety of professions and in everyday life. The ability to communicate in social environments often greatly influences a person's career development, can help resolve conflict, gain the upper hand in negotiations, or sway the public opinion. We focus our investigations on a special form of public speaking, namely public debates of socioeconomic issues that affect us all. In particular, we analyze performances of expert debaters recorded through the Intelligence Squared U.S. (IQ2US) organization. IQ2US collects high-quality audiovisual recordings of these debates and publishes them online free of charge. We extract audiovisual nonverbal behavior descriptors, including facial expressions, voice quality characteristics, and surface level linguistic characteristics. Within our experiments we investigate if it is possible to automatically predict if a debater or his/her team are going to sway the most votes after the debate using multimodal machine learning and fusion approaches. We identify unimodal nonverbal behaviors that characterize successful debaters and our investigations reveal that multimodal machine learning approaches can reliably predict which individual (∼75% accuracy) or team (85% accuracy) is going to win the most votes in the debate. We created a database consisting of over 30 debates with four speakers per debate suitable for public speaking skill analysis and plan to make this database publicly available for the research community.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Social and Behavioral Sciences**]

General Terms

Experimentation; Human Factors

MM'15, October 26–30, 2015, Brisbane, Australia.

ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806245.

Keywords

Machine Learning; Multimodal; Public Speaking; Nonverbal Behavior; Information Fusion

1. INTRODUCTION

The art of public speaking was practiced long before the Greeks wrote about it in their treatises more than 2500 years ago. For the Greek, it was a way of life, a way of being. The ancient Greeks were the first to systematize the art of public speaking, which they called *rhetoric*; the art of discourse. This art aimed to improve the capability of writers, lawyers, or politicians to inform, persuade, or motivate their audiences. An attest to the timeless role of public speaking in our culture are statistics accrued by online platforms such as Youtube[1], TED[2], or Intelligence Squared U.S.[3] providing access to historic and contemporary public speeches, debates, and arguments. In fact, public speeches to this day are an integral part of our lives and the ability to communicate in social environments often greatly influences a person's career development, can help build relationships, resolve conflict, gain the upper hand in negotiations, or sway the public opinion.

Proficient public speaking requires a different method of delivery than informal everyday conversations. While decisive arguments and a well-structured train of thought are important for a good public speaking performance, nonverbal behavior are just as, if not more, important for a speaker's success [27]. Relevant nonverbal behaviors include facial expressions, eye contact, posture, and gestures, as well as acoustic characteristics of the voice and speech, known as paralinguistic characteristics. These consist of voice quality, i.e. the coloring or timbre of the voice, as well as prosodic features, comprising pitch, rhythm, and intonation. An important aspect that is closely related to these behaviors is the display of emotion and affect. The importance of nonverbal behavior was shown in many domains that require proficiency in interpersonal skills, including healthcare, education, and negotiations where nonverbal communication was shown to be predictive of patient and user satisfaction [9], negotiation performance [29, 37] and proficiency in public speaking [38, 36, 35, 4, 7].

Within this work, we focus our investigations on a special form of public speaking, namely public debates of socioeconomic issues that affect us all. In particular, we analyze

[1] `http://youtube.com`
[2] `http://ted.com/talks`
[3] `http://intelligencesquaredus.org/`

performances of expert debaters recorded through the Intelligence Squared U.S. (IQ2US) organization. IQ2US invites subject matter experts on a regular basis to discuss issues of global and national impact in the Oxford style debate format, consisting of opening and closing statements, as well as a question and answer section. IQ2US collects high-quality audiovisual recordings of these debates and publishes them free of charge online. We mined this publicly available dataset and prepared it for automatic analysis. We extract audiovisual nonverbal behavior descriptors, including facial expressions, voice quality characteristics and a surface level linguistic analysis of sentiment and choice of words [40]. Within our experiments we investigate if it is possible to automatically predict if a debater or his/her team are going to sway the most votes after the debate - the Oxford style debating win condition - using automatic behavior analysis as well as multimodal machine learning and fusion approaches. Specifically, we identified three main research questions that we address within this work:

Q1: What nonverbal behaviors are indicative of successful debate performances and can we automatically and reliably extract such behaviors?

Q2: Which modality (audio, visual or the surface level linguistic analysis) is most indicative of successful debate performances and performs best in subject independent classification experiments?

Q3: Is it possible to reliably predict which individual is part of or what team is going to win the debate, i.e. sway the most votes, using multimodal information fusion?

The remainder of the paper is organized as follows. In Section 2 we provide an overview of related work. In Section 3 we go into detail on the creation of the dataset. In Section 4 we provide information about the feature extraction methods. In Section 5 we give a statistical evaluation of the features. In the following section (Section 6) we discuss our machine learning method. And in the remaining sections we provide the discussion and a conclusion.

2. RELATED WORK

Public speaking anxiety has received a lot of attention [32], but rather little work has been done on automatic assessment of public speaking skills. And to the best of our knowledge there is no work out there that attempts to automatically predict outcomes of debates similar to ours based on audiovisual nonverbal behavior descriptors, surface level linguistic analysis, and multimodal machine learning approaches. In the following, we summarize some of the related work found on automatic public speaking performance assessment and characterization.

Strangert and Gustafson [38] found in their study that the subjectively highest rated Swedish parliamentarians had a greater mean, standard deviation (SD) and range of the fundamental frequency (f_0) than the lower rated ones. This work is backed up by Hirschberg and Rosenberg [19], who observed a positive correlation between a greater mean and standard deviation of f_0 and charisma ratings for American politicians. Hincks [18] found that a greater f_0 standard deviation characterized speaker liveliness in her thesis studying computer aided pronunciation training.

Pfister and Robinson [31] developed a system that could classify between nine affective states. The same system was also used to detect six speech qualities in public speaking.

Scherer et al. [35] used the acoustic feature set in [38] along with measures of pause timings (i.e. average pause) and voice quality parameters. They identified that more tense voices are correlated with more persuasive and better political speeches overall. They furthermore analyzed motion energy on a global scale and found that it was positively correlated with the speaker's rating.

Weninger et al. [45] addressed the automatic assignment of 14 affective ratings to lectures from the TED talks website based on the relation between linguistic features and the ratings given by the audience on the website. They also performed a study to automatically determine different dimensions of leadership on a manually annotated YouTube corpus utilizing linguistic and acoustic features [44].

Chen et al. [7] used features for speech delivery, speech content (using transcripts) and non-verbal behavior to predict human holistic scores. They selected a subset of their features based on the Pearson correlation with the human rating scores on the entire data set.

Apart from trying to correlate subjective ratings with automatically extracted parameters, there has also been some effort put into creating systems that can automatically assess the speaker's skill and provide feedback. The results from our research can potentially be used to improve these type of systems.

In [24] a presentation coaching system was developed that detects the duration of utterances, the pitch (f_0) and the filled pauses. It furthermore uses a speech recognition engine to detect the speech rate. Face position and face orientation is tracked with the help of a marker. Both online and offline feedback is given to the user.

Cicero [4] is a platform that aims to train public speaking skills by providing automated feedback while speaking to a virtual audience. The system measures various vocal features such as pitch variety, volume and voice breathiness. It also tracks global arm and leg movement as well as gaze.

MACH (My Automated Conversation coacH) [20] is a system designed for social skill training. In particular the system was built for job interview training. It automatically reads facial expressions, speech, and prosody and responds via a virtual agent in real time. The researchers further developed their algorithms and developed an application using Google Glass to give online feedback to the user during a presentation based on their volume and speaking rate [39].

3. DATASET COLLECTION

We collected 36 debates published by the Intelligence Squared U.S. (IQ2US) organization from 2011 to 2014. We gathered video footage with audio of the debates, as well as manual transcripts provided by IQ2US. Voting results of each debate were also collected as well as the gender of the speakers. Debates were chosen solely on video quality, proper camera angles and audio quality. Twenty-three of the videos have a resolution of 720p (1280 by 720 pixels) and the remaining thirteen are 360p (640 by 360 pixels). All videos were extracted from the IntelligenceSquared Debates YouTube page[4]. Example motions include "Break up big banks" and "Genetically modify food". A full list of the debates is provided online[5].

[4] `https://www.youtube.com/user/IntelligenceSquared`
[5] `http://tinyurl.com/noof6vr`

3.1 Debate Structure

Debates are held in Oxford-style, a style derived from the Oxford Union society. In our case two teams of two are arguing the motion; two debating for and two against the motion. Both teams consist of professionals that have a significant amount of experience with public speaking and are leading experts in the field of the debated motion.

A strict structure is followed for these debates (Figure 2). Prior to any debate, the audience members cast their vote (for, against, or undecided) on the motion, which is shortly introduced by the organizer of the debate. Then each panelist takes turns in giving an uninterrupted opening statement. After the opening statements a question round takes place where the moderator presents the panelists with questions from both himself and the audience. In the final round each presenter gives a short closing statement. Finally the audience cast their post-debate vote. The winner is declared by looking at which team swayed the most audience members, based on the difference between the voting percentages before and after the debate. Speakers have a maximum of seven minutes for their opening statements and two minutes for their closing statement. For a few debates the maximum duration of the opening speech is six minutes. The question round varies in duration, but lasts around 45 minutes on average. Opening speeches were held standing up behind a lectern, while closing speeches were held sitting down behind a table (see Figure 1). The team debating for the motion always debated on the left for the audience and the other team on the right. The debates were held in large venues. Most debates took place in either the Merkin Concert Hall or Miller Theatre in New York, who seat 449 and 688 people respectively.

Figure 1: a) An overview of the stage. b) A typical frontal camera angle for the opening speech. c) A typical frontal camera angle for the closing speech. d) Example voting results.

3.2 Preprocessing

We decided not to include the question round in our work. During this question round speakers quickly change and interruptions take place. Given that this question round lasts a long time this would result in many short fragmented segments per speaker. However, we acknowledge it might be important to the outcome of the debates and plan to incorporate this round in our future investigations.

The opening and closing speech time frames have been annotated in ELAN [46][6] for all debates. These speeches have also been annotated on whether or not the camera had a frontal view of the speaker. Then, utilizing the slice functionality present in the SALEM toolbox [15], we extracted the time frames from the ELAN annotations to Matlab. As a next step we used the Matlab FFmpeg toolbox[7] to extract all the segments. The eight statement videos per debates where converted to 16 kHz mono wav audio files using FFmpeg[8]. From the transcripts we copied each statement to an individual text file. These transcriptions are clean versions, meaning that for example hesitations are not included. Transcripts are not time aligned with the audio files. The resulting files include for each speaker:

- Audio segments for the closing and opening statements.
- Full transcript for the closing and opening statements.
- All video segments with frontal data for the closing and opening statements.

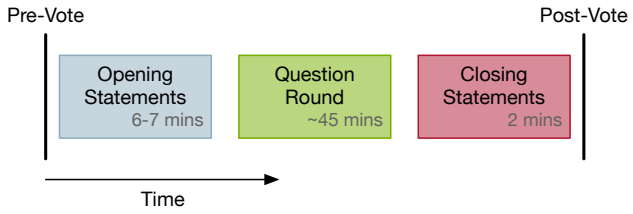

Figure 2: Oxford style debating overview figure. Prior to the start of the debate the audience votes on the motion. This motion is followed by opening statements by each speaker in turn. The opening statements are followed by a question round and the debate is closed by individual closing statements. After these statements a second vote is conducted. The team that swayed the most votes in their favor wins the debate.

3.3 Final Dataset

For our study we excluded six debates with the lowest voting difference (ranging from six to two) as these debates are close to being ties, i.e. it is unclear who the winner would be. This left thirty debates with a mean voting difference of 21.77 (SD=14.44) with a minimum difference of eight percentage points. We included all 120 speakers (19 female, 101 male) in our dataset from these debates, originating from various ethnicities and nationalities.

Ten females debated for a motion and nine against. A different group of ten females won the debate while nine lost. Furthermore the teams proposing the motion won fifteen times in total; the same amount as the teams opposing the motion. Four males partook in two debates, each time with a different partner. Two of these persons won both debates, one lost both and one won once and lost the other.

4. FEATURE EXTRACTION

4.1 Audio Features

Using COVAREP(v1.3.2), a freely available open source Matlab and Octave toolbox for speech analyses, we extracted several audio feature [8][9]. A large number of voice feature

[6] http://tla.mpi.nl/tools/tla-tools/elan

[7] http://www.mathworks.com/matlabcentral/fileexchange/42296-ffmpeg-toolbox

[8] https://www.ffmpeg.org

[9] http://covarep.github.io/covarep/

extraction methods have been implemented in this toolbox based on notable papers. We extracted the pitch, first and second formant as well as seven features that characterize voice qualities on a breathy to tense dimension. Breathy voice and tense voice are often considered to be on the opposite ends of the voice quality continuum [13].

- **Fundamental frequency (f_0)**: The fundamental frequency, or pitch, is tracked using the method suggested in [10] based on residual harmonics. This method simultaneously detects if a segment is voiced or unvoiced and is especially suitable for noisy conditions.
- **Formants (F_1 and F_2)**: [5] introduces the tracking of formants in detail. The first and second formants (F1 and F2) identify and characterize primarily vowels. These formants are the vocal tract frequencies that describe the first two spectral peaks with the lowest frequencies of the speech signal.
- **Normalized Amplitude Quotient (NAQ)**: The use of time-base parameters is one of the most common methods for glottal flow parameterization, which is used for voice quality estimation. Two commonly used parameters are the open quotient (OQ), the ratio between the open phase of the glottal pulse and the length of the fundamental period and the closing quotient (CQ), the ratio between the glottal closing phase and the length of this period.
  The NAQ [1] is a method to parametrize the glottal closing phase. It describes the ratio between the maximum of glottal flow and the minimum of its derivative, after being normalized by the fundamental frequency.
- **Quasi-Open Quotient (QOQ)**: The QOQ is related to the OQ and thus describes the relative open time of the glottis. It is measured by detecting the duration during which the glottal flow is 50% above the minimum flow. This is then normalized by the local glottal period [34].
- **H1-H2 ratio**: The H1-H2 parameter describes the amplitude of the fundamental frequency relative to that of the second harmonic. It is used as an indication of the open quotient [16].
- **Parabolic Spectral Parameter (PSP)**: The PSP is derived by fitting a parabolic function to the low-frequency part of the glottal flow spectrum. It provides a single value that describes how the spectral decay of an obtained glottal flow behaves with respect to a theoretical limit corresponding to maximal spectral decay [2].
- **Maxima Dispersion Quotient (MDQ)**: The MDQ is a parameter that quantifies the extent of the dispersion of the maxima derived from the wavelet decomposition of the glottal flow in the vicinity of the glottal closure instant (GCI) [22]. This dispersion is larger for a breathy voice than for a tense voice.
- **Peak Slope (PS)**: The PS involves decomposition of the speech signal into octave bands and then fitting a regression line to the maximum amplitudes at the different scales. The slope coefficient of this line is a measure for the voice breathiness [21].
- **Liljencrants-Fant model parameter Rd**: The final measure is one of the R-parameters of the Liljencrants-Fant model characterizing the glottal source [12]. The Rd captures most of the covariation of the LF model parameters. In [8] it was shown that the Rd confidence score also has discriminatory properties with respect to emotions and is therefore included in our work as a separate feature. We set the confidence threshold for the Rd parameter to 0.6. Rd values below this threshold are filtered out.

For all the above mentioned features we computed the mean and standard deviation over the entire opening or closing statement after removing the unvoiced data. Furthermore we computed the range of the f_0, F_1 and F_2 between the 25^{th} and 75^{th} percentile of these features. This gives a total of 25 audio features per segment.

4.2 Video Features

Paul Ekman showed in his studies that people can universally recognize the expressions of seven emotions. We extracted evidence for the presence of these seven basic emotions (joy, anger, surprise, fear, sadness, disgust and contempt) using Facet[10] from the frontal videos. Evidence for two more advanced emotions, confusion and frustration, was also extracted. Three overall sentiments were estimated as well; positive, neutral and negative. Due to the nature of the database baselining the application, feeding it neutral expressions per video, was not feasible. Therefore the results should be interpreted as facial expressions, rather than pure emotions. Nineteen features that describe elementary facial muscle actions, called action units (AU) in the Facial Action Coding System (FACS) [41], were computed as well. After combining the features from all the frontal video for each speaker we computed the mean and standard deviation over these features for a total of 46 features per statement.

Using Okao[11] we extracted nine more video features. These include horizontal and vertical face direction, face roll as well as horizontal and vertical gaze. The openness of the mouth and the level of smiling were also measured. We again computed the mean and standard deviation over these features. We ended up not using the mean of the horizontal gaze and face direction. This is due to the fact that those features are biased towards on which side of the stage the debater presented (see Figure 1a). This results in 16 remaining features.

4.3 Text Features

We applied a text analysis method called LIWC2007 [30] to extract features belonging to psychological and structural categories. LIWC is software that is developed to assess emotional, cognitive and structural text samples using a psychometrically validated internal dictionary [40, 42]. The software calculates the relative frequency to which a text sample belongs to a category. We used all 32 relevant categories, namely various pronouns, articles, as well as several psychological processes divided into social processes, affective processes, cognitive processes and perceptual processes[12] resulting in 32 features.

5. STATISTICAL EVALUATION

In order to investigate research question **Q1**, we conducted statistical analysis with the extracted audiovisual

[10] http://www.emotient.com/products/emotient-analytics/

[11] http://www.omron.com/r_d/coretech/vision/okao.html

[12] http://www.liwc.net/descriptiontable1.php

and surface level linguistic features. For each feature we computed a two tailed t-test as well as the effect size, using Hedges' g, over all segments. The g value denotes the estimated difference between the two population means in magnitudes of standard deviations [17]. Hedges' g is a commonly used standardized mean difference measure that can be transferred into other measures like Cohen's d [11]. In this section we summarize the features that differentiate the groups the most based on these tests. Due to the relatively small number of females we analyzed males and females together as one group, rendering the sample more heterogeneous. Given that the groups are balanced this shouldn't affect the results.

Table 1 summarizes the result of this statistical evaluation with respect to winning vs. losing debate performance characteristics and lists observed mean values, standard deviations, and effect sizes. Below we report major findings with respect to each modality separately.

Audio: The audio feature that distinguishes the winners and the losers the most is the pitch (f_0) range ($p<0.01$) and f_0 standard deviation ($p<0.01$). Furthermore the frequency of the second formant (F2) is higher for winners ($p<0.05$). The voice quality parameter that showed the biggest difference between the two groups is the QOQ with a near significant p-value of 0.0638. Given that winning speakers have an overall lower value than losing speakers, this indicates that voices of losers are more breathy. We also notice a larger standard deviation for H1-H2 and MDQ ($p<0.05$) for winners and a higher standard deviation of the confidence level for the Rd parameter ($p<0.05$) for losers.

Video: We found that the winning teams express less joy than the losing teams ($p<0.01$) this goes together with showing less overall positive expressions ($p<0.05$). Winners also show a larger standard deviation of evidence for disgust ($p<0.05$). Evidence for two action units, 18 and 20, showed significant differences between the two groups. Action unit 18 (*lip puckerer*) was detected more in the winning group, while action unit 20 (*lip stretcher*) was detected more in the losing group[13]. Action unit 20 its standard deviation also differs greatly between the two groups ($p<0.01$), being higher for the winners.

Utilizing Okao software we find that the standard deviation for the horizontal face movement is larger for the winners ($p<0.05$), while the deviation for the horizontal gaze movement is lower ($p<0.05$).

Text: Winners use more personal pronouns (e.g. *we* or *you*) than losers ($p<0.05$). Furthermore they use more words belonging to the discrepancy category such as *would*, *could* and *should* ($p<0.05$). Losers include more words from the perceptual category ($p<0.01$) and in particular related to the hearing (i.e. *listen*, *hearing*, $p<0.05$) category. Finally winners use somewhat more language involving social subjects ($p = 0.0742$) using words such as *mate*, *child*, and *story*.

When we compare the opening and the closing statements to each other we found a large number of differences between the two, which was expected due to their different nature (i.e. seated vs. standing, 2 mins vs. 6-7 mins, etc.). We opted to not include the results for this comparison within this work, as this is outside the scope of this paper. We, however, provide the results of the statistical tests between winners and losers on only the opening and closing speeches separately in Table 1. The observed values for both opening and closing statements separately indicate similar trends as shown with the Hedges' g values. In particular, positive g values represent a higher observation for the respective feature for the winners and lower values for losing debaters, vice versa for negative g values respectively.

[13]Examples of action units are found here: `http://www.cs.cmu.edu/~face/facs.htm`

6. EXPERIMENTS

In order to investigate research questions **Q2** and **Q3** we conduct speaker and debate team independent unimodal (**Q2**) and multimodal (**Q3**) machine learning experiments.

6.1 Method

Using all the features described in Section 4, we built a classifier to automatically determine the winning team of a debate. Given the large differences between the opening speech and closing speech features we decided to keep them separate in our work. Making use of the LIBSVM [6][14] package we trained eight separate support vector machine classifiers using second order polynomial kernels after testing Gaussian, linear, and several polynomial kernels.

We perform leave-one-debate-out testing where one debate with four speakers is kept for testing and the remaining 29 debates are used for training and validation. We conduct this validation approach in order to investigate the automatic classifier performance independent of the debate topic, debate teams, as well as individual debaters. All training data is scaled using min-max normalization. This same normalization is then applied to the test data.

In order to find the optimal regularization parameter (C-parameter) we perform 5-fold cross validation on the training set automatically searching in a range from 2^{-5} to 2^{20} with a step size of $2^{0.5}$. The C-value resulting in the highest performance is then chosen. This process is repeated 30 times, once for each debate excluded from the training set. We then take the median value of these 30 C-parameters as our one optimal C-parameter to train the classifier with. This final step essentially takes out outlying C-parameters that might have been the result of a random optimal 5 fold split, but aren't necessarily representative for the entire data set. This last step provides us with a more robust C-values for all cases.

Individual Debaters: In the next step eight models - one for each modality or data source and for opening and closing statements separately (see Figure 3) - are trained based on the leave-one-debate-out validation, i.e. each time with leaving out a different debate for classification. This provides us with eight labels per person that can then be used for decision fusion utilizing majority voting on an individual debater level. Given that we have an even number of inputs for the decision fusion classifier a tie might occur. In case a tie happens we take the best scoring initial classifier as having the final decision. This decision fusion step provides us with one label per person.

Debate Teams: For each debate, we then fuse the individual debater labels once more to come to a final conclusion on which team won the debate. As we have four speakers per debate a tie can also occur here. Such a tie occurs when all four persons in the debate are classified as either being winners, being losers, or if one person in each time is being

[14]`http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Table 1: Winning speakers versus losing speakers. Significance is denoted with * ($p < .05$) and ** ($p < .01$). The last two columns indicate the values when only including the opening or closing speeches.

Feature	Winner mean (SD)	Loser mean (SD)	Hedges' g	Hedges' g Opening	Hedges' g Closing
Audio Features					
f_0 Range	41.88 (15.10)	35.80 (13.71)	0.4203**	0.3893*	0.4696*
f_0 SD	32.45 (9.74)	29.02 (9.57)	0.3540**	0.2871	0.4374*
Rd Confidence SD	0.0595 (0.010)	0.0626 (0.009)	-0.3285*	-0.2606	-0.3879*
MDQ SD	0.0091 (0.0022)	0.0085 (0.0016)	0.3214*	0.2740	0.3641*
H1-H2 SD	2.786 (0.661)	2.620 (0.526)	0.2776*	0.2438	0.3121
Mean F2	1540.60 (61.60)	1523.16 (60.81)	0.2841*	0.3735*	0.2141
Mean QOQ	0.505 (0.062)	0.519 (0.058)	-0.2396	-0.2516	-0.2281
Video Features					
AU20 Evidence SD	0.455 (0.105)	0.415 (0.084)	0.4161**	0.4807**	0.3482
Mean AU20 Evidence	-0.1520 (0.311)	-0.030 (0.358)	-0.3628**	-0.4701*	-0.2683
Mean AU18 Evidence	-0.817 (0.684)	-1.045 (0.626)	0.3491**	0.4124*	0.3072
Mean Joy Evidence	-1.587 (0.720)	-1.333 (0.745)	-0.3460**	-0.4721*	-0.2388
Disgust Evidence SD	0.530 (0.168)	0.480 (0.142)	0.3244*	0.2987	0.3841*
Mean Positive Evidence	-0.177 (0.583)	0.018 (0.635)	-0.3191*	-0.4165*	-0.2326
Horizontal Face Direction SD	10.585 (3.055)	9.703 (3.458)	0.2695*	0.2539	0.3023
Horizontal Gaze Direction SD	17.990 (7.085)	20.423 (10.800)	-0.2655*	-0.3066	-0.2236
Text Features					
Perceptual Processes	1.273 (0.710)	1.596 (1.036)	-0.3628**	-0.2785	-0.4296*
Hear Category	0.550 (0.484)	0.732 (0.688)	-0.3055*	-0.2985	-0.3106
Discrepancy Category	2.006 (1.031)	1.751 (0.926)	0.2597*	0.2791	0.2636
Personal Pronouns	7.900 (2.810)	7.292 (2.276)	0.2551*	-0.0272	0.5325**
Social Category	10.318 (2.810)	9.693 (2.600)	0.2300	0.0944	0.3767*

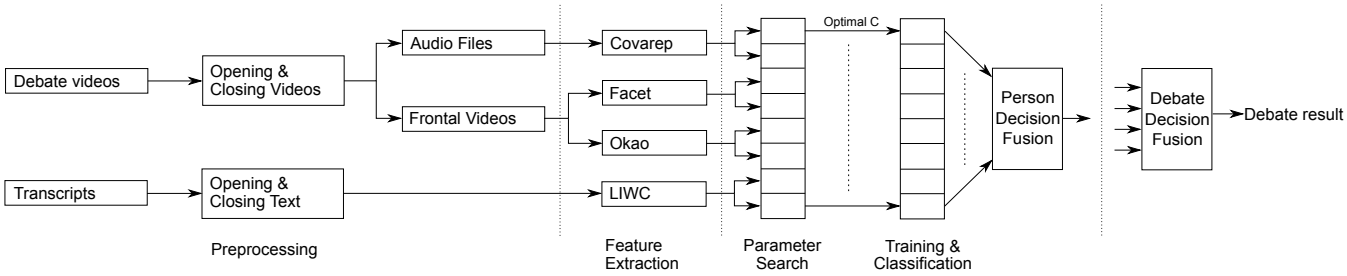

Figure 3: An overview of the entire system from preprocessing (Section 3), feature extraction (Section 4) to classification (Section 6). Opening and closing statements are individually processed, resulting in a total of eight initial classifiers.

classified as a winner. We do not further try to solve for these ties, but rather interpret them as being 50% correct as there are only two possible outcomes for each debate[15]. Alternatively, these ties could also be interpreted as reject cases, when no conclusive decision can be found by the classifier.

6.2 Results

The results of each unimodal classifier are presented in Table 2. The first column of numbers indicate the percentage of persons that were correctly classified by each individual support vector machine. The second column indicate the results from each individual classifier on a debate level. Overall, the support vector machine utilizing the acoustic domain features extracted using Covarep outperforms the other individual modalities for both the opening and closing statements (opening: 67.5% accuracy; closing: 65.0% accuracy). This result is considerably above chance level which is due to the setup of the dataset at 50% for all levels of investigation. We see that although a classifier might score the same as another on an individual level, it could score different on the debate level. This is due to the way the correctly classified persons are distributed over all the debates.

As a next step we combine multiple modalities to one label per person as discussed in the previous section. The number of correctly classified individuals for the different combinations is provided in Table 3. Both the fusion of audio and video as well as all three available modalities perform the best with 75.8% accuracy in both cases.

Table 3 further provides prediction accuracy on which team won the debate when fusing multimodal individual labels again. The debate team classification can either be correct, incorrect, or a tie might occur. In this case, the multimodal fusion over all modalities performs the best with 85.0% accuracy.

[15]This assumption statistically should also hold as a classifier could randomly choose the winner among these reject cases and still be correct about 50% of the time.

In total 22 debates are correctly classified, the classification of seven debates results in a tie and only one debate is misclassified. Figure 4 summarizes and visualizes the results of the multimodal fusion for both the individual debaters and the debate teams.

Table 2: The accuracy for each of the eight initial classifiers. The numbers in the brackets indicate (correct-tie-false) on the debate level.

Feature Group	Individual	Debate
Video		
Okao Opening	58.33%	61.7% (12-13-5)
Okao Closing	58.33%	61.7% (13-11-6)
Facet Opening	58.33%	66.7% (14-12-4)
Facet Closing	58.33%	60.0% (12-12-6)
Audio		
Covarep Opening	**67.50%**	71.7% (18-7-5)
Covarep Closing	65.00%	**73.3%** (15-14-1)
Text		
LIWC Opening	55.00%	53.3% (11-10-9)
LIWC Closing	60.00%	70.0% (19-4-7)

Table 3: The accuracies for individual speaker decision fusion and full debate decision fusion. The numbers in the brackets indicate (correct-tie-false).

Modalities	Individual	Debate
Video + Text	60.0%	66.7% (14-12-4)
Audio + Text	72.5%	76.7% (20-6-4)
Audio + Video	**75.8%**	83.3% (22-6-2)
Audio + Video + Text	**75.8%**	**85.0%** (22-7-1)

7. DISCUSSION

In this section, we discuss our results with respect to the research questions previously introduced:

7.1 Q1 - Nonverbal Indicators

Our first research question aims at identifying behavior indicators from audio, video and the surface level linguistic analysis. Our statistical evaluations revealed several interesting findings that we discuss in the following.

Audio: Based on the acoustic analysis we found that increased fundamental frequency (f_0) both measured as range as well as standard deviation is indicative of a successful debate performance. Intuitively, both measures indicate that a speaker with an increased expressivity is more successful than a speaker that sounds more monotonous. In fact, expressive speakers have been found to be more engaging and better overall in related work [38, 19, 35]. In addition, we could identify that speakers with less breathy voice quality are more likely to be in a winning team. This finding is confirmed in prior work that investigated political speakers in the German parliament [35]. The researchers found that speakers with tenser voice qualities were rated better overall and more persuasive and less insecure than those with more breathy voice qualities. While we already found a good number of interesting features with respect to the acoustic characteristics of successful debate performances, we believe there is plenty of room for improvement. For example we have not at all investigated timing based features (e.g. pause timings) or intensity features within our

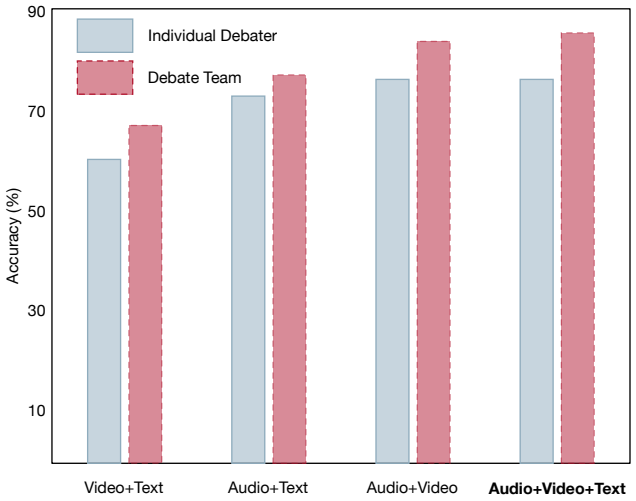

Figure 4: Results of the multimodal fusion in accuracy in % for both individual debaters as well as entire debate teams. Overall the multimodal fusion of all available modalities shows the most promising results and outperforms the other subsets of modalities.

work that have been found to be of relevance in the past [35]. We plan to incorporate such indicators in forthcoming investigations.

Video: With respect to video based behavioral descriptors we found that a decreased display of joy and less positive emotional displays overall are indicative of a successful debate performance. While this might sound counterintuitive, it is possible that a debater that performs more seriously during the opening and closing statements is regarded as more professional. A serious speaker might be regarded as having more powerful arguments than someone that appears to mask insecurity in his/her argument through the display of joy [3, 33].

We further found some interesting behavior indicators that are related to gaze patterns and debate performances. It appears that individuals that win their debates shift their entire face when addressing the audience during their opening and closing statements, while individuals on the losing teams shift their gaze more with their eye movement rather than gross head movements. This finding is in line with other work that has identified gestures and overall coarse body motion has a large impact on public speaking performances and their effect on audiences. Nguyen et al. [28] for example, state that emotion expression through body language is the most important cue to asses a speaker and built a system around this fact using the skeleton data from the Microsoft Kinect. Gross movement of the body was also identified to be related to proficient public speaking in political debates [23, 35]. In the future, we plan to incorporate behavioral features in our investigations as well. In particular, we aim to analyze gestures jointly with acoustic features. For example, visual beat gestures to emphasize a point should be temporally coordinated with acoustic emphasis in order to maximize the conveyed effect. We seek to identify behavioral factors that reveal synergies across modalities to explain qualitatively observed behavioral con-

cepts, such as increased anxiety, reduced expressivity, or lack of competence.

Text: With our limited surface level linguistic analysis using the software LIWC [40], we could only identify a few indicators of successful performances. For example increased use of words such as *should*, *would* and *could* is associated with a higher chance of success in the debate. However, it is difficult to assess the relevance of such findings given the small statistical difference between winners and losers. Our future analysis will aim to unravel more complex features such as argument structure or thought processes. In particular, the identification of the use of metaphors, examples, arguments, or facts could be of use. We plan to investigate novel natural language processing algorithms that have been successfully employed in a wide range of applications, such as document vectorization approaches to identify such patterns in language [26, 25].

7.2 Q2 - Unimodal Debate Classification

Our second research question is aimed at finding out which modality (audio, visual, or surface linguistic features) is most indicative of successful debate performances based on classification experiments.

In order to do so we trained eight separate support vector machines (one for each data source). We trained the opening and closing statements separately as our analysis indicated that there were significant differences between the two.

Audio: We found that audio was the best modality at differentiating winning speakers from those that lose. With the audio features we achieved an accuracy for individual debaters of 65.0% for the closing statement and 67.5% for the opening statement respectively. This result shows an accuracy, which is considerably above chance level (50%). On a debate level, the accuracy increases slightly to above 70% accuracy for both opening and closing statements. The opening statements appear to yield slightly improved results for the individual debater classification; for debate level classification this trend is reversed.

Given the promising results from these audio features we plan to investigate them further using more advanced machine learning methods. We plan to investigate sequential learning techniques such as recurrent neural networks. In particular the recurrent neural network toolkit that supports processing on GPU's called CURRENNT, allowing for much faster processing times, is of interest to us [43]. Recurrent neural networks have shown promising results in the modeling of speech and human behavior in general [14].

Video: From the video modality we extract two separate groups of features. While both feature sets originate from the same modality, i.e. video, they are qualitatively quite different, as one focuses on emotions, while the other focuses on head and gaze movement (cf. Section 4).

Our two types of video features both achieve an accuracy of 58.33% over both the opening and closing statements for the individual debater classification. We attempted to fuse both feature sets early (i.e. combining the features before classification), this however led to a lower performance than the initial 58.33%. This might indicate that both feature sets are able to classify a different set of speakers correctly, which holds a lot of potential for multimodal fusion techniques and error correcting algorithms (cf. Research question **Q3**).

On a debate level the classification results improve slightly, with the best performance observed for the emotion relevant features (i.e. Facet features) in the opening statement. This can be interpreted in a way that emotional display is in particular important for the opening statements of the debate rather than the closing statements.

Text: Utilizing the surface level linguistic analysis provided by the software LIWC [40], we find that in particular closing statements appear to be important to distinguish winning from losing debaters as well as teams respectively. For the opening statements, we barely reach accuracies above chance level, which indicates that these surface features (i.e. broad word categories) are not specific enough or do not capture important aspects of the performances. In fact, the opening statements are about three times as long as the closing statements and hence comprise a lot more data, which intuitively should result in better classification results. This, however, is not the case. We believe that the classification approach utilizing the debate transcripts has the largest potential for improvement in the future.

7.3 Q3 - Multimodal Debate Classification

Our third research question aimed to answer the question if we could improve classification results by combining modalities. In order to do so we applied decision (or late) fusion on the eight initial classifiers to obtain one label for each individual speaker. We then fused these labels once more to obtain a result on a debate level.

We found that using all three modalities (audio, video, and text) showed the most promising results with an accuracy of 75.8% on individuals and 85.0% on debates. Using only video and audio information also resulted in an accuracy of 75.8% on individuals, but gave slightly worse results on debates (83.33%; cf. Table 3 and Figure 4). This indicates that the text features do not add a lot of information to the fusion. Now while it is certainly possible to extract different text features, as explained in the discussion for **Q1**, our results in fact also indicate that it is possible to reliably determine the winners of debates without having access to manual transcripts of the debates, but merely based on nonverbal behavior. However, as argued earlier we consider linguistic information, such as argument structure, a very important source of information to determine successful debaters. In the present work, we only utilize surface level linguistic features.

The multimodal fusion further underlines the discussed importance of audio features in our investigations (cf. Discussion for question **Q2**), as the accuracy drops to only 60.0% on individuals and 66.7% on debates when using only surface linguistic and visual features. This is below that of using an unimodal approach based on the audio features.

When investigating the debates that led to wrong or undecided results we find the following: For the debates that aren't correctly classified we find that their average voting difference in percent of votes (mean 14.25) is well below that of the overall database (mean 21.77). With the one debate that is being misclassified, "Break up big banks", having the lowest voting difference in the dataset being eight percentage points. For the seven debates that are being classified as ties or undecided we found that one has a difference of 9 percentage points and three of them a difference of 10. This in fact indicates that in particular the close debates are difficult to classify. This can be explained intuitively as the speakers' performances in these debates in fact might be quite similar and on par across teams. Therefore the diffi-

culty of the classification is increased. In addition, it might be argued that in the closer debates one speaker is carrying their team, while his/her partner might be considered a poorer speaker and does not add to the team. As the debates are evaluated in teams both speakers are always given the same label. We plan to investigate individual speaker performances using post-hoc annotations per speaker at a later stage.

8. CONCLUSION

In this paper we investigated public speaking in the context of team based debates on a large variety of socioeconomic issues. The debates follow the standard Oxford style debating model in which the winner is decided by the percentage amount of swayed votes between a pre-debate and a post-debate vote. This paper aims to provide four additions to the state of the art on public speaking research: (1) we conduct a thorough analysis of nonverbal behavioral indicators of successful debate performances. Our findings confirm those of related work and extend the pool of features investigated considerably. (2) We identified that within our analysis the acoustic modality might have the strongest discriminative faculty and resulted in the highest observed accuracies for single modalities. We, however, acknowledge that there is considerable room for improvement in our work especially in the visual modality (e.g. behavioral information, gestures) and the linguistic analysis (e.g. argument structure, use of facts, blame, etc.). (3) A multimodal fusion approach was found to reliably predict winners of debates automatically both for individual debaters as well as debate teams with accuracies of around 75% for individuals and 85% for teams respectively. (4) In addition to the conducted investigations, we collected a novel multimodal database that we plan to make publicly available to help further research on public speaking assessment and evaluation. The database is based on a very active online platform named Intelligence Squared U.S. and the organization is publishing a novel debate online every few weeks, which renders the proposed database extensible and ever more challenging in the future.

Acknowledgements

We would like to thank Intelligence Squared U.S. for the immensely interesting debates and the excellent resource that they provide online. This material is based upon work supported by the National Science Foundation under Grants No. IIS-1421330 and U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government, and no official endorsement should be inferred.

9. REFERENCES

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

[2] P. Alku, H. Strik, and E. Vilkman. Parabolic spectral parameter–A new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

[3] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110(3):1581–1597, 2001.

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

[5] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Improved differential phase spectrum processing for formant tracking. pages 2421–2424, 2004.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee. Towards automated assessment of public speaking skills using multimodal cues. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 200–203. ACM, 2014.

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

[9] M. R. DiMatteo, R. D. Hays, and L. M. Prince. Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. *Health Psychology*, 5(6):581, 1986.

[10] T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976. ISCA, 2011.

[11] J. A. Durlak. How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, 34(9):917–928, 2009.

[12] G. Fant, J. Liljencrants, and Q. Lin. A four parameter model of glottal flow. *KTH, Speech Transmission Laboratory, Quarterly Report*, 4:1–13, 1985.

[13] C. Gobl and A. N. Chasaide. Acoustic characteristics of voice quality. *Speech Communication*, 11(4):481–490, 1992.

[14] A. Graves. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.

[15] M. Hanheide, M. Lohse, and A. Dierker. SALEM-statistical analysis of ELAN files in Matlab. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pages 121–123, 2010.

[16] H. M. Hanson and E. S. Chuang. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, 106(2):1064–1077, 1999.

[17] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981.

[18] R. Hincks. *Computer support for learners of spoken English*. PhD thesis, Royal Institute of Technology, 2005.

[19] J. B. Hirschberg and A. Rosenberg. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Eurospeech 2005*, pages 513–516, 2005.

[20] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. MACH: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706. ACM, 2013.

[21] J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA, 2011.

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

[23] M. Koppensteiner and K. Grammer. Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, 44(3):374–379, 2010.

[24] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 358–365. ACM, 2007.

[25] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1188–1196, 2014.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, 2013.

[27] A.-T. Nguyen, W. Chen, and M. Rauterberg. Feedback system for presenters detects nonverbal expressions. *SPIE Newsroom*, 2012.

[28] A.-T. Nguyen, W. Chen, and M. Rauterberg. Online feedback system for public speakers. In *IEEE Symp. e-Learning, e-Management and e-Services*, 2012.

[29] S. Park, P. Shoemark, and L.-P. Morency. Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization. In *Proceedings of the 18th International Conference on Intelligent User Interfaces (IUI '14)*, pages 37–46. ACM, 2014.

[30] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of LIWC2007, 2007.

[31] T. Pfister and P. Robinson. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2):66–78, 2011.

[32] C. B. Pull. Current status of knowledge on public-speaking anxiety. *Current opinion in psychiatry*, 25(1):32–38, 2012.

[33] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm. Spotting laughter in natural multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems: Special Issue on Affective Interaction in Natural Environments*, 2(1):4:1–4:31, 2012.

[34] S. Scherer, Z. Hammal, Y. Yang, L.-P. Morency, and J. F. Cohn. Dyadic behavior analysis in depression severity assessment interviews. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 112–119. ACM, 2014.

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

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

[37] H. S. Shim, S. Park, M. Chatterjee, S. Scherer, K. Sagae, and L.-P. Morency. Acoustic and paraverbal indicators of persuasiveness in social multimedia. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[38] E. Strangert and J. Gustafson. What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations. In *Proceedings of Interspeech 2008*, pages 1688–1691. ISCA, 2008.

[39] M. I. Tanveer, E. Lin, and M. E. Hoque. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 286–295. ACM, 2015.

[40] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[41] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[42] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*, volume 10, pages 178–185, 2010.

[43] F. Weninger, J. Bergmann, and B. Schuller. Introducing CURRENNT–the Munich open-source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research*, 16:547–551, 2015.

[44] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller. The voice of leadership: Models and performances of automatic analysis in online speeches. *IEEE Transactions on Affective Computing*, 3(4):496–508, 2012.

[45] F. Weninger, P. Staudt, and B. Schuller. Words that fascinate the listener: Predicting affective ratings of on-line lectures. *International Journal of Distance Education Technologies (IJDET)*, 11(2):110–123, 2013.

[46] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: a professional framework for multimodality research. In *International Conference on Language Resources and Evaluation*, pages 1556–1559, 2006.