UNIVERSITY OF TWENTE.

A Multimodal Predictive Model of Successful Debaters

or How I learned to Sway Votes

Master's Thesis in Human Media Interaction

MAARTEN BRILMAN

Faculty of Electrical Engineering, Mathematics and Computer Science
UNIVERSITY OF TWENTE
Enschede, The Netherlands 2015

Master's Thesis 2015

A Multimodal Predictive Model of Successful Debaters

or How I learned to Sway Votes

MAARTEN BRILMAN
m.f.brilman@alumnus.utwente.nl

Examination committee:
Prof. dr. D.K.J. Heylen, University of Twente
Dr. ir. H.J.A. op den Akker, University of Twente
Dr. rer. nat. S. Scherer, USC Institute for Creative Technologies

USC Institute for Creative Technologies

University of Southern California Institute for Creative Technologies
Computer Science Department
Playa Vista, CA, The United States of America

UNIVERSITY OF TWENTE.

Human Media Interaction
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
Enschede, The Netherlands

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.

Keywords: Machine Learning; Multimodal; Public Speaking; Nonverbal Behavior; Information Fusion.

Contents

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 [39]. 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 [11], negotiation performance [41, 51] and proficiency in public speaking [53, 50, 49, 5, 9].

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 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

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

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 [55]. 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?

1.1 Previous Work

Parts of this thesis have been adapted from the paper coauthored by Stefan Scherer, which has been accepted to appear in the ACM Multimedia 2015 conference proceedings:
M. Brilman and S. Scherer. A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes. *To appear in the Proceedings of the ACM International Conference on Multimedia, ACM, 2015.*

1.2 Thesis Outline

The remainder of the thesis is organized as follows. In Chapter 2 an overview of related work is given, as well as some general background information on public speaking. In Chapter 3 the dataset that was created for this study is extensively discussed. In Chapter 4 information about the feature extraction methods that were used is given. A statistical evaluation of the aforementioned features is provided in Chapter 5. In the following chapter (Chapter 6) the machines learning methods are discussed. And in the remaining chapters the discussion, including future work, as well as a conclusion are given.

2
Background

2.1 What defines a good public speaker?

In this section some general outlines will be given that are characteristic of a good public speaker.
One should keep in mind with general outlines that a performance should be adjusted to the context and content of the presentation. A speech for president election generally requires more exaggerated movements and a stronger voice than a presentation in a classroom. And speaking to a large group with 400 people requires a different approach than presenting the same content to a group of ten people.
The first characteristics that will be discussed relate to the voice.

2.1.1 Voice

Stephen Lucas describes a number of aspects of the voice that should be controlled to become a good public speaker; volume, pitch, rate, vocal variety and pauses [37].
First off it is important to speak at the right *volume*, even if todays microphones can be used to amplify one's voice. This means that it should not be too loud, which could come across as rude. But also not too softly, as people will not understand it. Volume can also be a powerful tool to emphasize a point.
A second aspect is the *pitch* of one's voice. Pitch is the highness or lowness of a speaker's voice. Changes in pitch are known as inflections that are used to convey meaning and emotion. People who do not use inflections are said to speak in a monotone. And although few people speak in an absolute monotone, many do use repetitious pitch patterns that should be avoided. One way to avoid this is to prevent each sentence from ending with the same inflection, so not to always increase or decrease the pitch, but rather to vary it.
Another aspect is the *speech rate*. People in the United States usually speak at a rate between 120 and 150 words per minute, however there is no uniform rate for effective speechmaking. And while Franklin Roosevelt spoke at 110 words per minute, John Kennedy used 180. Martin Luther King even opened his "I have a dream" speech with 92 words per minute, but finished it with 145. The best rate of speech therefore depends on several factors, such as the mood that should be created, the current subject of the speech and also the voice of speaker itself. This means that the speech rate should often even vary during a speech. The key is that the rate should not be too slow, but also not too fast depending on the situation. The latter is a problem that many public speakers have when they are nervous.
Pauses can be a useful when used properly in a speech. For example using a pause

at the end of an idea can give the the audience time to let it sink in. However pauses in the middle of the sentences are generally bad and even worse are vocalized pauses, or pause fillers, such as “uh” and “uhm". These pause fillers create negative perceptions about the speaker's intelligence. And excessive fillers can even distract the listener so much that they could lead to the message being lost entirely [52].
Varying the speech rate, pitch and volume contributes to a lively and expressive voice that keeps the public interested.

2.1.2 Body

The second important group of nonverbal behaviors for public speaking is related to the speaker's body. The five main aspects here are movement, posture, gestures, facial expressions and eye contact.
Movement is related to how the speaker moves on the podium. Being nervous manifests itself into distracting mannerisms that should be avoided [57]. One has to take care that one does not franticly paces across the podium as this comes across as being on edge. Although the same goes for standing completely still as if being a statue. Continuously shifting one's weight from one foot to the other has a negative influence on the performance as well. But movement can also be used in a positive way as long as moving is done with a reason. It can support and reinforce an idea and movement will almost always attract an audience's attention. Other behaviors that plague ineffective speakers include leaning on the lectern, fidgeting with notes or adjusting hair and clothing.
Posture is related to how someone positions his or her body. Posture reflects someone's attitude and whether they are confident when speaking. Good posture also helps with breathing properly and speaking loudly and clearly. Furthermore it provides a good starting point for gesturing. Good posture involves standing up straight and balancing the body weight properly on the balls of the feet. As with some of the voice characteristics it is all about balancing certain aspects. The posture should be relaxed, but not sloppy and alert, but not stiff. An aspect that is both related to posture and gestures is the position of the hands while in their rest position. A few generally bad positions are listed below [6], these are shown in Figure 2.1:

- Hands behind the back. This looks very formal and furthermore the hands have to first be brought forward before one can start gesturing. Making gesturing somewhat awkward.
- Hand in the pockets. This looks very casual and although it could be a good position when speaking to colleagues or friends, it can also come across as if someone does not care or even as cocky.
- Hands on the hip. This also looks very arrogant in most situations.
- Folding the arms or crossing the arms across the body. These are very defensive positions that makes it appear as if one is closed off.
- Keeping the hands folded all the way down in front of the body. Gesturing from this position is also hard as the hands have to travel a great distance before each gesture.

Figure 2.1: Some common bad poses.

Simply letting the arms hang naturally at the sides is a position that works well if this feels natural. Another position that works well is somewhat resting one hand in the other just above the belt area, as shown in Figure 2.2.

Figure 2.2: Some generally good poses.

Gestures are used in a different way by each person. Some great speakers gesture a lot while others gesture hardly at all. The important part is that the gestures should appear natural and spontaneous. In order to achieve this it is helpful to not suppress one's natural impulse to strengthen words with gestures, which is something everyone has. A number of other tips include:

- Gestures should match the verbal message as not doing so can come across as artificial.
- Proper timing is important, this means that the stroke (the part of the gesture that carries the actual meaning) should be aligned with the matching word most of the time.

- Gestures should also be lively and distinct and should be performed wholeheartedly as otherwise it suggests that the speaker lacks conviction. This also requires a gesture to be performed smoothly and at the right speed.
- Following a set pattern is detriment to proper gesturing.

The Toastmasters group, among others, defines four different types of gestures as being part of presenting [57]:

- *Descriptive* gestures are used to clarify a verbal message, such as visualizing the size of a certain object.
- *Emphatic* gestures try to communicate emotions, such as using a slumped body language to convey sadness.
- *Suggestive* gestures are used to convey a certain mood. An example would be to cross the arms to indicate that someone was very closed off.
- *Prompting* gestures are gestures that prompt the audience to do something, such as clapping or raising their hands.

Facial expressions communicate emotions and feelings while speaking. People are capable of recognizing these feelings (e.g. anger or joy) by just looking at the expressions [15]. Nervousness might exhibit itself into random facial expressions that should be avoided such as licking or biting the lips, raising the corners of the mouth or even in making twitching movements in parts of the face. Smiling at appropriate times is key to conveying friendliness, but other than that there are no real rules that define good facial expressions, as they rely heavily on the content of the talk.
Eye contact is not just a good tool to monitor the audience's response; speakers who fail to establish eye contact can also be perceived as less credible and less competent [32]. Therefore it is important to make effective eye contact, rather than just passing your gaze throughout the room. Although with large groups this might not always be possible it is still important to look at the entire audience. Furthermore it is good to remember to not follow a repeating pattern when gazing at an audience, such as constantly moving the head from the left to the right.

2.2 Related Work

Public speaking anxiety has received a lot of attention [44], 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, some of the related work found on automatic public speaking performance assessment and characterization is summarized.
Eva Strangert has studied rating public speaking skill based on vocals. In one of her works she had Swedish parliamentarians rated based on their public speaking abilities [53]. She found that the highest rated parliamentarians had a greater mean, standard deviation (SD) and range of the fundamental frequency (f_0) than the lower rated ones. Furthermore she found that disfluencies were a cause for negative ratings. Her work is backed up by Hirschberg and Rosenberg [24], who observed a positive correlation between a greater mean and standard deviation of f_0 and charisma ratings for American politicians. A faster speaking rate also correlated with increased charisma ratings. Hincks [23] found that a greater f_0 standard deviation characterized speaker liveliness in her thesis studying computer aided pronunciation training. By making use of this metric, that she termed the pitch variation quotient, she could help combat monotonous delivery.
Pfister and Robinson [43] developed a system that could classify between nine affective states (absorbed, excited, interested, joyful, opposed, stressed, sure, thinking and unsure). The same system was also used to detect six positive speech qualities (clear, competent, credible, dynamic, persuasive and pleasant). Rather than selecting one label, as with emotion detection, all classes were detected and were labeled with a probability for each sample. The training segments were labeled by an experienced speech coach. An average accuracy of 81% was achieved by comparing the automatic labels to labels given by an expert.
[49] used the acoustic feature set in [53] along with measures of pause timings (i.e. average pause) and voice quality parameters. Two of the new parameters introduced were the normalized amplitude quotient (NAQ) and the PeakSlope that both correlated strongly with the overall assessment of the speakers. The PeakSlope and NAQ are parameters that identify breathy regions of the speech (or voice tenseness) and thus measure voice quality. Pause time also correlated negatively with the speaker's perceptual rating. They furthermore analyzed motion energy on a global scale and found that it was positively correlated with the speaker's rating.
[62] address the automatic assignment of 14 affective ratings to online lectures from the TED talks website. This is based on the relation between linguistic features and the ratings given by the audience on the website. The features come from set of word and bag of word models. A set of words model deals with binary features, the presence or absence of a word, whereas in a bag of word model features relate to word counts. Naive Bayes and support vector machine approaches were used to train different classification models.
Weninger et al. [61] performed a study to determine different dimensions of leadership on a manually annotated YouTube corpus containing 409 minute-long speeches.

For the acoustic analysis a large set consisting of 1582 features was used. Example features include loudness, fundamental frequency, spectral features and voice quality features. A bag of words model was employed for the linguistic features.
Chen et al. [9] 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 [34] 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 [5] 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) [25] 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. It uses two types of feedback: summary feedback and focused feedback. For the summary feedback the following (voice) features were displayed: smiles, total pause duration, speaking rate, pitch variation and weak language. These are tracked over multiple sessions, each session shown in a different color. Weak language is implemented as a fixed list of (filler) words, such as "like", "basically", "umm" and "totally". For the focused feedback the recorded video is shown to the user, while showing the intensity of smiles, number of head nods and shakes, the spoken words with weak and strong language marked, loudness, emphasis and pauses. These 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 [54].

2.3 Multimodal Learning

While not a lot of work has been done yet on multimodal recognition of public speaking skills, there has been a large interest in multimodal detection of emotions. This field is generally known as affective computing. As both fields share similarities in this section an overview will be given of some of those techniques. First an introduction into multimodal fusion will be given.

2.3.1 Fusion Levels

In order to classify information from different sources several fusion levels have been proposed. The most widely used approaches are: early, hybrid and late fusion [3]. These approaches are illustrated in figure 2.3.
For early or feature level fusion features are extracted from the input signals and directly combined into one feature vector, after which an analysis unit performs the classification. This means that the features are combined at an early stage. Therefore early fusion can be used for signals that are highly dependent such as speech and lip movement [13].
With late or decision level fusion each modality is first classified using individual classifiers and afterwards these decisions are analyzed to reach a final decision. For this strategy features can have entirely different representations, unlike with early fusion, because the decisions generally end up having the same representation. Decisions level fusion is also much easier to upscale with extra features than early level fusion. This approach is for example suitable for fusion of speech and gestures as these modalities are less coupled.
Researchers have used hybrid fusion [28] as well. Here both feature and decision level fusion is applied to the data and afterwards fusion is applied again over both results.
A fourth method that has recently gotten more attention is mid-level fusion [45], where features are first combined into different mid-level concepts before classification [29, 46].

2.3.2 Fusion Methods

In [3] common methods for multimodal fusion are extensively discussed. Three categories are defined: estimation-based, rule-based and classification-based. Estimation-based methods are used to better estimate certain parameters (i.e. velocity) for applications like object tracking. These methods are thus not relevant for the current project. The other two categories deal with obtaining a decision based on observations.

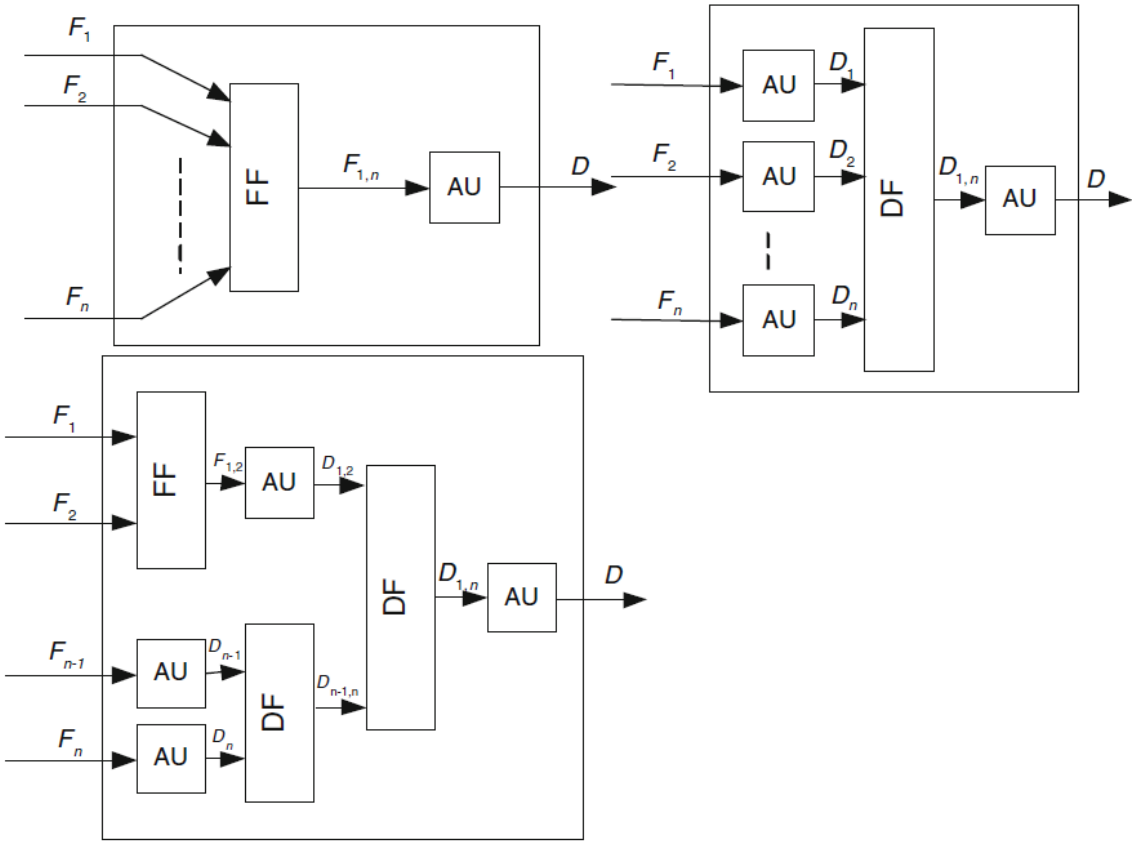


Figure 2.3: Early fusion (upper left), late fusion (upper right) and hybrid fusion (bottom). AU stand for analysis unit, DF for decision fusion and FF for feature fusion (adapted from [3]).

2.3.2.1 Rule-based Fusion

Rule-based fusion uses various rules to combine multimodal information. Linear weighted fusion is one of the simplest methods. Here feature vectors or decisions are combined (after normalization) using sum or product operators. Majority voting is a special case of weighted fusion with all weights being equal. The final decision is thus based on the one that the majority of the classifiers decide on. Custom-defined rules can also be used. These rules are context specific such as conditional (i.e. if-then) rules.

2.3.2.2 Classification-based Fusion

Classification-based fusion methods include techniques that are used to classify (multimodal) observations into one of the predefined classes. A number of these techniques are:

- *Support vector machine* is a supervised learning method and is used as an classifier where a set of input vectors is partitioned into one of two classes by a hyperplane. Various extensions exist to allow for more than two classes. For multimodal fusion this method can be used to solve a pattern problem with as input scores given by individual classifiers.
- For *Bayesian inference* the multimodal information is combined via rules of probability theory. Observations of multiple modalities are combined and a joint probability of a certain observation is inferred. The same way decisions from multiple classifiers can be combined, so this method can be applied to both feature and decision level fusion.
- *Hidden Markov Models* are the most popular form of a dynamic Bayesian network. These are represented by a graph in which nodes represent observations or states. Nodes depend on each other via certain probabilities indicated by the edges. These systems allow multiple dependencies between nodes and are very suitable for decisions involving time series data.
- A *neural network* consists of a network with three main types of nodes: input, hidden and output. Sensor data or decisions (based on the data) is fed to the input nodes. The fusion of the observations or decisions is given at the output nodes. Hidden nodes are neither input or output nodes that are part of the network architecture. The weights along the paths between the nodes along with the architecture determine the behavior of the network.

3
Dataset Collection

3.1 Intelligence Squared US

Intelligence Squared U.S. (IQ2US) is based on the Intelligence Squared debate program in London. It has presented over a 100 debates on a wide range of contemporary topics, ranging from clean energy and the financial crisis to the Middle East and the death of mainstream media. Invited speakers are well-known authorities on the discussed topics. The organization records all their debates and puts the videos on their website[1] along with information about the voting results, speakers and additional research on the topic. A transcript of the entire debate, prepared by National Capitol Contracting, is also made available. New debates are added every month.

3.2 Debate Collection

I manually collected video footage of 36 debates published by the Intelligence Squared U.S. organization from 2011 to 2014. Debates were chosen solely on the quality of the recordings, with the main three criteria being: video quality, proper camera angles and audio quality. The transcripts provided by IQ2US were also downloaded. Debates from before 2011 weren't selected as back then there was no standard format yet, resulting in for example bad camera angles and low quality footage.
Furthermore the voting results of each debate were collected as well as the gender of the speakers. 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[2]. Example motions include "Break up big banks" and "Genetically modify food". A full list of the debates, including topics, speakers and voting results, is provided in Appendix A.

3.3 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

[1] `http://intelligencesquaredus.org/`
[2] `https://www.youtube.com/user/IntelligenceSquared`

experience with public speaking and are leading experts in the field of the debated motion.
A strict structure is followed for these debates (Figure 3.2). 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 takes questions from the audience for the panelists and also asks the panelists questions himself. 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 3.1). 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 3.1: 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.

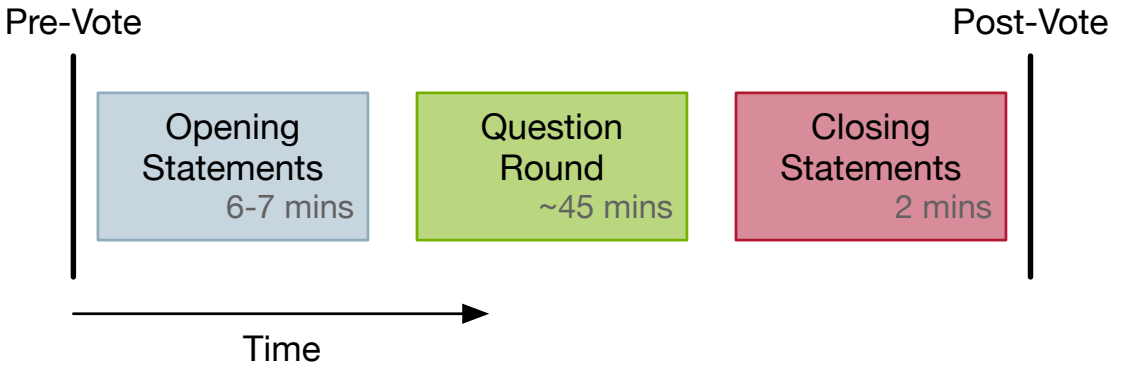


Figure 3.2: Oxford style debating overview figure. Prior to the start of the debate the audience votes on the motion. This motion is followed by opening statements by each speaker in turn. The opening statements are followed by a question round and the debate is closed by individual closing statements. After these statements a second vote is conducted. The team that swayed the most votes in their favor wins the debate.

3.4 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, as this question round might be important to the result of the debates it is something that could be included in future investigations.

3.4.1 Debate Annotation

I annotated the time frames for the opening and closing speech in ELAN [63][3] for all 36 debates. In a sub-tier these speeches have also been annotated on whether or not the camera had a frontal view of the speaker. As these frontal videos would later be used to extract facial data it was important that these video segments only included the current speaker. An example of this annotation is shown in Figure 3.3. ELAN files were given names in accordance with the date of the debate.

3.4.2 Video Extraction

In order to extract the correct video segments, as annotated, the SALEM toolbox [20] was used. This toolbox includes a slice function that can be used to extract time stamps to Matlab based on the ELAN annotations. A Matlab script was written to automatically import these time stamps into Matlab. In the next step these time stamps were used to extract the shorter video segments from the complete debate video making use of the functionality of the Matlab FFmpeg toolbox[4]. Naming the

[3] `http://tla.mpi.nl/tools/tla-tools/elan`
[4] `http://www.mathworks.com/matlabcentral/fileexchange/42296-ffmpeg-toolbox`

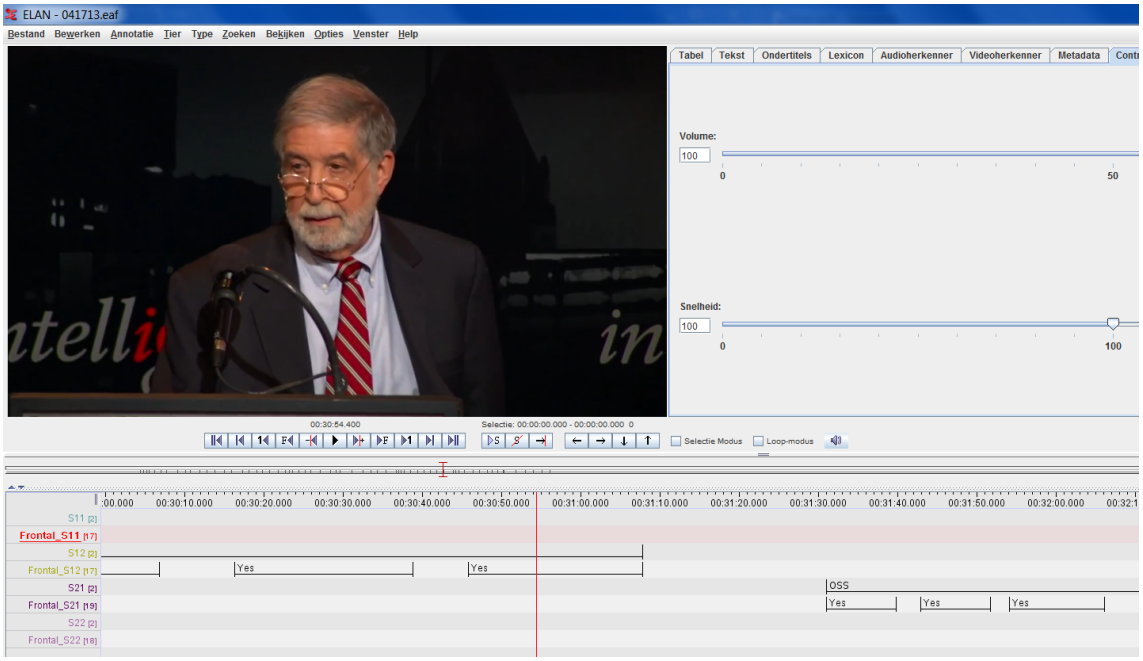


Figure 3.3: Example annotation showing the different tiers. Each debater has his/her own tier and sub-tier. Opening and closing statements are annotated on the same level.

video files was done automatically based on the type of video (i.e. opening or closing) and the name of the ELAN file. These steps result in having eight separate videos, one for each closing and opening speech per debate, as well as having separate video segments for all frontal video data present in the debate.

3.4.3 Audio Extraction

The eight statement videos per debates were converted to 16 kHz mono wav audio files using FFmpeg[5] through a small script in a batch file. This frequency was chosen as it is the required frequency for the toolbox that would later be used to extract the audio features. This toolbox is discussed in the next chapter.

3.4.4 Text Extraction

The text that was needed for the research was collected by copying each individual opening or closing statement to a separate text file from the complete debate transcript. These transcriptions are clean versions, meaning that for example hesitations are not included. Transcripts are also not time aligned with the audio files.

[5] `https://www.ffmpeg.org`

3.5 Final Dataset

The final dataset consists per speaker of:

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

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. All 120 speakers (19 female, 101 male) from these debates were included in the dataset, originating from various ethnicities and nationalities.

3.5.1 Dataset Balance

Given the small number of females it is not feasible to separately train classification methods for females and males. It is therefore important that the dataset is balanced around females. This happens to be the case as ten females won the debate while nine lost. Furthermore a different group of ten females debated for a motion while nine debated against it.
Another factor where the dataset could be skewed is if either the for or the against teams wins the debate a lot more than the other side. An example for where this might cause problems is in the emotion analysis. One could for example hypothesize that the team debating for the motion would be more positive than the one debating against the motion. However the teams proposing the motion won fifteen times in total, which is the same amount as the teams opposing the motion. All this makes for a well balanced dataset.
A few speakers spoke in multiple debates. 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 [10][1]. A large number of voice feature 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. A sample rate of 100 samples per second was used. Breathy voice and tense voice are often considered to be on the opposite ends of the voice quality continuum [17]. Voice quality is the timbre or auditory coloring of a person's voice. It is an important aspect in the perception of emotion in speech [18].

- **Fundamental frequency ($\mathbf{f_0}$)**: The fundamental frequency, or pitch, is tracked using the method suggested in [12] based on residual harmonics. This method simultaneously detects if a segment is voiced or unvoiced and is especially suitable for noisy conditions.

- **Formants ($\mathbf{F_1}$ and $\mathbf{F_2}$)**: [7] 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 [48].

[1] `http://covarep.github.io/covarep/`

- **H1-H2 ratio**: The H1-H2 parameters describe the amplitude of the fundamental frequency relative to that of the second harmonic. It is used as an indication of the open quotient [21].
- **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) [31]. 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 [30].
- **Liljencrants-Fant model parameter Rd**: The final measure is one of the R-parameters of the Liljencrants-Fant (LF) model characterizing the glottal source [16]. The Rd captures most of the covariation of the LF model parameters. In [10] 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.

Given the long duration of the segments it is not practical to keep the features on a per sample basis. Thus we decided to compute the mean and standard deviation over the entire opening or closing statement after removing the unvoiced data for these features. 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. Some tests were also performed with for example taking the mean and standard deviation over multiple shorter voiced segments (i.e. using 5 second fragments), thus resulting in far more features per individual, but this did not improve the machine learning results.

4.2 Video Features

Paul Ekman showed in his studies that people can universally recognize the expressions of seven emotions [15]. We extracted evidence for the presence of these seven basic emotions (joy, anger, surprise, fear, sadness, disgust and contempt) using Facet[2] 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. A number of these emotions as used by the Facet software are shown in Figure 4.1. The features were extracted at a sample rate of 30 frames per second. Facet outputs a comma-separated values (csv) file that stores tabular data per video, these files were then imported into Matlab.

Figure 4.1: Six of the basic emotions; from left to right and from top to bottom: anger, disgust, fear, joy, sadness, and surprise.

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

Nineteen features that describe elementary facial muscle actions, called action units (AU) in the Facial Action Coding System (FACS) [56], were computed as well. Each action unit has its own number, for example AU1, and it roughly corresponds to an individual muscle of the face. A full list of the action units used, including example images, is given in Appendix C.

After combining the features from all the frontal videos for each speaker we computed the mean and standard deviation over these features for a total of 46 Facet features per statement.

4.2.1 Face and Gaze Movement

Using Omron's Okao[3] we extracted nine more video features using a sample rate of 30 frames per second. Okao outputs a .txt tabular file for each video, which were imported into Matlab for further processing. The features 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 heavily biased towards on which side of the stage the debater presented (see Figure 3.1a). A total of 16 features are acquired this way.

4.3 Text Features

We applied a text analysis method called LIWC2007 [42] 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 [55, 58]. 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 resulting in 32 features. A list of all the LIWC features being used is given in Appendix B. LIWC outputs a csv file that lists the relative frequency per feature for each opening or closing speech.

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

5

Statistical Evaluation

In order to investigate research question **Q1** as to which nonverbal behaviors are indicative of successful debaters, we conducted statistical analysis with the extracted audiovisual 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. A t-test is used to determine if two sets of data are significantly different from each other. The effect size is a measure to indicate the strength of this difference. The value of g denotes the estimated difference between the two population means in magnitudes of standard deviations [22]. Hedges' g is a commonly used standardized mean difference measure that can be transferred into other measures like Cohen's d [14]. This measure was extracted with the help of the Matlab MES Toolbox[1].
In this section the features that differentiate the groups the most based on these tests are summarized. 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.

5.1 Winning versus Losing

Table 5.1 summarizes the result of this statistical evaluation with respect to winning versus losing debate performance characteristics and lists observed mean values, standard deviations, and effect sizes. Below major findings with respect to each modality is reported separately. In Chapter 7 these results are further discussed.
Audio: The audio features 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

[1] http://www.mathworks.com/matlabcentral/fileexchange/32398-measures-of-effect-size-toolbox

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. Action unit 20 its standard deviation also differs greatly between the two groups ($p<0.01$), being higher for the winners. Appendix C provides example images for these action units.

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*.

Table 5.1: Winning speakers versus losing speakers. Significance is denoted with * ($p < .05$) and ** ($p < .01$).

Feature	**Winner mean (SD)**	**Loser mean (SD)**	**Hedges' *g***
Audio Features			
f_0 Range	41.88 (15.10)	35.80 (13.71)	0.4203**
f_0 SD	32.45 (9.74)	29.02 (9.57)	0.3540**
Rd Confidence SD	0.0595 (0.010)	0.0626 (0.009)	-0.3285*
MDQ SD	0.0091 (0.0022)	0.0085 (0.0016)	0.3214*
H1-H2 SD	2.786 (0.661)	2.620 (0.526)	0.2776*
Mean F2	1540.60 (61.60)	1523.16 (60.81)	0.2841*
Mean QOQ	0.505 (0.062)	0.519 (0.058)	-0.2396
Video Features			
AU20 Evidence SD	0.455 (0.105)	0.415 (0.084)	0.4161**
Mean AU20 Evidence	-0.1520 (0.311)	-0.030 (0.358)	-0.3628**
Mean AU18 Evidence	-0.817 (0.684)	-1.045 (0.626)	0.3491**
Mean Joy Evidence	-1.587 (0.720)	-1.333 (0.745)	-0.3460**
Disgust Evidence SD	0.530 (0.168)	0.480 (0.142)	0.3244*
Mean Positive Evidence	-0.177 (0.583)	0.018 (0.635)	-0.3191*
Horiz. Face Direction SD	10.585 (3.055)	9.703 (3.458)	0.2695*
Horiz. Gaze Direction SD	17.990 (7.085)	20.423 (10.800)	-0.2655*
Text Features			
Perceptual Processes	1.273 (0.710)	1.596 (1.036)	-0.3628**
Hear Category	0.550 (0.484)	0.732 (0.688)	-0.3055*
Discrepancy Category	2.006 (1.031)	1.751 (0.926)	0.2597*
Personal Pronouns	7.900 (2.810)	7.292 (2.276)	0.2551*
Social Category	10.318 (2.810)	9.693 (2.600)	0.2300

5.2 Opening versus Closing

An interesting phenomenon that we observe are the large differences between the opening speeches and the closing speeches[2]. This could somewhat be expected due to the different nature of the speeches. For example the opening speech is longer and is given standing up, while the shorter closing speech is given sitting down. This also means that there is far less data available for the closing speech compared to the opening speech, so differences could be the caused due to this reason. Furthermore the opening speeches are given at the start of the debate, while the closing statements are given right before the final voting round. Table 5.2 summarizes the result of our statistical evaluation with respect to opening versus closing statements and lists observed mean values, standard deviations, and effect sizes. Below significant findings with respect to each modality is reported separately.

Audio: There is a large difference for the audio features between the two statements. In particular, the pitch is a lot higher for the opening speech ($p<0.01$) as well as the pitch range ($p<0.01$) and its standard deviation ($p<0.05$). We also notice that the opening speech is more breathy than the closing speech, with higher PSP ($p<0.01$), Peak Slope ($p<0.05$) and MDQ ($p<0.05$) values with the other voice quality features following the same trend. The mean F1 ($p<0.01$) and F2 ($p<0.01$) frequencies are higher as well for the opening speech, while the first formant range is lower ($p<0.05$). The standard deviation of F2 ($p<0.05$) and NAQ ($p<0.05$) is lower for the closing speech, while standard deviation of the Rd confidence parameter is much higher ($p<0.01$).

Video: For facial expressions extracted using Facet software, we notice a greater standard deviation ($p<0.01$) for a large portion of the emotion related features for the opening statement, with slightly less significant results for the standard deviation of anger and sadness ($p<0.05$). The only emotions that do not display this behavior in standard deviation are fear, confusion and frustration. Furthermore, speakers show less confusion ($p<0.05$) and frustration ($p<0.05$) during the closing speech. The latter might be due to the fact that the speech is shorter. Action units 7, 18 ($p<0.01$), 17 and 23 ($p<0.05$) show significant differences as well. With six action units (4,7,14,15,20 $p<0.05$ and 25 $p<0.01$) showing large differences in their standard deviation. The large difference for the standard deviations for both the emotions as well as the action units is possibly caused by the far smaller amount of data available for the closing speech compared to the opening speech. The values of all these standard deviations are not included in Table 5.2.

From the Okao software we gather that speakers look down more during the closing speech ($p<0.05$). Which could be due to the fact that many debaters appear to read from a sheet of notes during their closing statements. They also appear to look around more with a higher standard deviation for both the horizontal face direction ($p<0.01$) and gaze direction ($p<0.05$). The standard deviation for the smile ($p<0.05$) is also greater for the opening speech.

[2]Given these large differences we also computed the results of the statistical tests between winners and losers on only the opening or closing speeches for the features mentioned in Section 5.1. The observed values for both opening and closing statements separately indicate similar trends, which is shown with the Hedges' g values in Appendix D.

Text: Speakers use pronouns ($p<0.01$) and in particular the personal pronouns *I* ($p<0.01$), *We* ($p<0.01$), and impersonal pronouns ($p<0.01$) relatively more during the closing speech. They also use more verbs ($p<0.01$). Furthermore they show more positive emotions during the closing speech ($p<0.05$) and use the future tense more ($p<0.05$). Words belonging to cognitive processes ($p<0.05$) are used more during the closing speech as well, in particular words from the discrepancy category ($p<0.01$). During the closing statement language belonging to the inclusive category (e.g. *and*, *include*) is used more often ($p<0.01$) as well. Lastly language involving social subjects ($p<0.01$) is used more during the closing speech.

Table 5.2: Opening versus closing statements. Significance is denoted with * ($p < .05$) and ** ($p < .01$).

Feature	**Opening mean (SD)**	**Closing mean (SD)**	**Hedges' *g***
Audio Features			
Mean F2	1547.41 (58.70)	1516.36 (60.94)	0.5172**
Pitch (f_0) Range	42.35 (13.92)	35.34 (14.70)	0.4881**
Rd Confidence SD	0.0579 (0.0099)	0.0615 (0.0087)	-0.3871**
Mean F1	470.44 (33.59)	456.46 (40.06)	0.3770**
Mean PSP	0.368 (0.070)	0.341 (0.075)	0.3664**
Mean Pitch (f_0)	161.02 (32.83)	149.34 (33.45)	0.3513**
NAQ SD	0.0333 (0.0068)	0.0312 (0.0056)	0.3372*
F2 SD	271.87 (26.12)	263.50 (24.46)	0.3299*
Pitch (f_0) SD	30.62 (9.22)	27.69 (9,14)	0.3176*
F1 Range	145.38 (30.34)	155.34 (36.61)	-0.2952*
Mean MDQ	0.121 (0.008)	0.119 (0.008)	0.2625*
Mean Peak Slope	-0.325 (0.046)	-0.336 (0.043)	0.2617*
Video Features			
Horiz. Face Direction SD	11.053 (3,076)	9.235 (3.250)	0.5728**
Mean Vertical Face Direction	-2.573 (8.028)	-5.118 (8.135)	0.3139**
Mean Frustration Evidence	-0.622 (0.588)	-0.818 (0.703)	0.3016*
Mean Confusion Evidence	-0.566 (0.505)	-0.732 (0.638)	0.2883*
Smile SD	16.056 (6.539)	14.057 (7.295)	0.2877*
Vertical Face Direction SD	7.855 (2.259)	7.170 (2.934)	0.2610*
Horiz. Gaze Direction SD	20.398 (9.350)	18.016 (8.919)	0.2599*
Text Features			
Pronouns	14.212 2.684)	16.171 (3.280)	-0.6514**
Verb Frequency	14.284 (2.388)	15.601 (2.717)	-0.5130**
Social Category	9.450 (2.676)	10.561 (2.659)	-0.4157**
Discrepancy Category	1.705 (0.719)	2.052 (1.172)	-0.3555**
Inclusive Category	5.2368 (1.210)	5.771 (1.749)	-0.3543**
Cognitive Processes	17.553 (2.414)	18.494 (3.386)	-0.3190*
Future Tense	0.989 (0.592)	1.196 (0.831)	-0.2851*
Positive Emotions	2.891 (1.003)	3.235 (1.417)	-0.2791*

5.3 For versus Against

In the used dataset the teams debating for the motion win just as much as the teams debating against the motion. The importance of having the dataset (somewhat) balanced around this shows through in the differences in the video features between the two groups. As we find large differences for several facial expressions between the two groups. We find that speakers debating for the motion express more joy ($p<0.01$) and less sadness ($p<0.01$). They also show more overall positive emotions ($p<0.05$). Furthermore the expected bias (mentioned in Section 4.2.1) in the mean horizontal gaze and face direction shows heavily in the statistical tests. A p-value much smaller than 0.01 was found for these two features.

6
Experiments

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

6.1 Method

Using all the features described in Chapter 4, a classifier was built to automatically determine the winning team of a debate.

6.1.1 Algorithm

From the feature extraction step we obtained 25 audio, 46 video and 32 text features per opening or closing segment. Given that we deal with non-sequential features a number of machine learning techniques can be used including support vector machines (SVMs), feedforward neural networks and Naive Bayes (also see Section 2.3). Both SVM and feedforward neural networks have been applied to many areas with excellent generalization results [36]. Support vector machines have been shown to perform similar, if not better, than feedforward neural networks in common machine learning problems. Furthermore training a SVM is generally easier and quicker to compute [64]. Naive Bayes performs worse than the former two in traditional cases, as this method is less sophisticated [59]. In addition as there was no license present at the Institute for Creative Technologies for the Matlab Neural Network Toolbox, nor any other open source Matlab toolbox available, the choice was ultimately made to use support vector machines. Matlab has a toolbox for SVMs, but a faster and easier to use toolbox in the form of the Matlab LIBSVM package [8][1] was used for this project.

6.1.2 Parameters

A support vector machine tries to separate the data through an optimal hyperplane, creating the largest margin between the two groups. This is depicted for a two dimensional case in Figure 6.1. In order to properly use this method the user has to choose a suitable kernel (including its parameters), as well as a regularization parameter (C-parameter). The C-parameter sets how strongly a SVM should try to avoid misclassifying training samples.

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

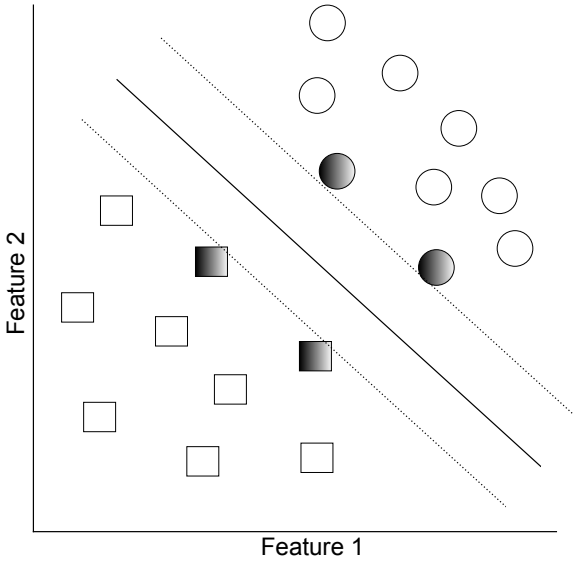


Figure 6.1: The idea behind a support vector machine for a two dimensional case. Squares indicate one class, while circles indicate the other. The shaded samples on the margin are called the support vectors. By using the support vectors the optimal margin between the two groups can be found.

A large C will choose a smaller margin between the two groups if this means that more of the training samples are separated. The reverse goes for a small value for C. Choosing a C that is too large means that the SVM will overfit to the training data, not allowing it to generalize across new data. A C that is too small will mean that the SVM will incorrectly classify samples, often even if it's possible to separate the data relatively easily.
Conceptually a kernel maps the original data to a higher-dimensional space allowing the data to be more easily separated by the algorithm using the linear hyperplane. This principle is illustrated in Figure 6.2. The optimal choice of this kernel is highly dependent on the data and the application and therefore should be found through testing [27]. There are several kernels, such as polynomial and Gaussian radial based function (RBF) kernels. Using no kernel at all is called using a linear kernel. A kernel has its own parameters; the degree for polynomial kernels or the gamma value for RBF kernels. The gamma value has a similar function as the C-value, as it controls how closely the algorithm should fit the training data. It can be seen as a parameter that determines the radius of influence of samples selected as support vectors.

6.1.3 Training and Testing

Given the large differences between the opening speech and closing speech features they were kept separate during training and testing. As there are four data sources; Facet video, Okao video, audio and text features, this gives a total of eight separate SVM classifiers in the case of decision fusion. Using this structure ended up providing the best results. Other structures, such as first feature fusing both video data sources, were tested as well but performed worse.

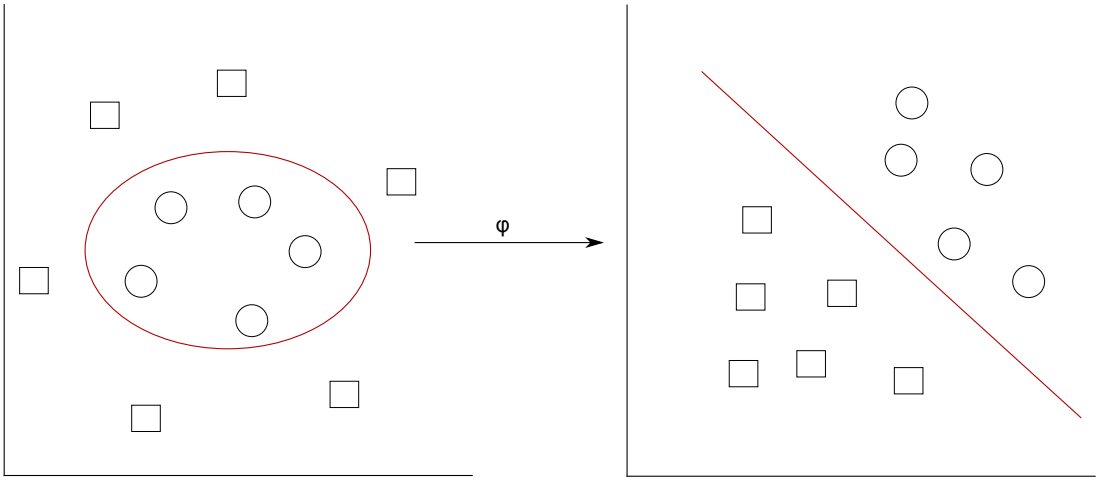


Figure 6.2: A kernel can be used to allow the computations to take place in a higher-dimensional feature space. For example this allows the function to the left to be computed in the space to the right.

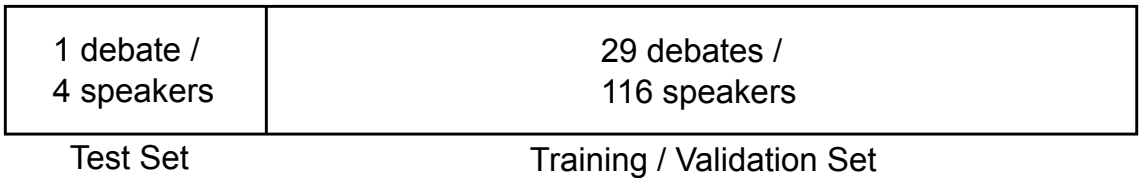


Figure 6.3: Leave-one-debate out testing. Each debate is used as a test set once.

In order to find the optimal kernel and its parameters a validation, training and testing procedure was followed; leave-one-debate-out testing was performed where one debate is kept for testing and the remaining 29 debates are used for training and validation (Figure 6.3). By using this validation approach the performance of the automatic classifier can be investigated independent of debate topics, debate teams, as well as individual debaters.
Min-max scaling was first applied to the training data to scale all the data to a range of [0,1]. This is done by subtracting the minimum value and then dividing by the difference of the maximum and minimum value for each feature (Equation 6.1). This same scaling is then applied to the test data. The main advantage of scaling is to prevent features that have a large numeric range to dominate features with a small range. Another advantage is that it can prevent numerical problems [26]. Min-max scaling is the recommended method for the LIBSVM toolbox.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{6.1}$$

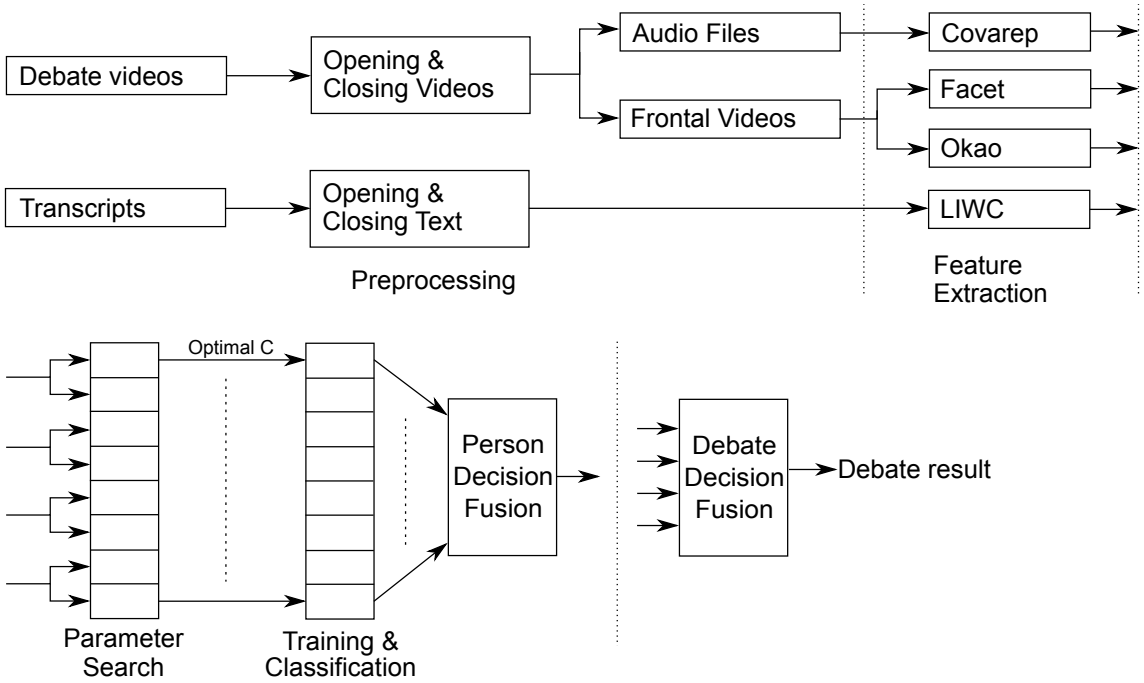


Figure 6.4: An overview of the entire system from preprocessing (Chapter 3), feature extraction (Chapter 4) to classification (Chapter 6). Opening and closing statements are individually processed, resulting in a total of eight initial classifiers.

The optimal kernel was found by testing polynomial kernels with a degree of 2 through 5, as well as the linear and RBF kernels. In order to find the optimal C-parameter per kernel we performed 5-fold cross validation[2] 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. Next the entire training set is used to train a model given the found C-parameter, which is then used to find the accuracy on the test set. The accuracy is defined as the number of speakers correctly classified. This process is repeated 30 times, once for each debate excluded from the training set. In case of the RBF kernel a grid search is performed where a combination of the gamma value and the C-value is tested per trial. This gamma value ranged from 2^{-15} to 2^{3} with the same step size of $2^{0.5}$. Through these tests it was found that the second order polynomial kernel performed the best for all the modalities and was therefore chosen.
As an extra step we took the median value of the 30 C-parameters as our one optimal C-parameter to train a 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 are not necessarily representative for the entire data set. This last step provides us with a more robust C-values for all cases.
The final structure of the entire pipeline is shown in Figure 6.4.

[2]For cross-validation the training set is split into K equal parts, where K is the number of folds. Each testing round one part is used for testing, while the other parts (K-1) are used for training. The average accuracy after testing on all the parts is then given as the result. By using this process the result more precisely represents the accuracy on unknown test data compared to only using one training and test set.

6.1.4 Fusion Structure

Individual Debaters: After finding the optimal kernel and parameters eight models - one for each modality or data source and for opening and closing statements separately (see Figure 6.4) - were 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 classified as a winner. We do not further try to solve for these ties, but rather interpret them as having 50% correct as there are only two possible outcomes for each debate[3]. 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 6.1 on the next page. The first column of numbers indicate the percentage of persons that were correctly classified by each individual support vector machine. The second column indicates 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 6.2. 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 6.2 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. In total 22 debates are correctly classified, the classification of seven debates results in a tie and only one debate is misclassified. Figure 6.5 summarizes and visualizes the results of the multimodal fusion for both the individual debaters and the debate teams.

[3]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.

Table 6.1: 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 6.2: 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)

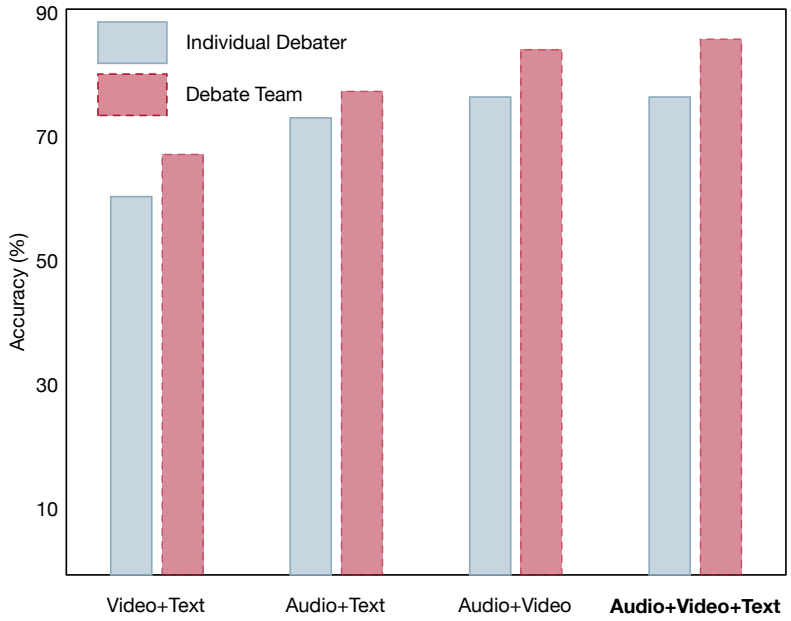


Figure 6.5: Results of the multimodal fusion in accuracy in % for both individual debaters as well as entire debate teams. Overall the multimodal fusion of all available modalities shows the most promising results and outperforms the other subsets of modalities.

7
Discussion

In this section the results with respect to the research questions presented in the introduction are discussed.

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 [53, 24, 49]. 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 [49]. 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 work that have been found to be of relevance in the past [49]. 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 given the serious nature of the discussed motions. 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 [4, 47].

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. [40] 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 [33, 49]. 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 concepts, such as increased anxiety, reduced expressivity or lack of competence.

Text: With our limited surface level linguistic analysis using the software LIWC [55], 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 [38, 35].

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 CUR-

RENNT, allowing for much faster processing times, is of interest to us [60]. Recurrent neural networks have shown promising results in the modeling of speech and human behavior in general [19].
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. Chapter 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 [55], 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 6.2 and Figure 6.5). 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 difficulty 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 thesis public speaking in the context of team based debates on a large variety of socioeconomic issues is investigated. 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 thesis 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.

Bibliography

[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] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[4] 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.

[5] 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.

[6] S. Bavister. What to do with your hands when you're presenting! `https://www.youtube.com/watch?v=ooOQQOQdhH8`.

[7] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Improved differential phase spectrum processing for formant tracking. In *Proceedings of Interspeech - ICSLP*, pages 2421–2424, 2004.

[8] 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.

[9] 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.

[10] 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*, pages 960–964, 2014.

[11] 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.

[12] 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.

[13] B. Dumas, D. Lalanne, and S. Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction*, pages 3–26. Springer, 2009.

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

[15] P. Ekman. Facial expressions. *Handbook of cognition and emotion*, 16:301–320, 1999.

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

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

[18] C. Gobl and A. N. Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.

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

[20] 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.

[21] 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.

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

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

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

[25] 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.

[26] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification, 2003.

[27] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, 37(4):543–558, 2004.

[28] M. S. Hussain, R. A. Calvo, and P. A. Pour. Hybrid fusion approach for detecting affects from multichannel physiology. In *Affective computing and intelligent interaction*, pages 568–577. Springer, 2011.

[29] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 215–222. ACM, 2013.

[30] 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.

[31] 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.

[32] C. L. Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[33] 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.

[34] 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.

[35] 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.

[36] M.-C. Lee and C. To. Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress. *International Journal of Artificial Intelligence & Applications*, 1(3), 2010.

[37] S. Lucas and Y. Suya. *The art of public speaking*. McGraw-Hill New York, 2012.

[38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26*, 2013.

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

[40] 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.

[41] 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.

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

[43] 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.

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

[45] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, et al. Multi-modal classifier-fusion for the recognition of emotions. *Coverbal Synchrony in Human-Machine Interaction*, page 73, 2013.

[46] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, and G. Palm. A generic framework for the inference of user states in human computer interaction. *Journal on Multimodal User Interfaces*, 6(3-4):117–141, 2012.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] W. H. Stevenson. Cutting out filler words. `https://www.toastmasters.org/Magazine/Articles/Cutting-Out-Filler-Words`.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Toastmasters International. Gestures: your body speaks. `http://web.mst.edu/~toast/docs/Gestures.pdf`.

[58] 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.

[59] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.

[60] F. Weninger, J. Bergmann, and B. Schuller. Introducing CURRENNT–the Munich open-source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research*, 15, 2014.

[61] 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.

[62] 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.

[63] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1556–1559, 2006.

[64] E. Zanaty. Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, 13(3):177–183, 2012.

A

List of Debates

The following table contains a list of all the debates as well as the voting percentages before and after the debate. Debates 3, 5, 12, 14, 15 and 20 weren't used for the current research as these fall below the threshold of having a minimum voting gain difference of 6 points between the two teams. The first 23 debates have a video resolution of 720p (1280 by 720 pixels) and the last thirteen are 360p (640 by 360 pixels). The first two speakers debated for the motion, the latter two against.

Nr.	Speakers	Pre	Post	Gain	Debate Motion
1	James Dobbins	25	37	12	Israel Can Live With a
	Reuven Pedatzur	25	37	12	Nuclear Iran
	Shmuel Bar	35	55	20	
	Jeffrey Goldberg	35	55	20	
2	Dr. Pamela Peeke	55	55	0	Obesity is the Government's
	Dr. David Satcher	55	55	0	Business
	John Stossel	19	35	16	
	Paul Campos	19	35	16	
3	Sheldon Krimsky	24	41	17	Prohibit Genetically
	Lord Robert Winston	24	41	17	Engineered Babies
	Nita Farahany	30	49	19	
	Lee Silver	30	49	19	
4	Alan Dershowitz	29	54	25	The President Has
	Michael Lewis	29	54	25	Constitutional Power
	Noah Feldman	44	39	-5	To Target And Kill
	Hina Shamsi	44	39	-5	U.S. Citizens Abroad
5	Ian Bremmer	25	35	10	Russia Is A Marginal Power
	Edward Lucas	25	35	10	
	Robert D. Blackwill	43	58	15	
	Peter Hitchens	43	58	15	
6	Frederic Mishkin	24	54	30	America Doesn't Need A
	John Taylor	24	54	30	Strong Dollar Policy
	Steve Forbes	29	37	8	
	James Grant	29	37	8	
7	David Brooks	65	65	0	The GOP Must Seize The
	Mickey Edwards	65	65	0	Center Or Die
	Laura Ingraham	14	28	14	
	Ralph Reed	14	28	14	

8	Kris Kobach	16	35	19	Don't Give Us Your Tired, Your
	Tom Tancredo	16	35	19	Poor, Your Huddled Masses
	Mayor Julian Castro	54	52	-2	
	Tamar Jacoby	54	52	-2	
9	Dr. Eben Alexander	37	42	5	Death Is Not Final
	Dr. Raymond Moody	37	42	5	
	Sean Carroll	31	46	15	
	Dr. Steven Novella	31	46	15	
10	Dr. Scott Gottlieb	24	53	29	The FDA'S Caution Is Hazardous
	Peter Huber	24	53	29	To Our Health
	Dr. Jerry Avorn	32	38	6	
	Dr. David Challoner	32	38	6	
11	Floyd Abrams	33	33	0	Individuals and Organizations
	Nadine Strossen	33	33	0	Have a Constitutional Right to
	Burt Neuborne	49	65	16	Unlimited Spending on Political
	Zephyr Teachout	49	65	16	Speech
12	Carmel Martin	50	67	17	Embrace the Common Core
	Michael Petrilli	50	67	17	
	Carol Burris	13	27	14	
	Frederick Hess	13	27	14	
13	Ahmed Rashid	23	23	0	The U.S. Drone Program Is
	John Kael Weston	23	23	0	Fatally Flawed
	Admiral Dennis Blair	34	64	30	
	General Norton Schwartz	34	64	30	
14	David Keating	19	22	3	Two Cheers For Super Pacs:
	Jacob Sullum	19	22	3	Money In Politics Is Still
	Trevor Potter	63	69	6	Overregulated
	Jonathan Soros	63	69	6	
15	Aaron David Miller	26	45	19	Flexing America's Muscles In
	Paul Pillar	26	45	19	The Middle East Will Make
	Michael Doran	31	45	14	Things Worse
	Bret Stephens	31	45	14	
16	Reuel Marc Gerecht	38	44	6	Better Elected Islamists than
	Brian Katulis	38	44	6	Dictators
	Daniel Pipes	31	47	16	
	Dr. M. Zuhdi Jasser	31	47	16	
17	Alex Abdo	46	66	20	Mass Collection of U.S. Phone
	Elizabeth Wydra	46	66	20	Records Violates the Fourth
	Stewart Baker	17	28	11	Amendment
	John Yoo	17	28	11	
18	Glenn Hubbard	28	30	2	The Rich Are Taxed Enough
	Arthur Laffer	28	30	2	
	Robert Reich	49	63	14	
	Mark Zandi	49	63	14	

19	Peter Singer	65	67	2	Legalize Assisted Suicide
	Andrew Solomon	65	67	2	
	Baroness Ilora Finlay	10	22	12	
	Dr. Daniel Sulmasy	10	22	12	
20	Paul Butler	45	58	13	Legalize Drugs
	Nick Gillespie	45	58	13	
	Asa Hutchinson	23	30	7	
	Theodore Dalrymple	23	30	7	
21	Robert Fraley	32	60	28	Genetically Modify Food
	Alison Van Eenennaam	32	60	28	
	Charles Benbrook	30	31	1	
	Margaret Mellon	30	31	1	
22	Dr. Neal Barnard	24	45	21	Don't Eat Anything With A Face
	Gene Baur	24	45	21	
	Chris Masterjohn	51	43	-8	
	Joel Salatin	51	43	-8	
23	Lawrence Krauss	37	50	13	Science Refutes God
	Michael Shermer	37	50	13	
	Ian Hutchinson	34	38	4	
	Dinesh D'Souza	34	38	4	
24	Dr. Scott Gottlieb	16	32	16	Obamacare Is Now Beyond Rescue
	Megan McArdle	16	32	16	
	Dr. Douglas Kamerow	53	59	6	
	Jonathan Chait	53	59	6	
25	Peter Schiff	17	9	-8	China Does Capitalism Better
	Orville Schell	17	9	-8	Than America
	Ian Bremmer	50	85	35	
	Minxin Pei	50	85	35	
26	Anant Agarwal	18	44	26	More Clicks, Fewer Bricks:
	Ben Nelson	18	44	26	The Lecture Hall is Obsolete
	Jonathan Cole	59	47	-12	
	Rebecca Schuman	59	47	-12	
27	W. Keith Campbell	18	38	20	Millennials Don't Stand a Chance
	Binta Niambi Brown	18	38	20	
	Jessica Grose	47	52	5	
	David D. Burstein	47	52	5	
28	Eli Pariser	28	53	25	When It Comes To Politics,
	Siva Vaidhyanathan	28	53	25	The Internet Is Closing Our
	Jacob Weisberg	37	36	-1	Minds
	Evgeny Morozov	37	36	-1	
29	Malcolm Gladwell	16	53	37	Ban College Football
	Buzz Bissinger	16	53	37	
	Tim Green	53	39	-14	
	Jason Whitlock	53	39	-14	

30	Peter Bergen	41	46	5	It's Time To End The War On
	Juliette Kayyem	41	46	5	Terror
	Michael Hayden	28	43	15	
	Richard Falkenrath	28	43	15	
31	Hannah Rosin	20	66	46	Men Are Finished
	Dan Abrams	20	66	46	
	Christina Hoff Sommers	54	29	-25	
	David Zinczenko	54	29	-25	
32	Mort Zuckerman	33	39	6	Grandma's Benefits Imperil
	Margaret Hoover	33	39	6	Junior's Future
	Jeff Madrick	32	55	23	
	Howard Dean	32	55	23	
33	Richard Fisher	37	49	12	Break Up Big Banks
	Simon Johnson	37	49	12	
	Paul Saltzman	19	39	20	
	Douglas Elliott	19	39	20	
34	Mark Zandi	45	69	24	Congress Should Pass Obama's
	Cecilia Rouse	45	69	24	Jobs Plan - Piece by Piece
	Daniel Mitchell	16	22	6	
	Richard Epstein	16	22	6	
35	Bryan Caplan	46	42	-4	Let Anyone Take A Job Anywhere
	Vivek Wadhwa	46	42	-4	
	Ron Unz	21	49	28	
	Kathleen Newland	21	49	28	
36	Richard Falkenrath	26	29	3	Spy On Me, I'd Rather Be Safe
	Stewart Baker	26	29	3	
	Michael German	41	62	21	
	David Cole	41	62	21	

B

Text Features

The table on the next page provides a list of all the LIWC categories used as the text features. A list of all available categories in the LIWC software is provided online[1].

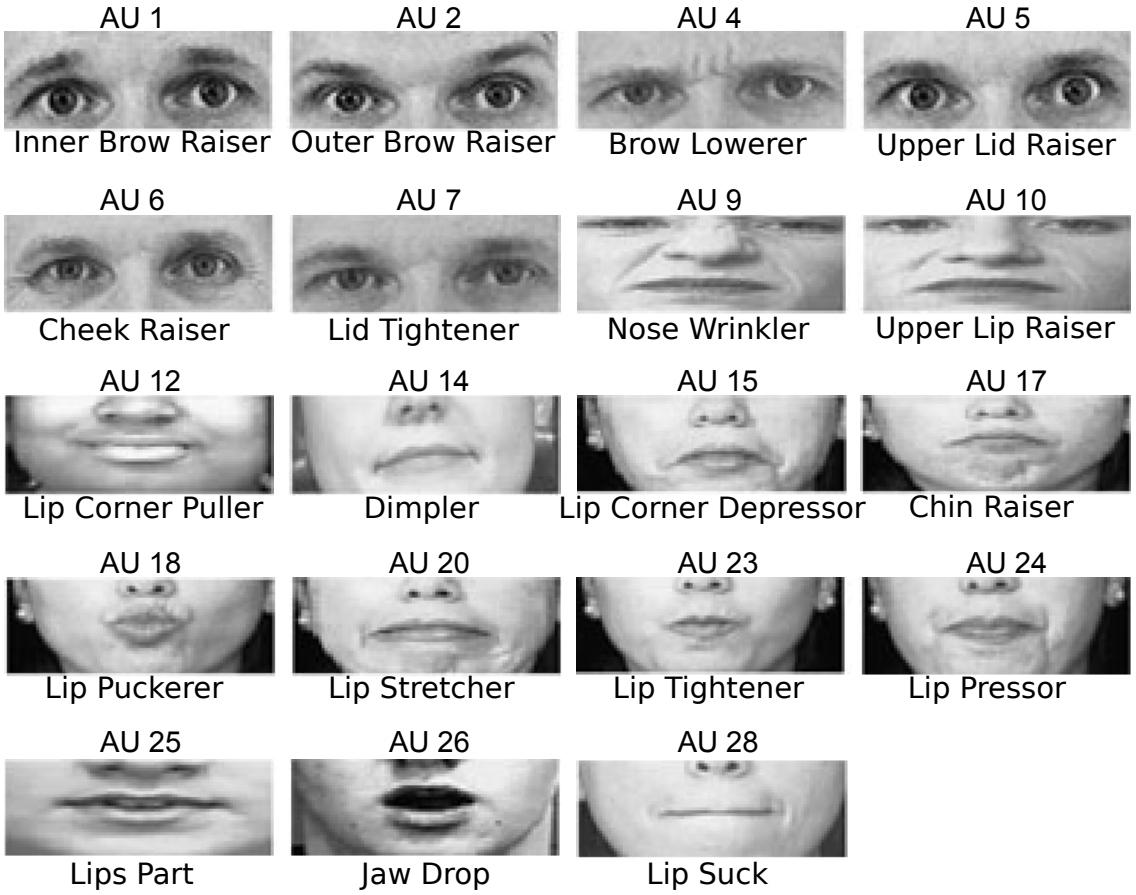
[1] http://www.liwc.net/descriptiontable1.php

Category	Examples
Linguistic Processes	
Total pronouns	I, them, itself
Personal pronouns	I, them, her
1st pers singular	I, me, mine
1st pers plural	We, us, our
2nd person	You, your, thou
3rd pers singular	She, her, him
3rd pers plural	They, their, they'd
Impersonal pronouns	It, it's, those
Articles	A, an, the
Psychological Processes	
Social processes	Mate, talk, they, child
Family	Daughter, husband, aunt
Friends	Buddy, friend, neighbor
Humans	Adult, baby, boy
Affective processes	Happy, cried, abandon
Positive emotion	Love, nice, sweet
Negative emotion	Hurt, ugly, nasty
Anxiety	Worried, fearful, nervous
Anger	Hate, kill, annoyed
Sadness	Crying, grief, sad
Cognitive processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain, stop
Inclusive	And, with, include
Exclusive	But, without, exclude
Perceptual processes	Observing, heard, feeling
See	View, saw, seen
Hear	Listen, hearing
Feel	Feels, touch

C

List of Action Units

The following is an overview of the action unit features as extracted by the Facet software[1].



S

[1]The images have been adapted from: `http://what-when-how.com/face-recognition/facial-expression-recognition-face-recognition-techniques-part-1`

D

Addition to Table 5.1

The following table is an addition to Table 5.1 and shows the results of the statistical tests between the winners and losers on only the opening or closing speeches. The Hedges' g values indicate similar trends as the combined closing and opening statement analysis[1]. 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. The combined Hedges' g is taken from Table 5.1 for easy comparison. Significance is denoted with * ($p < .05$) and ** ($p < .01$).

Feature	**Hedges' *g* Combined**	**Hedges' *g* Opening**	**Hedges' *g* Closing**
Audio Features			
f_0 Range	0.4203**	0.3893*	0.4696*
f_0 SD	0.3540**	0.2871	0.4374*
Rd Confidence SD	-0.3285*	-0.2606	-0.3879*
MDQ SD	0.3214*	0.2740	0.3641*
H1-H2 SD	0.2776*	0.2438	0.3121
Mean F2	0.2841*	0.3735*	0.2141
Mean QOQ	-0.2396	-0.2516	-0.2281
Video Features			
AU20 Evidence SD	0.4161**	0.4807**	0.3482
Mean AU20 Evidence	-0.3628**	-0.4701*	-0.2683
Mean AU18 Evidence	0.3491**	0.4124*	0.3072
Mean Joy Evidence	-0.3460**	-0.4721*	-0.2388
Disgust Evidence SD	0.3244*	0.2987	0.3841*
Mean Positive Evidence	-0.3191*	-0.4165*	-0.2326
Horiz. Face Direction SD	0.2695*	0.2539	0.3023
Horiz. Gaze Direction SD	-0.2655*	-0.3066	-0.2236
Text Features			
Perceptual Processes	-0.3628**	-0.2785	-0.4296*
Hear Category	-0.3055*	-0.2985	-0.3106
Discrepancy Category	0.2597*	0.2791	0.2636
Personal Pronouns	0.2551*	-0.0272	0.5325**
Social Category	0.2300	0.0944	0.3767*

[1]The main distinction being the difference for the Personal Pronouns category between the closing and opening speeches, which is caused by the fact that winners use *I* more during the closing speech ($p<0.05$), but less during the opening speech ($p<0.05$) compared to losers.