Investigating Textual Features in Automatic Real-time Feedback on Public Speaking Performance

Bart Laan
10812407

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligentie

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
dhr. drs. A. (Toto) van Inge

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 904
1090 GH Amsterdam

June 29th, 2018
Summary

Public speaking skills are important in numerous professions and contexts, and thus effectively evaluating them with the aim of improvement is often sought after. Assessment by experts has shown to be inconsistent and time-inefficient, spawning this project to pursue improvement on automated solutions by exploring what textual features have to offer. Several analyses were performed to investigate the relations between textual features and presentation ratings, revealing that most of them have a moderate correlation. An additional inspection in how these features compare to features extracted from real-time speech recognition software revealed that not all of them work well in this context. It is suggested to use the remaining features in combination with techniques shown to perform well in other studies to potentially create the best-functioning automated real-time public speaking feedback system. Finally, a prototype was made using the results to illustrate the ability for systems making use of speech recognition to perform in real time.
## Contents

Summary .................................................. 2
Acronyms ................................................. 4

1 Introduction ............................................ 5
   1.1 The Issue of Evaluating Public Speaking Performance .... 5
   1.2 Theoretical Background .............................. 5
   1.3 Investigating Textual Features ...................... 6

2 Research Method ...................................... 8
   2.1 Finding a Suitable Dataset ......................... 8
   2.2 Data Preprocessing .................................. 8
   2.3 Feature Extraction .................................. 10
   2.4 Experiments and Evaluation ......................... 11

3 Results ............................................... 14
   3.1 Feature Constancy Between Same Speaker Performances ... 14
   3.2 Feature-Rating Correlations ......................... 15
   3.3 Using Textual Features to Derive Polynomials ........... 16
   3.4 Comparing Manual and Automatic Transcriptions ........ 16
   3.5 Using Experiment Results in a Prototype Feedback System ... 18

4 Discussion and conclusion ......................... 20
   4.1 Discussion and Future Work ......................... 20
   4.2 Conclusion ......................................... 21

References ............................................. 22

5 Appendix .............................................. 24
**Acronyms**

ASC  Average syllable count. 9, 13–17, 23

ASWC  Average stopword count. 10, 13–15, 17, 25

CV  Coefficient of variation. 10, 13, 14

QWU  Quantitative word use. 9, 13–17, 19

SPM  Syllables per minute. 9, 13–17, 19, 25

WL  Word length. 9, 13–17, 23, 24

WO  Word obscurity. 10, 13, 14, 17, 19, 24, 25

WPM  Words per minute. 9, 13–15, 17, 26
1 Introduction

1.1 The Issue of Evaluating Public Speaking Performance

Public speaking is encountered in a large number of varying places and contexts. Examples include teaching in different stages and disciplines of education, performing presentations in a professional environment, or providing entertainment in stand-up comedy. Therefore, besides being an important communication device, performing well in public speaking contexts is seen as an important professional skill for numerous careers (Kyllonen, 2012). Consequently, improving - and by extension, assessing - these skills has been a valuable matter, and finding out how to do this effectively has been pursued for a long time (Fawcett and Miller, 1975; James, 1976). As a result, the improvement of presentation skills has become a part of the majority of modern education standards (Morley, 2001; Jackson and Ward, 2014).

In the past, evaluating the performance of students in a presentation context has been a challenging task, especially on a larger scale. To start, what exactly makes a good presentation is subjective and thus not always agreed upon, even among experts in the field (Ward, 2013). This naturally poses a problem, as one presentation could receive varying and inconsistent feedback from different teachers. Additionally, evaluating presentations is typically a particularly time-consuming task as students often have to present one at a time, with the alternative being to perform in front of different experts, which, again, has the problem of lacking feedback consistency. Hence, a consistent and automatic way of accurately assessing public speaking performance is desired. Addressing the issue of subjectivity, the idea of evaluating public speaking with respect to certain presentation types instead of a single golden standard could pose a solution, allowing for training according to varying standards while still offering objectivity.

Attempts to create standardized evaluation rubrics for public speaking have included language use (Schreiber et al., 2012), and combining this with the fact that over the past few years speech recognition features have seen a rise in availability and quality (Sak et al., 2015; Amodei et al., 2016; Chiu et al., 2017), this project was issued to investigate ways to aid the efforts in creating an automated feedback system for public speaking. Specifically, the question that will be researched is: how can textual information derived from presentations aid in automatically generating feedback in real-time?

1.2 Theoretical Background

As one might expect, the idea of automatically evaluating presentations is not new, and numerous efforts at doing so have been made in the past. For instance, an analysis was performed (Hincks, 2005) on the use of prosodic features with respect to the perceived liveliness of public speakers. These features concern aspects of speech that do not depend on the words themselves, such as pitch, intonation and pacing. It was hypothesized that the liveliness of speakers could
be measured by looking at the pitch variance of the speaker performances. A new metric dubbed pitch variation quotient was established to effectively measure the pitch variance, which can be derived by taking the pitch variance as a percentage of the mean pitch. Results from this study point towards a strong correlation between perceived speaker liveliness and the pitch variation quotient, a promising find for automatic feedback systems.

In a more recent study, an automated feedback system on nonverbal aspects of public speaking was implemented with some success, showing a high correlation of the analyzed features with expert evaluations (Wörtwein et al., 2015). The specific aspects that were analyzed included things like eye gaze, gesturing and voice quality, and the feedback was implemented as a virtual audience that reacts to these audiovisual cues.

Nuchelmans (2016) attempted to automatically evaluate speakers in real time using audio, which was done by deriving different features from speech data of TED Talks\(^1\) including pitch variance, average pitch height and speech tempo. These were evaluated with the ratings that were given to the speakers of which the data was taken from and once again, the primary discerning factor was pitch variance. The thesis also notably commented on the absence of textual features due to lack of a real-time transcription.

Building on this thesis, a study from the year after investigated the advantage of using prosodic features in addition to find out if a vector could be made from the data of speakers that could help characterize different types of speakers (Bos, 2017). This was done by measuring audio pitch range and pitch profiles of a speaker, and it was shown that they can discern between similar and different types of speakers.

1.3 Investigating Textual Features

It is possible to do real-time speech recognition on public speaking performances using a number of different speech-to-text APIs, such as Google Cloud\(^2\) and CMU Sphinx\(^3\), to generate computer-generated transcriptions. This means that textual features can be derived from speech performances, enabling an investigation into the possibility of using these features to create a real-time public speaking feedback system.

Approaching this issue, a number of different ways exist in which transcriptions could give insight into the quality of a presentation. For instance, talking speed could influence how involved the audience is in the presentation: they might get disinterested if the host speaks too slow, or might not be able to follow the narrative if the talking speed is too fast. The choice of words is another factor that might influence how well the audience can follow the presentation, as using words that the audience is not familiar with might limit their comprehension of the story. On the other hand, repeatedly using the same words or using

---

1.\(https://ted.com/talks\)
2.\(https://cloud.google.com/speech-to-text/\)
3.\(https://cmusphinx.github.io/wiki/\)
words with little meaning might irritate the listeners or make them disinterested, though this is admittedly a more subjective matter.

Although there are probably additional aspects that could be derived from presentation transcriptions, these are the factors this thesis will attempt to measure and evaluate. As previous research has suggested, aspects other than the mentioned textual features appear to be important factors in public speaking performances, which means that it is less likely for presentation quality to have a particularly strong connection with linguistic aspects of speech. It is therefore hypothesized that the aspects tested in this study have a mild to moderate, but not strong influence on the quality of presentations.

This thesis will first describe the method that is maintained to investigate the mentioned research question by detailing the dataset that is used, and how that data will then be analyzed and evaluated. Following up, the results of the concerning analyses are presented and explained, after which they will be evaluated in a discussion. Finally, the relation of these results to the main question and hypothesis is discussed in the conclusion.
2 Research Method

This chapter will describe how the research will be executed in such a way that it can be reproduced, which is approached by initially describing the selection and properties of a suitable dataset. Next, the preprocessing of the data as well as the selection and extraction of textual features are detailed, and lastly the setup and evaluation methods of the different experiments are discussed.

2.1 Finding a Suitable Dataset

To create a real-time public speaking feedback application, the usefulness of textual features will be analyzed and consequently, a dataset is required. TED (short for Technology, Entertainment and Design) is an annually held conference where dozens of speakers talk about their wildly varying areas of expertise, under the motto ‘ideas worth spreading’. Videos of these TED Talks are hosted online and are free for anyone to view, download and use under the Creative Commons License. Viewers can also vote for the videos on one or more rating categories from a set of fourteen. All TED Talks have an official manually transcribed text version, which are also freely available.

A dataset containing information about a selection of TED Talks including their ratings, metadata and transcriptions is available on Kaggle and will be used in this thesis, as it provides a moderately large dataset (n=2550) and includes the kind of data that is desired for this study (i.e. presentations including transcriptions and rating tags). Note however, that the hosts of Ted Talks are not arbitrarily selected and often have considerably good presentation skills, so the data is presumably biased towards more positive ratings. Moreover, perceived quality of a presentation is a subjective matter, and there is no information about the people that vote on TED Talks, or which of the videos each voter rated. That being said, the people that rate the talks are part of the target audience, arguably making the ratings meaningful regardless of their background. Still, the rating categories are not explained which leaves the interpretation of the rating’s meaning up to the voter, which is a potential cause for noise.

2.2 Data Preprocessing

Table 1 shows the fourteen rating categories that viewers of the TED Talks could vote on, and how their meaning was interpreted. As mentioned, the meaning of the rating types are not explained by TED, and are thus only interpreted as having either a positive, negative or neutral meaning.

In the dataset, the results of this voting process can be found for each TED Talk as the amount of votes each category received, which will be normalized in this study by the amount of views. Figure 1 depicts the distribution of these voting categories. Comparing this with table 1, it is apparent that the most popular

4https://www.kaggle.com/rounakbanik/ted-talks
Table 1: Interpretation of TED Talk rating categories

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beautiful</td>
<td>Confusing</td>
<td>OK</td>
</tr>
<tr>
<td>Courageous</td>
<td>Longwinded</td>
<td></td>
</tr>
<tr>
<td>Fascinating</td>
<td>Obnoxious</td>
<td></td>
</tr>
<tr>
<td>Funny</td>
<td>Unconvincing</td>
<td></td>
</tr>
<tr>
<td>Informative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingenious</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspiring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jaw-dropping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persuasive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Percentage of votes on different TED rating categories
voting categories are those with a positive interpretation, supporting the notion that the dataset has a bias towards more qualified presentation hosts.

To get an idea of the overall performance of the TED talk, the number of votes for each category will be added up to (for positive categories) or subtracted from (for negative categories) a total, which is subsequently normalized by the amount of views. For this rating and for individual rating categories, normalization by views was used instead of using the total amount of ratings because it is assumed that each viewer is equally likely to give a rating to a TED Talk, and that more exposure in itself does not necessitate that the talk was qualitatively better. This rating will be referred to as the ‘combined rating’ throughout this thesis, and should be a good indication of the overall quality of the TED Talks. Still, as we are also interested in the individual ratings with the prospect of using them as profiles, the textual features will additionally be compared to the original rating categories to find out if the features strongly relate to some of them in particular.

2.3 Feature Extraction

The broad aspects of speech that will be investigated are talking speed and choice of words, both discussed in section 1.3. These factors can potentially be measured using a number of different transcription features; for example, words per minute (WPM) measures how many words the speaker uses per minute, and could be indicative of the speaker’s talking speed. Another similarly defined measure of talking speed is the syllables per minute (SPM), which differs from WPM in that it could potentially more accurately depict the perceived talking speed, as the SPM additionally takes the word length in syllables into account. While deriving WPM is as straightforward as counting the words and dividing it by the duration in minutes, the determination of SPM is more complex as a method is required to count the number of syllables in a word. Oxforddictionaries.com defines a syllable as “A unit of pronunciation having one vowel sound”\(^5\), and using this definition, the amount of syllables can be determined by counting the number of vowel sounds. This will be put into practice with the help of the CMUDict pronouncing dictionary (included with the NLTK package for Python\(^6\)), as it provides information about phonemes and which of them are vowel sounds. To ensure compatibility with unknown words, CMUDict’s mean amount of syllables per letter is derived and subsequently multiplied by the length of unknown words to get an estimation of their syllable count.

Additional features can be derived to give insight into the word use of a presentation, such as average syllable count (ASC), measured as the mean amount of syllables per word. This could, together with word length (WL, measured as the average number of letters per word), indicate how long and complex the words are. Quantitative word use (QWU) indicates the mean amount of repeated instances of each word, normalized by the talk’s duration. This is meant to indicate the extensiveness of the speaker’s vocabulary, where lower values should represent greater word variation. Stopwords are described as certain

\(^5\)https://en.oxforddictionaries.com/definition/syllable
\(^6\)https://www.nltk.org/
highly common words that add little to a sentence semantically, and the average stopword count (ASWC) could therefore give an insight into the amount of words with little meaning surrounding the semantic core of the talk. The NLTK package\(^6\) includes a list of stopwords that is used here to determine the ASWC. To get an indication how uncommon the speaker’s words are, a feature will be calculated as the number of word occurrences in a text corpus normalized by the corpus’ mean word occurrence and then inverted. In this case, the talk transcriptions are used in addition to the Gutenberg corpus from NLTK to get the final text corpus from which this feature dubbed word obscurity (WO) is derived.

Considering that speech recognition techniques are not - and arguably never will be - perfect, a comparison between manually and automatically derived transcriptions is desired. As such, the Python SpeechRecognition\(^7\) wrapper module is used to interface with Google Speech Recognition in order to obtain transcriptions from a randomly selected subset of the TED dataset (n=108). The reason for the decision to use a subset (instead of the complete dataset) is that the time needed to automatically transcribe the whole dataset was deemed too long considering the scope of this project. A Python script was used to download the required audio files from the web pages of each talk, after which the files were converted to mono 16KHz WAV files to perform the speech-to-text conversion using the AudioSegment Python module\(^8\).

2.4 Experiments and Evaluation

Several experiments will be conducted to evaluate the usefulness of the textual features. First, the feature variance between different performances of the same speaker will be compared to the variance of all TED Talks. This is done by measuring the mean coefficient of variation (CV) of talks hosted by speakers with multiple data points in the dataset, and comparing that with the CV of the whole dataset. CV is computed by dividing the dataset’s standard deviation by its mean, making it immune to feature scaling and consequently enabling a more transparent data variance comparison. If the text feature works as expected, its CV between talks hosted by different speakers should be higher than between performances of the same speaker, as it is assumed that the performance of a speaker is relatively consistent. If the CV is not significantly higher or even lower, it means that the feature is probably not a good indicator of the speaker’s overall skill.

A second experiment will investigate the relations between features and different rating types of TED Talks. It is expected for most features that both their extremes (i.e. considerably low and high values) negatively influence the positive ratings of presentations, as it is difficult to imagine that, for example, ever decreasing/increasing talking speed or word length would always result in better ratings. However, it is presumed that in practice speakers have a tendency towards one side of the optimal range of features, which would result in a correlation between textual features and the resulting ratings. It might

\(^6\)https://github.com/UBERI/speech_recognition
\(^7\)https://github.com/jiaaro/pydub
\(^8\)https://github.com/jiaaro/pydub
therefore be interesting to evaluate these correlations by looking at their type (positive/negative), probability and strength. To measure these variables, the Pearson Product-moment Correlation Coefficient (commonly known and from here on referred to as the correlation coefficient) is initially calculated to find the correlation type and strength. Given the fact that the TED dataset has a sample size of 2434, the correlation coefficient then has to exceed .035 or be lower than -.035 for the correlation to be deemed significant, as the probability that the relation would not exist would otherwise exceed \( p = 0.05 \), which is the significance threshold maintained in this case.

A third experiment aims to find out if the features indeed have exactly one optimum, with the extremes of each feature both having a negative influence on the quality of presentations. To investigate this suspicion, second degree polynomials will be fitted to the different combinations of features and rating categories. Second degree polynomial functions with a single variable are of the form:

\[
f(x) = w_0 + w_1 x + w_2 x^2\]  

where \( x \) is the input variable and \( w_0, w_1 \) and \( w_2 \) depict the weights that characterize the specific function. This function type is used here for its property of having exactly one minimum or maximum, which can be found by computing \( \frac{-w_1}{2w_2} \). To get the function for a feature-rating combination with the smallest sum of squared errors to its data points, the textual feature will first be transformed to a set of features that fits the polynomial specification: \( x \rightarrow [1, x, x^2] \) where \( x \) represents the extracted textual feature. This new set of features is then used with the respective ratings as training data in the linear regression algorithm included in the scikit-learn\(^9\) toolkit for Python, which solves the least squares problem and outputs the produced weights. The plot of the resulting function is subsequently used to determine whether the relation is meaningful, and if the feature’s extremes are predicted to be negative or not. If the feature’s extremes behave in the expected manner, where they result in a negative prediction for positive typed ratings and vice-versa, the function’s optimum might then serve as the sweet spot for the concerning feature-rating combination.

In a fourth experiment, to verify that the speech recognition software works sufficiently well, the mean and standard deviation of the features derived from manual transcriptions will be compared to their counterparts computed from speech-to-text transcriptions. If the speech recognition works well, these values should not differ significantly.

Finally, if the features work correctly as specified by the preceding experiments, their results are used to create a prototype feedback system. This system will use either the combined rating or one of the rating categories to provide balanced feedback or feedback for a specific rating type, using a target value derived from the experiment results. How well this system would perform in helping students improve their presentation skills will not be tested, as that is outside the scope of this study. However, its ability to perform in real time will be evaluated.

\(^9\)scikit-learn.org
The discussed concepts were mapped into a diagram showing their relations to better illustrate the data flow and external influences present in the experiments. This diagram is shown in figure 2.

Figure 2: Data flow of the experiments
Table 2: Statistic attributes of TED Talk features between same speaker performances

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Stdev</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words per minute</td>
<td>11.253</td>
<td>0.081</td>
</tr>
<tr>
<td>Syllables per minute</td>
<td>15.360</td>
<td>0.075</td>
</tr>
<tr>
<td>Quantitative word use</td>
<td>0.075</td>
<td>0.247</td>
</tr>
<tr>
<td>Word length</td>
<td>0.109</td>
<td>0.025</td>
</tr>
<tr>
<td>Average syllable count</td>
<td>0.040</td>
<td>0.027</td>
</tr>
<tr>
<td>Word obscurity</td>
<td>0.149</td>
<td>0.221</td>
</tr>
<tr>
<td>Average stopword count</td>
<td>0.016</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 3: Statistic attributes of TED Talk features between different speakers

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Mean</th>
<th>Stdev</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words per minute</td>
<td>149.192</td>
<td>26.811</td>
<td>0.180</td>
</tr>
<tr>
<td>Syllables per minute</td>
<td>219.710</td>
<td>39.022</td>
<td>0.178</td>
</tr>
<tr>
<td>Quantitative word use</td>
<td>0.266</td>
<td>0.107</td>
<td>0.402</td>
</tr>
<tr>
<td>Word length</td>
<td>4.393</td>
<td>0.218</td>
<td>0.050</td>
</tr>
<tr>
<td>Average syllable count</td>
<td>1.475</td>
<td>0.080</td>
<td>0.054</td>
</tr>
<tr>
<td>Word obscurity</td>
<td>0.685</td>
<td>0.391</td>
<td>0.571</td>
</tr>
<tr>
<td>Average stopword count</td>
<td>0.510</td>
<td>0.029</td>
<td>0.056</td>
</tr>
</tbody>
</table>

3 Results

This chapter will show and describe the results from the different experiments discussed in chapter 2. Note that it was found that a number of entries substantially deviated from the norm while inspecting the dataset. In particular, there were some data points that wholly consisted of musical performances, which are not relevant for this study and were therefore removed from the dataset. In addition, 86 entries did not have a transcript and were dropped as a result. What is left is a dataset with a sample size of 2434, which should still be more than sufficiently large for this study.

3.1 Feature Constancy Between Same Speaker Performances

266 out of the total of 2088 speakers in the dataset had multiple TED Talk entries and from these talks, the textual features and their standard deviation in addition to CV were extracted. Table 2 shows the standard deviation (seen as Stdev) and CV of the concerning talks. Comparing these with table 3, which shows these measures in addition to the mean for all TED Talks in the dataset, it is apparent that the CV for each feature is significantly lower for talks of the same speaker compared to the CV for the whole dataset, with the difference
Table 4: Correlation coefficients of textual features with rating categories. Bold entries mark correlations strong enough to be considered significant.

<table>
<thead>
<tr>
<th>Rating name</th>
<th>WPM</th>
<th>SPM</th>
<th>QWU</th>
<th>WL</th>
<th>ASC</th>
<th>WO</th>
<th>ASWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>0.089</td>
<td>0.056</td>
<td>-0.111</td>
<td>-0.100</td>
<td>-0.112</td>
<td>-0.073</td>
<td>0.085</td>
</tr>
<tr>
<td>Beautiful</td>
<td>-0.143</td>
<td>-0.190</td>
<td>-0.002</td>
<td>-0.149</td>
<td>-0.164</td>
<td>0.064</td>
<td>0.058</td>
</tr>
<tr>
<td>Confusing</td>
<td>0.070</td>
<td>0.068</td>
<td>-0.002</td>
<td>-0.035</td>
<td>-0.008</td>
<td>-0.003</td>
<td>0.085</td>
</tr>
<tr>
<td>Courageous</td>
<td>-0.081</td>
<td>-0.118</td>
<td>-0.139</td>
<td>-0.136</td>
<td>-0.123</td>
<td>-0.041</td>
<td>0.096</td>
</tr>
<tr>
<td>Fascinating</td>
<td>0.191</td>
<td>0.190</td>
<td>-0.099</td>
<td>-0.003</td>
<td>-0.006</td>
<td>-0.035</td>
<td>0.106</td>
</tr>
<tr>
<td>Funny</td>
<td>0.030</td>
<td>-0.029</td>
<td>0.074</td>
<td>-0.216</td>
<td>-0.199</td>
<td>0.101</td>
<td>-0.030</td>
</tr>
<tr>
<td>Informative</td>
<td>0.280</td>
<td>0.346</td>
<td>-0.090</td>
<td>0.248</td>
<td>0.218</td>
<td>-0.139</td>
<td>-0.046</td>
</tr>
<tr>
<td>Ingenious</td>
<td>0.149</td>
<td>0.145</td>
<td>0.128</td>
<td>-0.037</td>
<td>-0.017</td>
<td>-0.032</td>
<td>0.042</td>
</tr>
<tr>
<td>Inspiring</td>
<td>-0.006</td>
<td>-0.051</td>
<td>-0.144</td>
<td>-0.138</td>
<td>-0.148</td>
<td>-0.108</td>
<td>0.117</td>
</tr>
<tr>
<td>Jaw-dropping</td>
<td>-0.006</td>
<td>-0.020</td>
<td>-0.050</td>
<td>-0.073</td>
<td>-0.055</td>
<td>0.021</td>
<td>0.053</td>
</tr>
<tr>
<td>Longwinded</td>
<td>0.076</td>
<td>0.068</td>
<td>-0.221</td>
<td>-0.047</td>
<td>-0.027</td>
<td>-0.039</td>
<td>0.150</td>
</tr>
<tr>
<td>Obnoxious</td>
<td>0.003</td>
<td>-0.008</td>
<td>-0.222</td>
<td>-0.056</td>
<td>-0.039</td>
<td>0.037</td>
<td>0.024</td>
</tr>
<tr>
<td>OK</td>
<td>0.062</td>
<td>0.046</td>
<td>0.235</td>
<td>-0.067</td>
<td>-0.057</td>
<td>-0.028</td>
<td>0.045</td>
</tr>
<tr>
<td>Persuasive</td>
<td>0.138</td>
<td>0.163</td>
<td>-0.144</td>
<td>0.100</td>
<td>0.086</td>
<td>-0.137</td>
<td>0.007</td>
</tr>
<tr>
<td>Unconvincing</td>
<td>0.026</td>
<td>0.039</td>
<td>0.030</td>
<td>0.019</td>
<td>0.039</td>
<td>-0.039</td>
<td>0.027</td>
</tr>
</tbody>
</table>

ranging from 38% (QWU) to 61% (WO). This seems to suggest that all features work reasonably well to discern between unique speakers, which is a good sign for their reliability. Another minor thing to note is that the pairs WPM/SPM and ASC/WL have highly comparable CV values, which should not be surprising as they concern similar metrics.

3.2 Feature-Rating Correlations

The aim of experiment 2 was to investigate the correlations between textual features and rating categories. As such, the correlation coefficients between features and the different ratings - including the combined rating - were derived, and subsequently entered into table 4, its bold entries marking correlation values that are more significant than $r = .035$. As predicted, no correlations are deemed considerably strong, as none of the correlation coefficients are stronger than $r = .500$. In addition, most feature-rating combinations can be considered significant, and a good number of them have a mild to moderate correlation, interpreted as values more significant than $r = .100$.

Interestingly, ratings like “informative” and “beautiful” seem to be explained better by the different features than some other categories. This phenomenon appears to be more prominent for the more common rating categories (see figure 1). Some entries can intuitively be interpreted, such as that “informative” talks commonly use longer words (WL) to effectively communicate their ideas, or that “beautiful” talks often have a lower talking speed (SPM), but the table also shows some counterintuitive correlations. For example, one might expect a positive correlation between the “longwinded” rating and the QWU feature as
repeated use of the same words could be perceived negatively, but the correlation turns out to be negative instead. Furthermore, the correlation types (i.e. positive/negative) for the WPM and SPM features notably vary within ratings of the same class; for instance, “fascinating” and “ingenious” correlate positively with these features, while features like “beautiful” and “courageous” correlate negatively. This finding seems to support the notion that different types of talks are characterized in different ways by the aspects of speech.

It was theorized in section 2.4 that these correlations could be the result of that, when outside of the feature’s optimal range, most speakers tend towards one side of the feature’s mean. Figure 3c, which shows the TED data points as blue dots, seems to support this conjecture. It shows that data points with a SPM measure below the mean value of 220 (seen in 3) on average have lower “informative” ratings, while SPM values higher than 220 have a higher rating on average. Similarly, figure 3d shows a mirrored effect, where higher QWU values have a lower average “inspiring” rating. As one would expect, these phenomena are also supported by their correlation values from table 4. In practice, this could be used by recommending a change in the direction of the correlation when on this suboptimal side of the mean, or not recommending anything when on the other side.

### 3.3 Using Textual Features to Derive Polynomials

The polynomial functions were derived as described in chapter 2 from a total of \(7 \times 15 = 105\) feature-rating combinations. Figure 3 depicts some of those results, where the blue dots represent the data points and the green lines are the plotted polynomial functions. Section 2.4 explained that the functions were expected to have maxima for positive ratings and minima for negative ratings, with the optima existing near the highest ratings. As the example from figure 3a shows, some functions are indeed of the expected quality. However, the large majority of functions either resulted in a flat line such as the plot in figure 3c, or even result in an inverted function like in figure 3d where the feature’s extremes result in positive instead of negative rating predictions. Interestingly, the functions that behave like expected largely come from the features that measure the word length, i.e. WL and ASC. In total, only 21 of the 105 feature-ratings combinations had the expected function properties, which means that using the optima as target values is not an option for most of them. Still, it appears these 21 optima would prove useful targets for use in an automated feedback system. The plots of the 21 optima that satisfy the described specifications can be found in the appendix, section 5.

### 3.4 Comparing Manual and Automatic Transcriptions

As discussed, the reliability of the used speech-to-text software is evaluated by looking at the difference between manual and speech-to-text transcriptions by comparing their mean and standard deviation, and the results of this comparison can be seen in tables 5 and 6 respectively. Looking at the percentual differences,
Figure 3: Polynomial plots between features and ratings

(a) ASC vs. Fascinating rating

(b) WL vs. combined rating

(c) SPM vs. Informative rating

(d) QWU vs. Inspiring rating
Table 5: Percentual difference of feature means between manual and speech-to-text transcriptions

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Mean (manual)</th>
<th>Mean (speech-to-text)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPM</td>
<td>149.192</td>
<td>134.707</td>
<td>-9.71%</td>
</tr>
<tr>
<td>SPM</td>
<td>219.710</td>
<td>191.961</td>
<td>-12.63%</td>
</tr>
<tr>
<td>QWU</td>
<td>0.266</td>
<td>0.178</td>
<td>-33.29%</td>
</tr>
<tr>
<td>WL</td>
<td>4.393</td>
<td>4.286</td>
<td>-2.43%</td>
</tr>
<tr>
<td>ASC</td>
<td>1.475</td>
<td>1.424</td>
<td>-3.44%</td>
</tr>
<tr>
<td>WO</td>
<td>0.989</td>
<td>0.901</td>
<td>-8.92%</td>
</tr>
<tr>
<td>ASWC</td>
<td>0.510</td>
<td>0.510</td>
<td>+0.03%</td>
</tr>
</tbody>
</table>

Table 6: Percentual difference of standard deviations between features from manual and speech-to-text transcriptions

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Stdev (manual)</th>
<th>Stdev (speech-to-text)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPM</td>
<td>26.811</td>
<td>28.248</td>
<td>+5.36%</td>
</tr>
<tr>
<td>SPM</td>
<td>39.022</td>
<td>41.669</td>
<td>+6.78%</td>
</tr>
<tr>
<td>QWU</td>
<td>0.107</td>
<td>0.836</td>
<td>-22.03%</td>
</tr>
<tr>
<td>WL</td>
<td>0.218</td>
<td>0.179</td>
<td>-17.59%</td>
</tr>
<tr>
<td>ASC</td>
<td>0.080</td>
<td>0.064</td>
<td>-19.43%</td>
</tr>
<tr>
<td>WO</td>
<td>0.571</td>
<td>0.250</td>
<td>-56.30%</td>
</tr>
<tr>
<td>ASWC</td>
<td>0.029</td>
<td>0.027</td>
<td>-5.02%</td>
</tr>
</tbody>
</table>

only the ASWC feature seems to stay relatively consistent, with WL and ASC as possible contenders. The largest difference is seen in the WO feature, with over a 50% decrease in standard deviation. A possible explanation for this could be that the speech recognition software itself has a limited database of words, and that particularly uncommon words are therefore either ignored or get recognized as other, more common words instead.

3.5 Using Experiment Results in a Prototype Feedback System

Following the described results, a prototype feedback system was implemented by determining the two most relevant features for a selectable rating category, and extracting them from a transcription generated in real time using Google Speech Recognition. Then, a change to a feature is recommended in the direction of a polynomial optimum if the feature-rating combination is included in one of the 21 usable optima, and in the direction of a significant correlation otherwise. However, as discussed in section 3.2, when using the correlation a feature change will only be recommended when its value is on the suboptimal side of the mean. Furthermore, the ASC, QWU and WO features were omitted as options as they were deemed too unreliable considering the results from section 3.4.
In an early version of the prototype, the performance analysis could not be performed in real time as the time needed by the speech recognition software to extract a transcription varied roughly between 0.5 and 2.5 times the length of the audio buffer. To solve this issue, the prototype was enhanced with multiple threads to enable the speech recognition to take as long as $t$ times the buffer length, where $t$ is equal to the number of threads used. In this experiment, using three threads was enough to run the prototype without problems. Figure 4 shows what the prototype looks like in practice. In this screenshot, the blue region indicates the feature’s value measured at that point, while the red, orange and green regions specify whether or not a change in the respective feature is suggested (see the figure’s legend for further details). The sizes of these ranges are determined using the standard deviations found in table 3.
4 Discussion and conclusion

4.1 Discussion and Future Work

By looking at the plots from the third experiment (section 3.3), it might be concluded that other functions like the gaussian function would have fit the data better, as the tails of the data often seem to flatten out comparable to a gaussian function, rather than quickly increase or decrease like a polynomial. This mismatch could help explain why so few of the fitted functions had the expected properties. Moving on, regarding the performance of the used speech-to-text algorithm: the ASC, QWU and WO measures all differed too greatly between the manual and automatic transcriptions to be considered reliable for use in the automated feedback system. Still, the results of these features extracted from manual transcriptions showed potential, and with the prospect of the industry’s steady growth in speech recognition performance, they might perform better with automatic transcriptions in the future. Additionally, the speech recognition software that was used in this project is only one example, and the effect of other implementations could be looked into as well.

Although the features derived from transcriptions do not provide enough information to provide a complete evaluation, they might still play an important role if used in conjunction with other implementations, such as the ones discussed in section 1.2. In addition, though the talks were analyzed as a whole in this study, there is a possibility that looking at different stages of presentations (i.e. introduction, main body, conclusion) might provide additional information about how to perform effectively in a public speaking context (Jackson and Ward, 2014), and the effect of this in practice could be researched in future studies. Further ways of processing the data can also be explored, such as performing data clustering on speaker and/or rating types, which is not limited to transcription analyses. Additional transcription features such as sentence length could be derived from the manual transcriptions, but, due to their absence of sentence markers, not as easily from the automatic transcriptions. This could however be implemented by recognizing pauses as sentence markers.

Another potential that features extracted from automatic transcriptions offer is the ability to predict ratings using all transcription features, by using neural networks for example. This was purposely not done here as the aim was to provide feedback to the user, and when combining the features in an algorithm it becomes difficult to track down what aspects played which role in determining a score, and thus what aspects the speaker should try to change. Still, predicting ratings using transcriptions could prove valuable in other contexts. Lastly, it was not tested whether the feedback system has any influence on public speaking performance in practice were not tested, which is strongly recommended before using any implementation of these systems.
4.2 Conclusion

To sum up, the relations of seven textual features were analyzed with presentation ratings in order to evaluate their usefulness in an automated feedback system for public speaking. This was done in a couple of experiments, looking at the features’ ability to define unique speakers, what kind of relation the features have with presentation ratings, and how well the features work with speech recognition techniques. Results indicate that all seven features work reasonably well to define unique speakers, and some of them evaluate the presentation qualities well enough to give feedback. However, a portion of them differed too greatly between automatic and manual transcriptions and can therefore not be recommended at this time for real-time feedback applications. Finally, it was shown that a prototype feedback system was able to perform in real time, showing that feedback using transcriptions is not limited in that regard.
References


5 Appendix

Figure 5: Selection of polynomial plots between features and ratings

(a) ASC vs. Combined rating  
(b) ASC vs. Fascinating rating

(c) ASC vs. Informative rating  
(d) ASC vs. Ingenious rating

(e) ASC vs. Persuasive rating  
(f) WL vs. combined rating
(a) WL vs. Confusing rating

(b) WL vs. Fascinating rating

(c) WL vs. Informative rating

(d) WL vs. Ingenious rating

(e) WL vs. Persuasive rating

(f) WO vs. Fascinating rating
(a) WO vs. Funny rating  
(b) SPM vs. Funny rating  
(c) SPM vs. Inspiring rating  
(d) ASWC vs. Informative rating  
(e) ASWC vs. Ingenious rating  
(f) ASWC vs. Persuasive rating
(a) WPM vs. Confusing rating

(b) WPM vs. Courageous rating

(c) WPM vs. Longwinded rating