# ACOUSTIC AND PARA-VERBAL INDICATORS OF PERSUASIVENESS IN SOCIAL MULTIMEDIA

*Han Suk Shim, Sunghyun Park*, Moitreya Chatterjee*, Stefan Scherer, Kenji Sagae and Louis-Philippe Morency*

USC Institute for Creative Technologies, Playa Vista, 90094, CA, USA

## ABSTRACT

Persuasive communication and interaction play an important and pervasive role in many aspects of our lives. With the rapid growth of social multimedia websites such as YouTube, it has become more important and useful to understand persuasiveness in the context of online social multimedia content. In this paper, we present our results of conducting various analyses of persuasiveness in speech with our multimedia corpus of 1,000 movie review videos obtained from ExpoTV.com, a popular social multimedia website. Our experiments firstly show that a speaker's level of persuasiveness can be predicted from acoustic characteristics and para-verbal cues related to speech fluency. Secondly, we show that taking acoustic cues in different time periods of a movie review can improve the performance of predicting a speaker's level of persuasiveness. Lastly, we show that a speaker's positive or negative attitude toward a topic influences the prediction performance as well.

***Index Terms***— Persuasiveness, persuasion, speech analysis, social multimedia, online social multimedia content

## 1. INTRODUCTION

Persuasive communication that affects and alters opinions of others is an important aspect of many types of social interaction [1, 2]. Especially in the current society, communication through online social multimedia websites occupies a large portion of our daily lives. Since the communication medium heavily influences our decision-making process, it has now become very important and useful to understand persuasiveness in the context of online social multimedia. Many past studies were conducted to investigate factors in persuasive speech [3, 4, 5], yet only a few studies have investigated this topic from the computational perspective and from the context of online multimedia content.

In this paper, we show the relationship between acoustic / para-verbal characteristics and persuasiveness in online social multimedia. We present an extensive set of experiments with our multimedia corpus, which consists of 1,000 movie review videos obtained from a popular social multimedia website called ExpoTV.com. This paper particularly focuses on analyzing those speakers who are perceived to be in two extreme sides of highly or lowly persuasive speakers. We expect that comparing and analyzing the two contrary groups can provide a deeper understanding of the factors that make a speech more persuasive.

The work depicted here is sponsored by the National Science Foundation under Grant No. IIS-1118018 and the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statement and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

* The indicated authors made equal contribution to this work.

We firstly show that computational descriptors derived from acoustic characteristics in speech and fluency-related para-verbal cues can be predictive of persuasive speech. Secondly, we show the narrative characteristic of persuasiveness in speech by taking acoustic descriptors from different time periods of a speech. Lastly, we show that a speaker's positive or negative attitude toward a movie can affect the prediction performance when using both types of acoustic and para-verbal descriptors together. The main distinction of our work compared to past research in persuasion lies in investigating computational models of persuasiveness in social multimedia using acoustic and para-verbal descriptors.

## 2. RELATED WORK

Past research in psychology pointed out the importance of voice tone in persuasive communication [3]. Researchers showed that more intonation and higher speech volume affect perceived persuasiveness [6] and found that vocal pleasantness is positively correlated with perceived persuasiveness [7]. Some reported different effects of vocalic cues from conversational and dynamic speech styles on the perception of credibility and persuasiveness [4].

There are research findings that highlight the importance of speech fluency as well. Pause-fillers and speech rate were studied as speech behavior [8], and rapid speech rate was found to have a positive influence in persuasion [9]. In addition, higher speech rate and less halting speech were found to contribute to perceived persuasiveness [6].

Researchers interested in the affective analysis of speech have made much progress toward automatic recognition of emotion using computational indicators of speech [10]. On a similar topic of sentiment analysis, researchers have shown the possibility of automatically analyzing sentiment in speech using computational descriptors of prosody [11]. Furthermore, other researchers also analyzed charismatic speech from a corpus of American political speech and examined the correlation between acoustic descriptors and a speaker's charisma ratings [5].

Motivated by such findings, we present our experimental results of using computational descriptors of acoustic and para-verbal cues in speech in order to predict persuasiveness of a speaker in the context of online social multimedia.

## 3. RESEARCH HYPOTHESES

Our experiments were designed to address the following three main hypotheses.

**Acoustic Cues and Para-Verbal Cues of Fluency**: Para-verbal cues of speech fluency, such as pause-fillers and stuttering, might have a close relationship with perceived persuasiveness of a speaker. Past research [3] also pointed out the relationship between voice tone and

persuasion. Motivated by these findings, we tested the following hypothesis with our experiments:

*Hypothesis 1 (H1)*: Acoustic characteristics and para-verbal cues of fluency in speech are predictive of persuasiveness.

**Time-Segmented Information of Speech**: When we intend to persuade others, we structure our narrative in a non-constant fashion. For instance, we may put more emphasis in the beginning of our narrative and we may build up our narrative with a strong persuasive statement at the end. Thus, we built our second hypothesis on acoustic behavior:

*Hypothesis 2 (H2)*: Acoustic descriptors can lead to an improved prediction performance when we divide speech time into quarters and use them for predicting persuasiveness.

**Speaker's Preference on Topic**: People may speak in different speech styles when they try to persuade someone in favor of or against a topic. From this intuition, we hypothesized that the prediction model of persuasiveness would perform differently when it is trained solely on a group in which speakers are trying to persuade one positively or negatively.

*Hypothesis 3 (H3)*: Performance of persuasiveness prediction models varies with a speaker's attitude toward a topic.

## 4. DATASET

ExpoTV.com is a popular website housing videos of commercial product reviews. Each product review video has a speaker talking about a particular product as well as his/her direct rating of the product on an integral scale from 1 star (for most negative review) to 5 stars (for most positive review). This direct rating is useful for the purpose of our study because this allows us to study perceived persuasiveness in different directions of persuasion (in favor of or against). For instance, a speaker in a 5-star movie review video would most likely try to persuade his/her audience in favor of watching the movie while a speaker in a 1-star movie review video would argue against watching the movie. We collected a total of 1,000 movie review videos that were either highly positive or negative. The dataset consists of the following:

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

Each video in the corpus has a frontal view of one person talking about a particular movie. The average length of the videos is about 94 seconds with the standard deviation of about 32 seconds, and total length of all the videos combined in the corpus is around 26 hours. The corpus contains 372 unique speakers and 600 unique movie titles and is available to the research community for academic purpose .[1]

### 4.1. Evaluation of persuasiveness

Amazon Mechanical Turk (AMT), which is a popular online crowdsourcing platform, was used to obtain subjective evaluations of each speaker's persuasiveness [12, 13, 14]. For each video in the corpus, we obtained 3 evaluations on the level of persuasiveness of the speaker by asking the workers on AMT to give a direct rating on each speaker's persuasiveness on a Likert scale from 1 (lowly persuasive) to 7 (highly persuasive). A total of 50 native English-speaking workers based in the United States participated in the evaluation process online, and the task was evenly distributed among the 50 workers. To minimize gender influence, the task was distributed such that the workers only evaluated speakers of the same gender.

Once the evaluation was complete, we used the mean persuasiveness score for each video as the ground-truth measure of each speaker's perceived persuasiveness. In this initial effort, we focused on videos that were rated near the two end points in the persuasiveness level spectrum: highly persuasive or lowly persuasive. More specifically, videos with a mean evaluated score of equal to or greater than 5.5 were taken as highly persuasive ones, while those with a mean score of equal to or less than 2.5 were taken as lowly persuasive ones. We ended up with a total of 300 videos, specifically 157 videos of positive reviews (75 highly persuasive and 82 lowly persuasive) and 143 videos of negative reviews (62 highly persuasive and 81 lowly persuasive).

### 4.2. Transcriptions

Using AMT and 18 participants from the same worker pool for the persuasiveness evaluation, we obtained verbatim transcriptions of the 300 videos analyzed for this paper, including pause-fillers and stutters. Each transcription was reviewed and edited by in-house experienced transcribers for accuracy.

## 5. COMPUTATIONAL DESCRIPTORS

In this section, we give details about the extraction and computational encoding of acoustic descriptors and para-verbal descriptors of speech fluency, which were examined as potential candidates for accurately capturing the level of persuasiveness in speech.

### 5.1. Acoustic descriptors

Following common approaches for extracting descriptors and conducting speech analysis [15], we extracted various speech descriptors using publicly available software called Covarep [16].

- **Fundamental frequency (F0)** is closely tied to the affective aspect of speech [17].
- **Formants**, referring to the resonance frequencies of the vocal tract, are commonly used for speech recognition and emotion recognition. We explored formants F1 through F3.
- **Voice quality related features**: Many studies show a strong relation between voice quality and perceived emotion [18], and it is widely used for emotion recognition in speech. We used various voice quality related features as descriptors including normalized amplitude quotient (NAQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), quasi-open quotient (QOQ), difference between the first two harmonics (H1-H2), and Peak-Slope. For more details, readers are referred to [19, 20, 21].
- **Mel frequency cepstral coefficients (MFCC)** are widely used for speech and emotion recognition [22]. We explored MFCC 1 though MFCC 12 in our experiments.

From the above raw features, we derived statistical functionals comprising of the mean, standard deviation, range, skewness, 25th percentile (substitute for min), and 90th percentile (substitute for max). A total of 144 descriptors were encoded and used to explore their feasibility in capturing persuasiveness in speech.

### 5.2. Para-verbal descriptors of speech fluency

From the verbatim transcriptions of the dataset, we observed a set of frequent para-verbal cues that could be associated with the level of persuasiveness.

[1]Dataset available online: http://multicomp.ict.usc.edu/

- **Pause-Fillers**: Human speech is interspersed with various pause-fillers, such as um or uh. To account for the varying length of each review, we normalized the count of all instances of pause-fillers by the number of words spoken in the video.
- **Pause**: Human speech is interspersed with pauses. We computed this descriptor by normalizing the total duration of silence during speech by the length of each video.
- **Stutter**: Stuttering is when a speech flow is interrupted by involuntary repetitions. We counted all instances of stuttering in each video and normalized them by the number of words spoken in the video.
- **Articulation Rate**: Articulation rate is the rate of speaking in which all pauses are excluded from calculation [23], and it was computed by taking the ratio of the number of spoken words in each video to the actual time spent speaking.
- **Speech Disturbance Ratio**: Disruption of speech flow such as pause-fillers and stuttering can be considered as the same category [24]. We computed speech disturbance ratio by dividing the number of speech disturbance instances (pause-fillers and stutters) by the total number of words.

## 6. EXPERIMENTAL SETUP

This section gives detail on the experimental methodology, particularly on our prediction models and the experimental conditions that we designed for testing our research hypotheses.

### 6.1. Methodology

For our prediction model, the support vector machine (SVM) classifier with a linear kernel was used [25]. In all of our experiments, we performed 4-fold cross-validation with one fold for validating the model parameter (C) and another fold for testing. We repeated the cross-validation experiments 3 times with randomly generated folds. We note that our folds were created such that no 2 folds contained the samples from the same speaker. This speaker-independent approach allows better generalizability of our prediction models.

We used the following feature selection scheme for all of our experiments. We first selected a list of promising descriptors by performing paired-sample t-test for each descriptor compounding the highly persuasive group from the lowly persuasive group. Then, we included only those descriptors that showed marginal significance at $p < 0.1$. Using these descriptors, we performed a greedy-forward feature selection. Descriptor inclusion was stopped when the prediction accuracy dropped twice consecutively.

### 6.2. Experimental conditions

In order to address the first hypothesis (H1), we trained and tested prediction models on different groups depending on the types of computational descriptors:

- Acoustic: acoustic descriptors only.
- Para-verbal: descriptors of speech fluency only.
- Early fusion: both types of descriptors together.

To address the second hypothesis (H2), we conducted additional experiments with the acoustic descriptor group by computing the descriptors separately within different time periods in speech. More specifically, we divided each review video into quarters (4 different time periods of equal length) and computed the descriptors separately over each quarter time period:

- Time-segmented: acoustic descriptors (quarter-based) where the descriptors of all 4 quarter time periods are concatenated together, quadrupling the number of descriptors.

| Descriptor Group | Entire set of Reviews | Positive Reviews | Negative Reviews |
|---|---|---|---|
| Acoustic | 69.89 | 71.77 | **68.74** |
| Para-Verbal | 66.89 | 70.91 | 64.82 |
| Early Fusion | **74.89*** | **82.98***** | 68.26 |
| Majority Baseline | 54.33 | 52.23 | 56.64 |

**Table 1**. Prediction accuracies (in percentage) for 3 different sets of computational descriptors (H1) for positive and negative reviews separately (H3). Paired-sample t-tests showed statistical significance between early fusion and acoustic descriptors with $p^* < 0.05$, $p^{***} < 0.001$.

- Global: acoustic descriptors (entire length of review video) where descriptors are computed over the whole interaction as in H1.

In addition, we only used time-segmented descriptors from a single quarter at a time to see if any specific time period has influence on our prediction problem of overall period.

To address the third hypothesis (H3), we first performed all experiments under above conditions over entire set of reviews. Then, we repeated the experiments in two different groups depending on speakers' attitude (e.g. positive or negative opinions) on movies.

After conducting all the experiments, we analyzed each descriptor individually (e.g. F0 alone or pauses alone) among the acoustic and para-verbal descriptors to discuss the performance of each individual descriptor on predicting persuasiveness.

## 7. RESULTS AND DISCUSSIONS

In this section, we report and discuss our results in light of our research hypotheses, followed by a detailed descriptor analysis.

### 7.1. Acoustic and para-verbal descriptors (H1)

Table 1 shows that both acoustic and para-verbal fluency descriptors are contributing factors in predicting persuasiveness. For the prediction using the entire set of reviews (i.e. both positive and negative reviews combined), early descriptor-level fusion of the acoustic and para-verbal descriptors yielded 74.89% accuracy, which was well above the baseline performance of 54% and outperformed acoustic descriptors alone with 69.89% accuracy and para-verbal descriptors with 66.89% respectively. Early fusion models significantly outperformed the unimodal approaches with paired-sample t-tests showing $p < 0.05$. These findings confirmed our first hypothesis (H1).

### 7.2. Time-segmented acoustic descriptors (H2)

The prediction results of the global and time-segmented acoustic descriptors are summarized in Table 2. Across all samples, the time-segmented acoustic descriptors performed significantly better than the global acoustic descriptors, confirming the usefulness of adding time specific information to the acoustic descriptors. For the prediction of the entire set of reviews, time-segmented acoustic descriptors led to 72.78% accuracy with a statistically significant improvement

| Descriptor Group | Entire set of Reviews | Positive Reviews | Negative Reviews |
|---|---|---|---|
| Global | 69.89 | 71.77 | 68.74 |
| Time-segmented | **72.78*** | **78.98**** | **77.43**** |
| Majority Baseline | 54.33 | 52.23 | 56.64 |

**Table 2**. Prediction accuracies (in percentage) for the global and time-segmented (quarter-based) acoustic descriptors. Paired-sample t-tests showed statistical significance between the 2 types of descriptors with $p^* < 0.05, p^{**} < 0.01$.

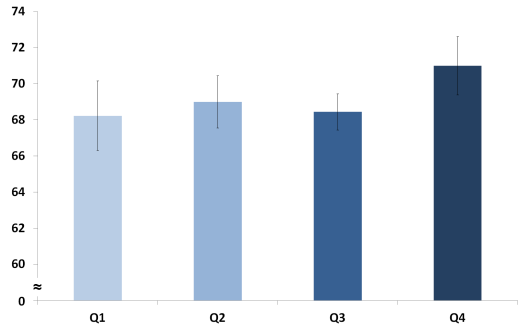


**Fig. 1**. Prediction performance using time-segmented acoustic descriptors within each time quarter only (both positive and negative reviews combined).

over those of global descriptors at 69.89% accuracy (paired-sample t-test with $p < 0.05$). This result confirmed our second hypothesis (H2).

Another interesting finding was that the performance of time-segmented acoustic descriptors varied considerably when used individually from each quarter of the review video (Fig. 1). In other words, instead of quadrupling the number of time-segmented acoustic descriptors (as the time-segmented model in Table 2), we only used data from a single quarter at a time to see if any specific time period showed better performances. We observed a general trend that the last quarter tended to have the best performance. We suspect that this might be due to people's tendency to emphasize the end of a speech to maximize the impact of their attempted persuasion.

### 7.3. Speaker's preference on topic (H3)

Table 1 also shows the prediction accuracy across different conditions of positive reviews only and negative reviews only. The early-fusion descriptor group showed a strong variance. The predictor, which is trained and tested using only the positive reviews, led to 82.98% accuracy. This was higher than the predictor trained on negative reviews only (68.26% accuracy), and those using both positive and negative reviews (74.89% accuracy). This result confirmed our third hypothesis (H3). It suggests that our computational approach can capture indicators of positive persuasiveness, i.e. the speaker tries to persuade the listener to watch a movie, better than negative persuasiveness or both types together. It would be possible that when a speaker in negative reviews group is criticizing a movie, people may focus more on the content of the speaker's statements instead of his/her vocal or fluency cues. Thus, acoustic and para-verbal characteristics could be less predictive than verbal characteristics in negative reviews group.

### 7.4. Descriptor analysis

Table 3 summarizes the top performing descriptors. These descriptors led to accuracies around 60% when used for prediction of persuasiveness. All the descriptors in Table 4 showed statistical significance on the Pearson correlation coefficient test with $p < 0.05$ except standard deviation and range of MFCC 4 in negative reviews.

Across all 3 groups (i.e. entire set of reviews, positive reviews, and negative reviews), MFCC descriptors that emphasize lower frequency regions - in particular MFCC 2 and MFCC 4 - stood out for predicting persuasive speakers in both positive and negative reviews.

Among all acoustic and para-verbal descriptors, the pause descriptor always showed the best individual performance, showing the importance of para-verbal cues in predicting persuasiveness.

| **Top Performing Descriptors** | **Prediction Accuracy (%)** | **Correlation Coefficient** |
|---|---|---|
| *Entire set of Reviews* | | |
| Pause | 66.67 | - 0.37 |
| MFCC 4 (standard deviation) | 65.00 | 0.25 |
| MFCC 4 (range) | 62.89 | 0.23 |
| F2 (range) | 60.33 | 0.24 |
| MFCC 2 (range) | 60.22 | 0.20 |
| *Positive Reviews* | | |
| Pause | 67.04 | - 0.39 |
| MFCC 4 (range) | 66.44 | 0.39 |
| MFCC 8 (range) | 64.54 | 0.29 |
| MFCC 4 (standard deviation) | 63.88 | 0.40 |
| MFCC 5 (range) | 62.45 | 0.26 |
| Peak-Slope (range) | 62.39 | 0.30 |
| MFCC 4 (skewness) | 61.61 | - 0.23 |
| MFCC 2 (Standard deviation) | 60.26 | 0.29 |
| Articulation rate | 59.46 | 0.29 |
| *Negative Reviews* | | |
| Pause | 62.72 | - 0.35 |
| F2 (range) | 61.44 | 0.25 |
| MFCC 4 (standard deviation) | 61.30 | 0.14 |
| F0 (standard deviation) | 60.96 | - 0.22 |
| MFCC 4 (range) | 60.53 | 0.11 |

**Table 3**. List of top performing computational descriptors with prediction performance when used individually. Pearson correlation coefficient between each descriptor and persuasiveness ratings.

The pause descriptor was negatively correlated with persuasiveness across all 3 groups, which suggests that persuasive speakers use fewer pauses. Analogously, articulation rate has a positive correlation with persuasiveness across all groups. This finding implies that persuasive speakers tend to speak faster.

Especially in the positive reviews group, the range of Peak-Slope showed high performance as a voice quality related descriptor. This finding suggests that persuasive speakers have a higher expressiveness in their voice. This descriptor also showed a positive correlation with persuasiveness in all 3 groups, but individually, it was not one of the top performing descriptors of persuasive speech for the entire reviews and negative reviews groups.

For the negative reviews group, the standard deviation of F0 descriptor (which is closely related to the variation of pitch) showed some predictive power in distinguishing highly persuasive and lowly persuasive speakers. Many speakers in the negative reviews group tended to frequently change their pitch when they tried to dissuade us from watching the movies. We suspect it was the reason that the predictive power of F0 was especially higher on negative reviews group.

## 8. CONCLUSIONS AND FUTURE WORK

We presented our experimental results of studying acoustic and para-verbal descriptors in the context of persuasiveness and social multimedia. In particular, we showed that both acoustic and para-verbal cues can be computationally encoded to predict persuasiveness in speech. For acoustic descriptors, we showed the benefit of adding time-segment information, which increased their prediction performance. In addition, we showed that a speaker's preference toward a topic affects the prediction performance of his/her persuasiveness. Promising future directions include exploring more automatic acoustic descriptors, more refined temporal modeling, and multi-modal fusion techniques with visual and language information together.

## 9. REFERENCES


[1] Kathleen Kelley Reardon, *Persuasion in practice*, Sage, 1991.

[2] Michael E Roloff and Gerald R Miller, *Persuasion: New directions in theory and research*, Sage Beverly Hills, 1980.

[3] Judith A Hall, "Voice tone and persuasion.," *Journal of Personality and Social Psychology*, vol. 38, no. 6, pp. 924, 1980.

[4] W Barnett Pearce and Bernard J Brommel, "Vocalic communication in persuasion," *Quarterly Journal of Speech*, vol. 58, no. 3, pp. 298–306, 1972.

[5] Julia Bell Hirschberg and Andrew Rosenberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proceedings of Eurospeech'05*, 2005.

[6] Albert Mehrabian and Martin Williams, "Nonverbal concomitants of perceived and intended persuasiveness.," *Journal of Personality and Social Psychology*, vol. 13, no. 1, pp. 37, 1969.

[7] Judee K Burgoon, Thomas Birk, and Michael Pfau, "Nonverbal behaviors, persuasion, and credibility," *Human Communication Research*, vol. 17, no. 1, pp. 140–169, 1990.

[8] Klaus R Scherer and Ursula Scherer, "Speech behavior and personality," *Speech evaluation in psychiatry*, pp. 115–135, 1981.

[9] Norman Miller, Geoffrey Maruyama, Rex J Beaber, and Keith Valone, "Speed of speech and persuasion.," *Journal of Personality and Social Psychology*, vol. 34, no. 4, pp. 615, 1976.

[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[11] François Mairesse, Joseph Polifroni, and Giuseppe Di Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5093–5096.

[12] Gelareh Mohammadi, Sunghyun Park, Kenji Sagae, Alessandro Vinciarelli, and Louis-Philippe Morency, "Who is persuasive?: the role of perceived personality and communication modality in social multimedia," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 19–26.

[13] Sunghyun Park, Gelareh Mohammadi, Ron Artstein, and Louis-Philippe Morency, "Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface," in *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*. ACM, 2012, pp. 29–34.

[14] Sunghyun Park, Philippa Shoemark, and Louis-Philippe Morency, "Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 37–46.

[15] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge.," in *INTERSPEECH*. Citeseer, 2009, vol. 2009, pp. 312–315.

[16] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep-a collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP14), Florence, Italy*, 2014.

[17] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 582–596, 2009.

[18] Christer Gobl and Ailbhe Nı Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1, pp. 189–212, 2003.

[19] John Kane, Stefan Scherer, Matthew Aylett, Louis-Philippe Morency, and Christer Gobl, "Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7982–7986.

[20] John Kane, Christer Gobl, Stefan Scherer, and Louis-Philippe Morency, "A comparative study of glottal open quotient estimation techniques," *BDL*, vol. 178, no. 15.17, pp. 0–41, 2013.

[21] Stefan Scherer, John Kane, Christer Gobl, and Friedhelm Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech & Language*, vol. 27, no. 1, pp. 263–287, 2013.

[22] Peipei Shen, Zhou Changjun, and Xiong Chen, "Automatic speech emotion recognition using support vector machine," in *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*. IEEE, 2011, vol. 2, pp. 621–625.

[23] Jana Dankovičová, "Articulation rate variation within the intonation phrase in czech and english," *age*, vol. 2, pp. 2, 1999.

[24] George F Mahl, "Disturbances and silences in the patient's speech in psychotherapy.," *The Journal of Abnormal and Social Psychology*, vol. 53, no. 1, pp. 1, 1956.

[25] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.