Bachelor Artificial Intelligence

University of Amsterdam

Automatic Descriptive Scoring of Public Speaking Performance Based on Prosodic Features

Towards Automatic Assessment of Public Speaking Performance

Max N. Bos

June 30, 2017

Supervisor: Dhr. drs. A. (Toto) van Inge
Abstract

Current human evaluation of public speaking performance is cost-inefficient, may be unreliable, and yields variable provision of constructive feedback. The present thesis aimed to investigate whether machine-generated prosodic features could predict human holistic scores in descriptive-word public speaking assessment, towards research on the development of a system for automatic assessment of public speaking performances. Experiments have been performed on which prosodic features can be utilized as speaker representative features, and results suggest that prosodic features related to the pitch range and pitch profile of a speaker can be speaker representative, and can be used to discriminate between similar and different speakers. In addition, clustering of similar feature vectors has performed, using the mean Euclidean distance of same speaker comparison as a non-similarity threshold, suggesting that speaker representative feature vectors can be classified as similar using a distance measure as threshold. Furthermore, the classification and regression models, utilizing Support Vector Machines (SVM), show promising results when focusing on target data related to prosodic features, suggesting the feasibility of automatic prediction of human holistic scores on public speaking performances.
Contents

1 Introduction 4

2 Related Work 6
   2.1 Automatic Assessment of Public Speaking Performance . . . . . . . . . . 6
   2.2 Analysis of Verbal Delivery . . . . . . . . . . . . . . . . . . . . . . . . . 8
      2.2.1 Prosogram . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

3 Method 10
   3.1 Data Collection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   3.2 Prosodic Feature Extraction . . . . . . . . . . . . . . . . . . . . . . . . . 11
   3.3 Data Pre-Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
   3.4 Comparison of Speaker . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
   3.5 Predicting Human Holistic Descriptive Scoring . . . . . . . . . . . . . . 15
      3.5.1 Classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
      3.5.2 Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

4 Experiments and Results 17
   4.1 Analysis of Prosodic Feature Vectors . . . . . . . . . . . . . . . . . . . . . 17
   4.2 Descriptive Word Classification . . . . . . . . . . . . . . . . . . . . . . . 21
   4.3 Predicting Descriptive Word Percentage Labels . . . . . . . . . . . . . . . 24

5 Discussion 27

6 Conclusion 29
   6.1 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

7 Acknowledgements 31

References 32
Communication and presentation skills are important to master, since they are essential for employability (Fallows & Steven, 2000; Kyllonen, 2012) and true academic study as they lead students to enter into debate and sustained reasoning (Morley, 2001). Public speaking is widely known to be the most feared form of communication (Pull, 2012) and proven efficacy of treatment programs show the cost of negative evaluation and negative views of one’s skills and appearance (Rapee, Gaston, & Abbott, 2009). Presentation training is the assessment of public speaking skills by observing and grading. It facilitates the receiving of evaluation and can thus be utilized to improve confidence in one’s communication and presentation skills. However, since current public speaking assessment relies on human scoring, the evaluation is cost-inefficient and may be unreliable.

First, since public speaking assessment rubrics may include measurements of non-speech related criteria, such as form and media, compensation of weak public speaking ability with non-speech related criteria may occur, weakening the assessment of the actual public speaking ability (Ward, 2013). Second, the assessment methods used are generally not consistent between assessed events within a programme, department or institution. In addition, feedback related to the results of the assessment to students is quite variable, resulting in students not having the ability to improve their public speaking performance by receiving constructive feedback.

This thesis will describe research using multimodal technologies toward developing an assessment system for automated scoring of presentation skills. Chen et al. (2016) recently performed research toward an assessment system for automated scoring of presentation skills using multimodal sensing. They developed a scoring model based on basic features in speech content, speech delivery, and hand, body, and head movements that significantly predicts human rating, suggesting the feasibility of using multimodal technologies in the assessment of public speaking skills.

While most current rubrics for evaluating public speaking performance attend to the content of the speech, the verbal delivery and nonverbal delivery of the content, this thesis will focus solely on the verbal delivery of the content using audio analysis. Furthermore, instead of utilizing data generated by human raters using extensive assessment rubrics, this research will be conducted over an oral presentation dataset that includes 559 individual audio files with corresponding human rated final holistic scores. The
holistic score represents an overall judgment by a human rater of a given presentation as a combination of descriptive words, combating the previously mentioned unreliability issues concerning human assessment by utilizing assessment rubrics. Recently, Chen et al. (2016) performed research on predicting holistic numerical scores of oral presentations, using machine learning methods to create several models based on basic features in speech content, speech delivery, and hand, body, and head movements, suggesting the feasibility of automatically predicting final holistic scores for public speaking performances.

The primary research question, that will be investigated in this thesis, is whether machine-generated prosodic features can predict human holistic scores in descriptive-word public speaking assessment. First, research will be performed on which prosodic features could be utilized as speaker representative features, by comparing speaker representative features of same and different speakers. Second, research will be performed on whether a speaker representative feature vector could be used as a subjective standard for speaker assessment, by classifying feature vectors in clusters of similar feature vectors using a non-similarity threshold. Third, the feasibility of utilizing prosodic features to predict human holistic scores using machine learning methods will be examined, using differing classification and regression approaches.

The present thesis is organized as follows: Section 2 provides a literature review on previous research performed on the automatic assessment of public speaking performance, and provides an analysis of the verbal delivery aspect of a presentation. Section 3 provides a description on the data collection, prosodic feature extraction, data pre-processing and approaches taken to answer the stated research questions. Section 4 describes the performed experiments, their corresponding results, and evaluation. Finally, Sections 5 and 6 summarize the findings of this thesis, aim to answer the stated research questions, and provide future research directions.
CHAPTER 2

Related Work

2.1 Automatic Assessment of Public Speaking Performance

Current human-scoring assessment of public speaking performance is made utilizing rubrics identifying what their authors consider to be core competencies for the practice of public speaking. According to Schreiber et al. (2012), the central core competencies to include in any measure of public speaking proficiency, evident in the literature on public speaking assessment, include determination of topic and purpose, use of supporting material, organization of ideas, speech introduction, speech conclusion, language use, verbal delivery, nonverbal delivery, and audience adaptation. There is an emerging body of literature on using unimodal and multimodal resources, representing these core competencies, to evaluate public speaking performance. This section will provide a review of such research, in particular regarding the results of and reasoning for utilizing the respectively selected prosodic features.

An example of unimodal research toward developing a system for oral presentation evaluation is a patent publication by Silverstein and Zhang (2003). Silverstein and Zhang claim a system for providing feedback information in real time to characteristics of oral presentation to a speaker. The evaluation, which can be used by the speaker to improve its presentation, is performed by analyzing representations of the audio signal corresponding to the oral presentation. The system evaluates an oral presentation by first processing the provided audio signal, subsequently the results are passed to an oral presentation analyzer, thresholding values that concern speech volume, pace, filler words, tone, long pauses and time.

Hincks (2005) analyzed prosodic variables in a corpus of eighteen oral presentations to prove the hypothesis that speakers with high pitch variation would be perceived as livelier speakers. The paper also investigates rate of speech and fluency variables, and concludes that the use of intonation by a speaker correlates with the perceived liveliness of a speaker. She proposed an application for this research in presentation skills training, where computer feedback could be provided for speaking rate and the extent to which speakers have been able to use their voices in an engaging manner. Nuchelmans (2016) performed research on developing the application as suggested by Hincks as part of a graduation project. The paper describes the implementation of an application for giving real-time feedback on one’s verbal presentation skills. The feedback is generated based on
three metrics, namely average pitch, pitch variation and speech tempo. For evaluation of the developed application, a collection of TED talks, videos of expert speakers on a topic in their field of research, is used. From the evaluation is concluded that the differences in pitch variation between speakers were significant.

Multimodal research on public speaking performance has been conducted by Rosenberg and Hirschberg (2005), which presents a study of charismatic speech, based upon elicited subject ratings of charisma and other personal attributes of speakers in a corpus of American political speech. Rosenberg and Hirschberg examined the correlation between the lexical and prosodic characteristics of speech tokens rated highly for charisma and reported significant correlations between charisma ratings and duration of token in words, seconds, and number of internal phrases; the number of first person pronouns; the complexity of lexical items in the token measured in number of syllables per word; pitch features including mean, normalized mean, standard deviation, and maximum pitch value; mean intensity; and speaking rate.

In 2014, a dataset of oral presentations, including corresponding slides and multimodal recordings, was provided to the participants of the Third International Multimodal Learning Analytics (MLA 2014) grand challenge and workshop. This dataset, known as the Oral Presentation Quality Corpus (ESPOL, 2014), was composed by 448 multimodal recordings on 86 oral presentations of undergraduate student groups, where each student group contained an average of four speakers. In addition, the dataset included human performed rating, on four point scales, on the individual students using a rubric that measured: speech organization, volume and voice quality, use of language, slides presentation quality, body language, and level of confidence during the presentation.

Using this dataset, Luzardo et al. (2014) performed an analogous research to Rosenberg and Hirschberg, on automatic prediction of the values assigned by the human raters using prosodic characteristics and features extracted from digital slides of 448 presentations. The prosodic characteristics from a presentation were calculated from the representing audio signal, involving the minimum, maximum and average pitch value; pitch standard deviation value; speech rate; articulation rate; and the average syllable duration. Machine learning methods were used to create several models that classify students in classes of low and high performers, based on the features extracted from the corresponding presentations. The models, created using only the audio features, reached up to 69% accuracy, with pitch and filled pauses related features being the most significant.

Chen et al. (2014), performed similar research, using the ESPOL dataset, on a more extensive set of features, namely features extracted from slides, speech, posture and hand gestures, as well as head poses. They examined the dimensionality of the human scores, which could be concisely represented by two Principal Component (PC) scores, for delivery skills and slides quality, and demonstrated that multimodal cues can predict human scores on presentation tasks.

The observation can be drawn from the mentioned research that the analogous research goal was to predict human scores on oral presentations based on data consisting of performed presentations with their corresponding human scored evaluations measured using an assessment rubric. Research on the assessment of public speaking (Ward, 2013) shows that assessment by humans utilizing assessment rubrics can be unreliable. Instead of utilizing data from extensive assessment rubrics, the automatic prediction could be based on given holistic scores. The holistic score would represent an overall judgment by a human rater of a given presentation as a numerical or descriptive category.
Recently, Chen et al. (2016) performed research on predicting human holistic scores of oral presentations, using two machine learning methods, namely polynomial Support Vector Machines (SVM) and Glmnet, to create several models based on basic features in speech content, speech delivery, and hand, body, and head movements. Experiments have been performed on the text, speech and visual features separately and combined, where the combination of text and speech features yielded an accuracy of 0.571, utilizing a polynomial SVM approach, and the combination of the three feature classes showed the most promising result, with an accuracy of 0.750.

2.2 Analysis of Verbal Delivery

Verbal delivery, the manner in which a speaker utilizes their voice to effectively convey meaning, is one of the central core competencies to include in any measure of public speaking proficiency. Prosody is the study concerned with those features of speech, such as intonation, stress, tempo, rhythm, and pause, that are used contrastively to communicate meaning, reflecting various features of the speaker or the utterance, such as the emotional state of the speaker, the form of the utterance, emphasis, and contrast. Prosodic features are referred to as suprasegmentals, since they are elements of speech that are not individual phonetic segments but are properties of syllables and larger units of speech. These features are calculated by combining prosodic variables, such as pitch, length and loudness of the voice which in acoustic terms are respectively known as the fundamental frequency (F0), duration and intensity. This section includes a description of the Prosogram, a system for semi-automatic transcription of prosody, in section 2.2.1.

2.2.1 Prosogram

Prosogram (Mertens 2004) is a system for semi-automatic transcription of prosody based on a stylization of the fundamental frequency data for syllabic nuclei, written by Piet Mertens, of the Department of Linguistics at the University of Leuven. The stylization is a simulation of human listener tonal perception (d’Alessandro & Mertens 1995), resulting in a more accurate representation of the intonation contour as perceived by human listeners, opposed to the fundamental frequency contour, since the auditory perception of pitch variations depends on many factors other than F0 variation. In addition, the system provides measurements of prosodic features for individual syllables, such as duration, pitch, pitch movement direction and size, as well as prosodic properties of longer stretches of speech, such as speech rate, proportion of silent pauses, pitch range, and pitch trajectory. The system is written as an executable script for Praat, a program for acoustic and phonetic research, written by Paul Boersma and David Weenink, of the Institute of Phonetic Sciences at the University of Amsterdam (Boersma et al. 2002).

The Prosogram system provides functionality to configure the analysis and plotting options when generating a Prosogram for a specified sound file. Figure 2.1 depicts a Prosogram with a wide rich view using automatic segmentation into syllable-sized units based on an audio sample representing French speech. It provides a prosodic transcription of the speech, annotated on a phonetic, syllabic, and word level, with the stylization (black bars), F0 (blue line), intensity (green line), intensity of the band-pass
filtered speech signal (magenta line), and voicing, as well as the phonetic segmentation. The vertical axes, define the pitch magnitude in semitones (ST) and hertz (Hz), on the left and right, respectively. The calculated prosodic data can be stored in files, including the stylization and intensity data, and holistic prosodic parameters concerning the pitch, phonation, pausing, and speech rate. The calculated prosodic parameters can be categorized into two different classes, namely pitch range related features, such as bottom, top, mean, median pitch values, and pitch profile related features, such as pitch movement direction, speech rate, and pitch trajectory.

From the previous research, as described in Section 2.1, the observation could be drawn that the analogous research goal was to predict human scores on oral presentations based on data consisting of performed presentations with their corresponding human scored evaluations measured using an assessment rubric. Since research on the assessment of public speaking (Ward, 2013) shows that assessment by humans utilizing assessment rubrics can be unreliable, this project will aim to predict human holistic scores of public speaking performances. Recently, Chen et al. (2016) examined this approach by prediction of numerical holistic scores based on basic features and speech content, utilizing polynomial Support Vector Machines (SVM). This thesis will use a dataset consisting of TED talks, an approach applied by Nuchelmans (2016), in combination with human descriptive word ratings related to each performance. The prosodic features of each speaker in a presentation will be extracted using Prosogram, and the utilization of the resulting two different classes of prosodic features, pitch range and pitch profile, will be examined.
CHAPTER 3

Method

This section provides a description on the data collection, prosodic feature extraction, data pre-processing and approaches taken to answer the research questions as stated in Section 1. As mentioned in Section 2, this thesis will aim to predict human holistic descriptive word ratings of TED talks. First, an analysis of utilizing different compositions of prosodic feature vectors, as extracted by Prosogram, will be provided. Subsequently, different SVM machine learning approaches will be utilized for the examination of automatic prediction of holistic descriptive word ratings, inspired by the results as achieved by Chen et al. (2016), using SVM models on a combination of textual and speech features.

3.1 Data Collection

The data collected for this thesis consists of 559 TED talks, high quality recorded videos of expert speakers on a topic in their respective fields of research. The talks have been performed by 492 different speakers, of which 52 speakers contribute more than one presentation. The talks originate from TED (Technology, Entertainment, Design), a non-profit media organization which posts talks online that may be used for research purposes under the Creative Commons License. TED provides transcripts related to these talks, with timestamps denoting the start of paragraphs and phrases. In addition, TED provides the functionality for viewers to rate a talk by selecting a maximum of three descriptive words from a predefined list of words, resulting in manual percentage-based classification of each talk. Figure 3.1 depicts the viewer rating results for a TED talk titled How to speak so that people want to listen by Julian Treasure (Treasure, 2017), and shows that the presentation is best described as persuasive, informative and inspiring.

The TED talks are additionally uploaded by TED to YouTube, an American video-sharing website. A custom Python program was written to automatically download videos from YouTube, using the youtube-dl command-line program, by traversing a manually constructed list of YouTube URIs that serve TED talks. Subsequently, the audio channel from each downloaded video file was extracted and stored as a WAV audio formatted file. In addition, the written program included the functionality to
automatically cut redundant sound before and after the actual speech representing the presentatio

As mentioned, TED provides a transcript of each talk, with timestamps denoting the start of paragraphs and phrases. The transcript of each talk was extracted from the TED website, and stored as Timed Text Markup Language (TTML), preserving timing information. YouTube auto-generates transcriptions, and relates a spoken word to the timestamp of utterance. This information could be used to relate text, specific words and phrases, to specific audio samples. Therefore, the auto-generated transcription of each talk was extracted from its representing YouTube page, and stored as TTML.

Furthermore, the manual classification data, representing percentage-based labeling by descriptive words, related to each video, was extracted from the TED website. Figure 3.2 depicts a boxplot representing the performed human-rating by descriptive words on the 559 TED talks. It visualizes the distribution of the percentages over the 559 talks for each descriptive word. From the figure, it can be concluded that the words Inspiring, Informative and Fascinating have high variance, with the highest median and maximum percentage values. In addition, descriptive words with a negative meaning, such as Long-winded, Obnoxious and Confusing, have small variance, with significantly low median and no significantly high maximum percentage values.

### 3.2 Prosodic Feature Extraction

After the collection of the TED talks data, the extracted audio files were processed for further analysis, using Prosogram (Section 2.2.1). For each talk, the representing audio file was automatically processed by Prosogram, using a command line call to the Prosogram Praat script, yielding a data file containing prosodic features that compose the pitch range of the speaker and, pitch and duration profile of the speaker. The parameters of Prosogram were set to calculate intermediate data files (no graphics files),
analyse the full time range, automatic detection of F0 range to a maximum of 450Hz, store the full parameter calculation in file, a frame period of 0.005s, and automatic segmentation based on acoustic syllables. The prosodic features regarding the pitch range of the speaker, calculated by Prosogram, include:

- **Range**: Estimated pitch range, in semitones (ST), based on 2%-98% percentiles of data in nuclei without discontinuities.
- **Bottom**: Lowest detected pitch value.
- **Mean**: Average of pitch values.
- **Median**: Median of pitch values.
- **Top**: Highest detected pitch value.
- **MeanOfST**: Mean of pitch values, where values are min and max pitch in ST for each syllable.
- **StdevOfST**: Standard deviation of pitch values, where values are min and max pitch in ST for each syllable.

The features regarding the pitch and duration profile of the speaker, include:

- **Gliss**: Proportion of syllables with large pitch movement, where the absolute distance is greater than or equal to 4ST.
- **Rises**: Proportion of syllables with pitch rise, greater than or equal to 4ST.
• **Falls:** Proportion of syllables with pitch fall, lower than or equal to -4ST.

• **TrajIntra:** Pitch trajectory, sum of absolute intervals, within syllabic nuclei, divided by duration, in ST/s.

• **TrajInter:** Pitch trajectory, sum of absolute intervals, between syllabic nuclei (except pauses or speaker turns), divided by duration (in ST/s).

• **TrajPhon:** Sum of **TrajIntra** and **TrajInter**, divided by phonation time, in ST/s.

### 3.3 Data Pre-Processing

This thesis contains research on automatic prediction of human-scored percentage labels for public speaking performance. Each label is a descriptive word with differing meaning, of which the word could be classified as a possible consequence of the conveyed content, or the manner in which the speaker utilizes verbal and nonverbal delivery to convey meaning. Table 3.1 lists the dictionary definition for each descriptive word, and its antonym, if available in the list. Based on the definition of the word, the word is classified to be a possible consequence of verbal delivery.

The hypothesis is made that 13 of the 14 descriptive words, namely Persuasive, Inspiring, Informative, Fascinating, Ingenious, Beautiful, OK, Courageous, Longwinded, Jaw-dropping, Unconvincing, Obnoxious, Confusing, have a low correlation with the verbal delivery aspect of a public speaking performance, and a more probable correlation with the content aspect of a speech. The list of descriptive words, for the dataset in the present thesis, includes the word Funny, a synonym for humorous. Research performed by Purandare et al. (2006), suggests the existence of correlation between humorous speech and the verbal delivery aspect of speech. They analyzed humorous spoken conversations from a classic comedy television show, by examining acoustic-prosodic and linguistic features and their utility in automatic humor recognition, and reported significant differences in prosodic characteristics, such as pitch, tempo and energy, of humorous and non-humorous speech.

Since not all descriptive words can be ascribed to be selected by a viewer as a consequence of verbal delivery, experiments will be performed where only a subset of the descriptive words, the words that could be a possible consequence of verbal delivery, are considered. The percentages related to the remaining labels of each presentation, are normalized to sum to unity, by dividing the percentage value of each label by the sum of all percentage values related to each presentation. The data related to the presentation is discarded from the dataset, when all label percentage values of a presentation are equal to zero, since this is considered to be sparse data.

Standardization of the prosodic feature data is performed, since the prosodic features are of differing units of measurement. First, the mean value of each feature is removed, after which the feature is scaled by dividing non-constant features by their standard deviation, resulting in the data to have zero mean and unit variance.
Table 3.1: Lists the 14 descriptive words composing the human-scored percentage labels for the public speaking performance data. For each word the definition and the antonym, if available in the list, is provided. The third column lists if a descriptive word could be an obvious consequence of verbal delivery.

<table>
<thead>
<tr>
<th>word</th>
<th>definition</th>
<th>possible direct consequence of verbal delivery</th>
<th>antonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persuasive</td>
<td>good at persuading someone to do or believe something through reasoning or the use of temptation.</td>
<td>No</td>
<td>Unconvincing</td>
</tr>
<tr>
<td>Inspiring</td>
<td>having the effect of inspiring someone.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Informative</td>
<td>providing useful or interesting information.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Fascinating</td>
<td>extremely interesting.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Ingenious</td>
<td>(of a person) clever, original, and inventive.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Beautiful</td>
<td>pleasing the senses or mind aesthetically.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>OK</td>
<td>denoting approval, acceptance, agreement, assent, or acknowledgment.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Funny</td>
<td>causing laughter or amusement; humorous.</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Courageous</td>
<td>not deterred by danger or pain; brave.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Longwinded</td>
<td>(of speech or writing) continuing at tedious length.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Jaw-dropping</td>
<td>amazing.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Unconvincing</td>
<td>failing to make someone believe that something is true or valid.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Obnoxious</td>
<td>extremely unpleasant.</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Confusing</td>
<td>bewildering or perplexing.</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

3.4 Comparison of Speaker

Current human assessment of public speaking performance is made by grading a student against criteria and standards utilizing an assessment rubric. Usage of textual assessment rubrics allows for assessor subjective grading, therefore leading to inconsistent grading between different assessors and variable provision of constructive feedback (Ward, 2013). Instead of human comparison of public speaking performance by criteria in a textual rubric, the assessment could be made automatically by direct comparison of the performance with a subjective golden standard. For instance, in teacher training, the public speaking performance of an aspirant teacher could be directly assessed by comparing its features with the features of a public speaking performance by a lead teacher. The features would comprise textual properties that characterize the Flesch-Kincaid Grade Level by readability analysis of speech content, and speech and visual properties that characterize the manner in which the teacher conveys meaning using verbal and nonverbal delivery, respectively. This section provides a description on research methods to determine which prosodic features could be used as speaker representative features.

As mentioned in Section 3.1, the available speaker data consists of prosodic features of 559 presentations related to the pitch range and pitch profile of the speaker, from 492 different speakers, including 52 speakers that contributed more than one presentation. Features that successfully represent a speaker would have similar values for different presentations given by the same speaker, assuming that the speaker has not changed its pitch range and profile. Therefore, one would expect that the distance between two feature vectors representing two presentations given by the same speaker would be small. The distance would be measured using Euclidean distance, a straight-line distance between two points in Euclidean space. Similarly, the distance between two features vectors representing two presentation given by different speakers would be small when
two speakers have a similar pitch range and profile. The distance magnitude would vary related to the similarity of the pitch range and profile of the speakers.

3.5 Predicting Human Holistic Descriptive Scoring

The previous section discussed a method for direct feature vector comparison of a single speaker or multitude of speakers, as a solution to the assessor subjective grading by textual assessment rubrics. In addition, a public speaking performance could be assessed by combining the feature vectors with target vectors that comprise holistic scores on the representing presentation. The holistic score would represent an overall judgment by a human rater of a given presentation as a numerical or descriptive category. In this manner, the feature vector representing a presentation could be related to a target vector representing a holistic score. Recently, Chen et al. (2016) performed research on predicting human holistic numerical scores of oral presentations, using two machine learning methods to create several models based on basic features in speech content and speech delivery, yielding an accuracy of 0.571 using Support Vector Machines (SVM) and 0.652 using Glimnet. The data collected for this thesis will be used to predict a holistic score by classifying a feature vector into one or more classes, where each descriptive word defines a class, by creating several classification models using Support Vector Classification (SVC), and OneVsRestClassifier (OVRC) methods, as will be discussed in Section 3.5.1. In addition, the data will be used to predict holistic scores by calculating the percentages for each descriptive word by using a Multi Output Regressor (MOR) in combination with Support Vector Regression (SVR), as will be discussed in Section 3.5.2. The machine learning methods are implemented using scikit-learn, a free software machine learning library for the Python programming language.

3.5.1 Classification

Two types of classification will be performed on the data, namely single-class and multi-class classification. Single-class classification will classify a presentation to a single class representing a descriptive word, such as Funny or Obnoxious, utilizing an SVC with a polynomial kernel with a degree of three. The target vectors in the dataset consist of 14 percentage-valued classes for each presentation, this will be reduced to one class representing 14 different labels, by only retaining the index of the class with the maximum percentage value. Multi-class classification will classify a presentation to a combination of three classes, by reducing the target vector for each presentation to the three classes with the highest related percentage values, such as Fascinating, Funny and Inspiring when these are the classes with the highest percentage values related to a presentation. This type of classification will be performed utilizing a hybrid between a OneVsRestClassifier and an SVC with a polynomial kernel with a degree of four.

Evaluation Metrics

**Precision** The proportion of relevant instances among the retrieved instances.

**Recall** The proportion of relevant instances that have been retrieved over the total relevant instances.
\textbf{F1-score} A weighted harmonic mean of the precision and recall, where the score can vary between 0 and 1, 1 denoting a good score:

\[
F_1 = 2 \frac{\text{precision} \ast \text{recall}}{\text{precision} + \text{recall}}
\]

3.5.2 Regression

Multi-output regression will be performed on the data, to predict the percentage value related to an arbitrary number of classes, utilizing a hybrid between a \texttt{MultiOutputRegressor} and a Nu Support Vector Regressor (NuSVR). In addition, single-output regression will be performed on single target classes, to predict percentage-values belonging to one single class, by converting the target data of fourteen descriptive words for each presentation to a single descriptive word. For instance, prediction of the percentage-value related to Funny for a sample presentation will be achieved by training a model with presentation data consisting of prosodic features and solely their related Funny percentage-value.

\textbf{Evaluation Metrics}

\textbf{The coefficient of determination $R^2$} The coefficient $R^2$ is defined as $(1 - \frac{SS_{res}}{SS_{tot}})$, where $SS_{res}$ is the regression sum of squares, $\sum_i (\hat{y}_i - \bar{y})^2$, and $SS_{tot}$ is the residual sum of squares, $\sum_i (y_i - \hat{y}_i)^2$. The score can range between 1.0 and an arbitrarily negative value, where 1.0 is the best possible score, 0.0 denotes a constant model that consistently predicts the expected value $y_i$, and negative if the model is arbitrarily worse.
CHAPTER 4

Experiments and Results

This section provides a description on the performed experiments, their corresponding results, and evaluation. Experiments have been performed on three compositions of prosodic features, namely pitch range related features, pitch profile related features and a combination of pitch range and profile features of the speaker. Section 3.2 provides a description of the specific prosodic features that compose the pitch range and pitch profile of a speaker.

4.1 Analysis of Prosodic Feature Vectors

As mentioned in Section 3.4, one would assume that the Euclidean distance between two feature vectors related to different presentations by the same speaker would be small. Similarly, the distance between two feature vectors of presentations by different speakers would be small, when the two speakers have similar pitch range and profile. The Euclidean distances between different presentations by the same speaker and distances between different presentations by different speakers have been examined.

The Euclidean distance was calculated, between each of the feature vectors for the presentations performed by the same speaker, for all 52 speakers with more than one presentation in the dataset. This is performed, for each multi-talk speaker, by iterating over the list of their feature vectors, and calculating the distance between the present feature vector and each remaining corresponding feature vector, resulting in a list of Euclidean distances between all feature vectors related to the same speaker.

Furthermore, the Euclidean distance between each feature vector of a multi-talk speaker and all other feature vectors, not including the data related to the multi-talk speaker, was calculated, resulting in a list of Euclidean distances between all feature vectors of multi-talk speakers and remaining feature vectors.

As mentioned in the introduction of this section, the feature vector comparisons of the same speaker and different speakers have been performed on three different compositions of prosodic features. Table 4.1 lists the mean Euclidean distance for the same speaker and different speaker comparisons for each composition of features. Notably, the mean distance for same speaker comparison is lower than the mean distance for different speaker comparison, confirming the hypothesis that the distance between feature
Table 4.1: Speaker comparison by Euclidean distance measure between feature vectors of presentations of same and different speakers, for three different feature sets. The second and third column depict the mean Euclidean distance (ED) for the same and different speaker, respectively.

<table>
<thead>
<tr>
<th>Features</th>
<th>Same Speaker (mean ED)</th>
<th>Different Speaker (mean ED)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>1.27925</td>
<td>3.45617</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>1.96443</td>
<td>3.60776</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>2.44059</td>
<td>5.19265</td>
</tr>
</tbody>
</table>

vectors related to the same speaker would be low and the variation in distance between feature vectors of different speakers would be high. Figure 4.1 depicts a combined density plot of the Euclidean distances for same speaker and different speaker comparison, for each composition of features.

Subsequently, experiments have been performed, determining the feature vectors most similar to each presentation in the dataset, using the mean Euclidean distance between same speaker feature vectors as a non-similarity threshold, resulting in a multitude of sets of similar feature vectors. The distribution of percentage-based rating, related to the similar feature vectors, was plotted, for each set of feature vectors similar to a presentation. Figure 4.2 depicts a boxplot for the ratings belonging to one presentation, given by Sophie Scott titled Why We Laugh, similar to a talk performed by Erin McKean, showing that the distribution of the percentage-valued ratings of the two talks is similar as well. From the figure can be concluded that the variance in percentages rated on the labels Inspiring, Informative, and Fascinating is high compared to the variance of Funny, Obnoxious, Persuasive and Unconvincing. This could be due to the former labels being more related to speech content features than verbal delivery features. The features for these presentations that contributed the lowest magnitude of similarity distance are PitchBottom, PitchMeanOfST, and PitchMean. Figure 4.2 shows the distributions of the percentage-valued ratings of three talks performed by Pico Iyer, of which the representing feature vectors have a similarity distance lower than the threshold. The figure shows that the variance for Inspiring, Informative and Fascinating are relatively low compared to the variance in the distribution of Figure 4.2, an explanation would be that the content of the presentations performed by Pico Iyer have similar characteristics that constitute an Inspiring, Informative and Fascinating speech.

Figure 3.2 depicts that the dataset includes presentations with discriminative percentage ratings for Funny. As mentioned in Section 3.3, research performed by Purandare et al. (2006) suggests the feasibility of detecting humorous and non-humorous speech using acoustic-prosodic features. For this reason, several experiments have been performed on a dataset, including 24 presentations related with Funny as the highest percentage-rated descriptive word, from 20 speakers, of which three contribute multiple presentations. Table 4.2 lists the mean and standard deviation Euclidean distance for the same speaker and different speaker comparisons for each composition of features. From the three feature combinations, the mean and standard deviation for pitch range features is notably low compared to pitch profile features and the combination of pitch range and profile features. Similar to the experiments on the original dataset, experiments have been performed determining the feature vectors most similar to each presentation in the dataset of 24 Funny presentations, using the mean Euclidean distance between same
Figure 4.1: Speaker comparison by Euclidean distance measure between feature vectors of presentations of same and different speakers, for three different feature sets.
Figure 4.2: Boxplot of the percentage-rated labels for sets of similar feature vectors, obtained by calculating a distance measure between feature vectors of presentations of different speakers, for two different feature sets.
Table 4.2: Speaker comparison by Euclidean distance measure between feature vectors of Funny presentations of same and different speakers, for three different feature sets.

<table>
<thead>
<tr>
<th>Features</th>
<th>Same Speaker</th>
<th>Different Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Pitch Range</td>
<td>0.995</td>
<td>0.429</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>2.726</td>
<td>1.565</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>3.011</td>
<td>1.410</td>
</tr>
</tbody>
</table>

speaker feature vectors as a non-similarity threshold, resulting in a multitude of sets of similar feature vectors. Figure 4.3 depicts two instances of the distribution related to each set of feature vectors for each composition of prosodic features. The combination of pitch range and pitch profile features seems to yield clusters with small variance in Funny percentage-ratings, with significantly small variance in the values of the pitch range related features.

4.2 Descriptive Word Classification

This section provides a description of experiments performed on single-class and multi-class classification of public speaking performances.

The ratings belonging to each sample in the dataset consist of fourteen percentage-valued classes. To be able to use this rating data for single-class classification, the class data of each sample is transformed to only contain the index of the class with the highest percentage value, resulting in a single class with fourteen different possible labels as target data.

Figure 4.4 depicts the frequency of each label after converting the fourteen descriptive word classes to a single class containing the corresponding label of the maximum percentage value. Notably, the target data contains a significantly large amount of data for Informative and Inspiring, a small amount data of data for Persuasive, Unconvincing, Beautiful, and Ingenious, and no data for Longwinded, Jaw-dropping, OK, Obnoxious and Confusing. Therefore, the classification model is solely trained on data related to Informative, Inspiring, Funny, Fascinating, and Courageous. The remaining dataset was split, using a 60-40 distribution, in random train and test subsets, respectively.

Table 4.3 lists classification reports for classification models trained on the three different compositions of prosodic features, using a Support Vector Classification (SVC) approach with a polynomial kernel with a degree of four. The pitch range and profile classification model is the best performing model, with an average precision of 0.53 and an accuracy of 0.52. Although the amount of data related to Funny speech is relatively small, the reports suggest the feasibility of using prosodic features for automatic classification of humorous speech.

Multi-class classification has been performed using a hybrid between a OneVsRestClassifier and an SVC with a polynomial kernel with a degree of four. The target data for each sample, related to the classes Informative, Inspiring, Funny, Fascinating, and Courageous, was transformed to solely contain the labels of three classes with the highest percentage values. For instance, when a presentation has an evaluation with the highest
Figure 4.3: Boxplot of the percentage-rated labels for sets of similar feature vectors, obtained by calculating a distance measure between feature vectors of Funny presentations of different speakers, for three different feature sets.
Table 4.3: Classification reports for classification models trained on three different compositions of prosodic features, using a Support Vector Classification (SVC) approach with a polynomial kernel with a degree of four.

(a) 

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funny</td>
<td>0.50</td>
<td>0.33</td>
<td>0.40</td>
<td>9</td>
</tr>
<tr>
<td>Inspiring</td>
<td>0.54</td>
<td>0.87</td>
<td>0.66</td>
<td>113</td>
</tr>
<tr>
<td>Courageous</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>Fascinating</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13</td>
</tr>
<tr>
<td>Informative</td>
<td>0.40</td>
<td>0.17</td>
<td>0.24</td>
<td>84</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.44</td>
<td>0.51</td>
<td>0.44</td>
<td>224</td>
</tr>
</tbody>
</table>

(b) 

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funny</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>8</td>
</tr>
<tr>
<td>Inspiring</td>
<td>0.49</td>
<td>0.94</td>
<td>0.65</td>
<td>104</td>
</tr>
<tr>
<td>Courageous</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>9</td>
</tr>
<tr>
<td>Fascinating</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>17</td>
</tr>
<tr>
<td>Informative</td>
<td>0.62</td>
<td>0.17</td>
<td>0.27</td>
<td>86</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.47</td>
<td>0.50</td>
<td>0.41</td>
<td>224</td>
</tr>
</tbody>
</table>

(c) 

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funny</td>
<td>1.00</td>
<td>0.11</td>
<td>0.20</td>
<td>9</td>
</tr>
<tr>
<td>Inspiring</td>
<td>0.53</td>
<td>0.88</td>
<td>0.66</td>
<td>113</td>
</tr>
<tr>
<td>Courageous</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>Fascinating</td>
<td>0.33</td>
<td>0.08</td>
<td>0.12</td>
<td>13</td>
</tr>
<tr>
<td>Informative</td>
<td>0.53</td>
<td>0.19</td>
<td>0.28</td>
<td>84</td>
</tr>
<tr>
<td>avg / total</td>
<td>0.53</td>
<td>0.52</td>
<td>0.45</td>
<td>224</td>
</tr>
</tbody>
</table>
Figure 4.4: Frequency of each label after converting the fourteen descriptive word classes to a single class containing the corresponding label of the maximum percentage value.

Table 4.4: Accuracy results for multi-class classification performed on three different sets of prosodic features combined with target data, where each sample is classified to a collection of the three most corresponding descriptive words.

<table>
<thead>
<tr>
<th>features</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>0.625</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>0.625</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>0.607</td>
</tr>
</tbody>
</table>

percentages for Informative, Funny, and Inspiring, it receives these three classes as its labels. Table 4.4 lists the achieved accuracies using the trained SVC models. Interestingly, the models trained on pitch range and profile related features achieved an accuracy of 0.607, which is lower than the accuracy achieved with the models trained with either pitch range or pitch profile features. Since the accuracies of the models trained with either pitch range or pitch profile features are in agreement, with a value of 0.625, the lower achieved accuracy on their combination could be ascribed to the skewed distribution of the dataset, where a large number of the presentations are highly labeled as either Informative or Inspiring.

4.3 Predicting Descriptive Word Percentage Labels

Experiments have been performed on the automatic prediction of human provided percentage-values related to the descriptive words belonging to a presentation. Two different machine learning approaches have been tested, namely multi-output regression, and single-output regression, utilizing a hybrid between a `MultiOutputRegressor` and a
Table 4.5: Results, measured by the coefficient of determination $R^2$ of prediction, for multi-output regression models trained on different compositions of features, where each presentation is related to target data containing percentage-rated values for Informative, Inspiring, Courageous, Fascinating, and Funny.

<table>
<thead>
<tr>
<th>features</th>
<th>coefficient of determination $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>0.062</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>0.056</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>0.108</td>
</tr>
</tbody>
</table>

NuSVR.

Multi-output regression has been performed on the presentation data in combination with the presentation related percentage-values for the descriptive words, Informative, Inspiring, Courageous, Fascinating, and Funny. An explanation for the selection of this subset of descriptive words is described in Section 4.2, where the frequency of high variance ratings for the descriptive words is examined, and concluded that the target data contains a small number of samples for Persuasive, Unconvincing, Beautiful, Ingenious, Longwinded, Jaw-dropping, OK, Obnoxious and Confusing. Table 4.5 lists the coefficient of determination $R^2$ of the predictions for models trained on three different feature sets. The $R^2$ coefficient of determination is considerably low for each model, while the highest result is obtained utilizing a combined feature set of pitch range and profile related features. The low performance could be ascribed to the descriptive words Informative, Inspiring, Courageous, and Fascinating, having a low correlation with the verbal delivery aspect of a presentation and a greater correlation with the content of a presentation.

The hypothesis that, the descriptive words Informative, Inspiring, Courageous, and Fascinating have a low correlation with prosodic features compared to Funny, could be examined by performing single-output regression to predict the corresponding percentage-values on each class individually. The classes Informative and Inspiring are well represented in the dataset, with a high variance in percentage-values, therefore, it is plausible to yield a high prediction utilizing single-output regression conditioning on that a high correlation exists between these classes and prosodic features representing the verbal delivery aspect of a presentation.

Table 4.6 lists the coefficient of determination $R^2$ for the different feature sets for single-output regression performed on Funny, Informative, and Inspiring. Since the dataset contains a relatively small number of samples with a high percentage-value for Funny, the dataset is split using a 90-10 distribution, of training and test data, respectively, for the single-output regression experiments performed on Funny. An $R^2$ coefficient of determination of 0.688 is achieved, when utilizing a combined feature set of pitch range and profile features, suggesting the feasibility of the hypothesis that humorous speech is correlated with verbal delivery.

Interestingly, a significantly low $R^2$ coefficient of determination of 0.106 is achieved when using solely pitch profile features, while a coefficient of 0.492 is achieved when using solely pitch range features, suggesting that pitch range features are more related to humorous speech. The single-output regression performed on solely Informative,
Table 4.6: Results, measured by the coefficient of determination $R^2$ of prediction, for single-output regression models trained on different compositions of features, where each presentation is related to target data containing a percentage-rated value for a single descriptive word. The results are listed for the models trained on the individual percentage-values for Funny, Informative, and Inspiring.

(a) Funny

<table>
<thead>
<tr>
<th>features</th>
<th>coefficient of determination $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>0.492</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>0.106</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>0.688</td>
</tr>
</tbody>
</table>

(b) Informative

<table>
<thead>
<tr>
<th>features</th>
<th>coefficient of determination $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>0.032</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>0.162</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>0.165</td>
</tr>
</tbody>
</table>

(c) Inspiring

<table>
<thead>
<tr>
<th>features</th>
<th>coefficient of determination $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Range</td>
<td>0.049</td>
</tr>
<tr>
<td>Pitch Profile</td>
<td>0.132</td>
</tr>
<tr>
<td>Pitch Range and Profile</td>
<td>0.022</td>
</tr>
</tbody>
</table>

achieved an $R^2$ coefficient of 0.165 using pitch range and profile features, confirming the hypothesis that the descriptive word Informative has a low correlation with verbal delivery. Notably, the $R^2$ coefficient for the prediction of the percentage-value related to Inspiring yields a higher result when solely pitch profile features are used, compared to using a combination of pitch profile and pitch range features. Interestingly, for both Informative and Inspiring, a higher $R^2$ coefficient is achieved when using pitch profile features instead of pitch range features, suggesting that there exists some correlation between pitch profile features and these descriptive words. Although there might exist some correlation between pitch profile features and the descriptive words Informative, and Inspiring, the results show that this correlation is significantly low compared to the correlation between humorous speech and prosodic features.
The research conducted in this thesis aimed to discover the feasibility of utilizing prosodic features for representing a speaker and the automatic prediction of human holistic scores for public speaking performances, towards research on the development of a system for automatic assessment of public speaking performance. A summarization of the achievements in this study is provided below.

First, presentation data collection was performed, using a custom program for downloading 559 TED talks, performed by 492 different speakers, of which 52 speakers contributed more than one presentation. The audio channel from each downloaded video file, representing a TED talk, was extracted and stored as a WAV audio formatted file. Furthermore, the manual classification data, representing percentage-based labeling by descriptive words, related to each video, was extracted from the TED website. It was hypothesized that 13 of the 14 descriptive words in the rating data would have a low correlation with the verbal delivery aspect of a public speaking performance. Solely the word Funny would have a probable correlation with this aspect of a presentation, since the descriptive word Funny is synonymous to the presence of humorous speech.

Second, prosodic features were extracted from each presentation representing audio file, using Prosogram \cite{Mertens2004}, a system for semi-automatic transcription of prosody based on a stylization of the fundamental frequency data for syllabic nuclei. This extraction yielded a data file containing prosodic features, that composed the pitch range of the speaker and, pitch and duration profile of the speaker, for each presentation.

Third, the feasibility of utilizing differing compositions of prosodic feature vectors as a speaker representation was examined. One would assume that the Euclidean distance between two feature vectors related to different presentations by the same speaker would be small. Similarly, the distance between two feature vectors of presentations by different speakers would be small, when the two speakers have similar pitch range and profile. The Euclidean distance between each of the feature vectors for the presentations given by the same speaker was calculated, for all 52 speakers with more than one presentation in the dataset. Furthermore, the Euclidean distance between each feature vector of a multi-talk speaker and all other feature vectors, not including the data related to the multi-talk speaker, was calculated. Notably, the mean distance for same speaker comparison is lower than the mean distance for different speaker comparison, confirming the hypothesis that the distance between feature vectors related to the same speaker would be low and
the variation in distance between feature vectors of different speakers would be high. In addition, the mean Euclidean distance for same speaker comparison was used as a non-similarity threshold, for the classification of feature vectors in clusters of similar feature vectors. The distribution of the target data related to these clusters showed that the descriptive words corresponding to the feature vectors can be significantly similar as well.

Finally, the feature vectors in combination with the related target data was used to train multiple machine learning models for the automatic prediction of human holistic scores. Four different machine learning approaches were examined, namely single-class classification, multi-class classification, single-output regression and multi-output regression. Using an SVC approach, promising results were achieved on a pitch range and profile single-class classification model, suggesting the feasibility of using prosodic features for automatic classification of humorous speech. The accuracies for the multi-class classification approach using either pitch range or pitch profile features were in agreement, and lower than the accuracy for the combined feature set, indicating that these models were subject to bias. The multi-output regression approach yielded a significantly low performance, which could be ascribed to the descriptive words Informative, Inspiring, Courageous, and Fascinating, having a low correlation with the verbal delivery aspect of a presentation and a greater correlation with the content of a presentation. For the single-output regression experiments performed on Funny, the results suggest the feasibility of the hypothesis that humorous speech is correlated with verbal delivery, and that pitch range features are more related to humorous speech than pitch profile features. The results for both Informative and Inspiring, suggest that there exists some correlation between pitch profile features and these descriptive words. Although there might exist some correlation between pitch profile features and these descriptive words, the results show that this correlation is significantly low compared to the correlation between humorous speech and prosodic features.
The primary research question, investigated in this thesis, was whether machine-generated prosodic features could predict human holistic scores in descriptive-word public speaking assessment. The research was divided into three sub-questions, the first questions which prosodic features could be utilized as speaker representative features. The results on the analysis of three different compositions of prosodic feature vectors, suggests that prosodic features related to the pitch range and pitch profile of a speaker can be speaker representative, since the representations can be used to relate similar speakers, and conversely, to classify different speakers as not similar.

The second sub-question was, whether a speaker representative feature vector could be used as a subjective standard for speaker assessment. The results on clustering similar feature vectors, using the mean Euclidean distance of same speaker comparison as a non-similarity threshold, suggests that speaker representative feature vectors can be classified as similar using a distance measure as threshold. However, using a distance measure as non-similarity threshold, does not provide an extensive comparison on the differences between each individual feature between subjects.

Lastly, the third sub-question was, whether prosodic features could predict human holistic scores using machine learning methods. The classification and regression models show promising results when focusing on target data related to prosodic features, suggesting the feasibility of automatic prediction of human holistic scores on public speaking performances.

6.1 Future Work

Although, the rating data related to humorous speech was relatively small, the present research suggests confirmation of the hypothesis that correlation exists between humorous speech and prosodic features. Since the current rating data merely includes criteria related to verbal delivery, the rating data should be extended with criteria for the evaluation of the actual verbal delivery aspect of one’s public speaking performance. In addition, the present research suggests confirmation of the hypothesis that merely some correlation exists between the descriptive words other than Funny and prosodic features, moreover a probable higher correlation with the content aspect of one’s public speaking
performance. Therefore, to investigate this hypothesis, the prosodic feature vectors representing each presentation should be extended with features that represent the content of the speech. Furthermore, the dataset of TED talks should be extended, especially with talks that are highly evaluated as Longwinded, Jaw-dropping, OK, Obnoxious, and Confusing, since these are not presented in the current dataset. Concludingly, the current research utilized features that provided average pitch value characteristics of a speaker in a speech, in addition, research could be performed on the analysis of the exact pitch trajectory for an entire speech, to indicate critical moments in a performance that define how a speaker variates or is perceived similar to a subjective speaker standard.
CHAPTER 7

Acknowledgements

First and foremost, I would like to thank my supervisor, Anthony (Toto) van Inge, for his dedicated support, insight, and constructive feedback. Second, I am thankful for the development of Prosogram, by Piet Mertens, which enabled the extraction of the prosodic features in a relatively short time frame. Lastly, credits should be provided to TED, the owner of the utilized talks, for the provision of the open distribution of their high quality recorded public speaking performances under the Creative Commons License.
References


methods in natural language processing (pp. 208–215).
derived improvements in the treatment of social phobia. Journal of Consulting and
Clinical Psychology, 77(2), 317.
charismatic speech. In Interspeech (pp. 513–516).
Silverstein, D., & Zhang, T. (2003, October 30). System and method of providing evalua-
tion feedback to a speaker while giving a real-time oral presentation. Google Patents.
App. 10/132,980)
Treasure, J. (2017). How to speak so that people want to listen. Retrieved from
https://www.ted.com/talks/julian_treasure_how_to_speak_so_that_people_want_to_listen
([Online; accessed June 14, 2017])
2013 12th international conference on information technology based higher educa-
tion and training (ithet) (p. 1-5). doi: 10.1109/ITHET.2013.6671050