Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Perception and prediction of speaker appeal – A single speaker study☆

Q1 Ailbhe Cullen[a,*], Andrew Hines[b], Naomi Harte[a]

Q2 [a] Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
[b] School of Computer Science, University College Dublin, Ireland

Received 23 December 2017; received in revised form 23 February 2018; accepted 14 April 2018
Available online xxx

Abstract

Q3 In this paper we explore the automatic prediction of speaker appeal from recordings of political speech. The database used contains recordings of a single speaker in a wide range of situations (interview, election ally et.) which has been annotated for six speaker traits: boring; charismatic; enthusiastic; inspiring; likeable; and persuasive. The aim of this study is to predict these ratings using acoustic features of the speech. We offer three key contributions in this paper. Firstly, we explore the effect of acoustic environment on the perception of speaker ability. We find significant biases in the perception of all six traits, with interview speech being consistently rated as less appealing, and election rally speech as more appealing. In our second contribution, we attempt to exploit this bias by modelling speech from each situation separately, which gives a significant improvement in classification performance. Finally, the database covers 7 years. Thus, our third contribution is an analysis of the variance in both annotations and acoustic features over time to uncover temporal trends in speaker appeal. We find significant trends which show a decline in the speaker's prosodic activity over time, which mirror a decline in the perception of speaker appeal as measured by the database annotations.

Keywords: Computational paralinguistics; Speaker trait detection; Speaker appeal; Political speech; Speech processing

1. Introduction

A politican's ability to speak well in a range of situations contributes significantly to their success or failure (Gregory and Gallagher, 2002; Slatcher et al., 2007). However, what exactly makes a good speaker remains an open question (Finlayson and Martin, 2008; Rosenberg and Hirschberg, 2009; Signorello et al., 2012). Studies in this area have explored both the content (Hart and Lind, 2010; Kaplan and Rosenberg, 2012; Pennebaker and Lay, 2002) and delivery of speech (Finlayson and Martin, 2008; Rosenberg and Hirschberg, 2009; Signorello et al., 2012). However, Q4 such studies typically focus on a single speech setting, for example the stump speech (Hart and Lind, 2010), the conference address (Finlayson and Martin, 2008; Heritage and Greatbatch, 1986), or the parliamentary debate (Strangert

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: cullena3@tcd.ie (A. Cullen).

https://doi.org/10.1016/j.csl.2018.04.004

and Gustafson, 2008). Furthermore, they tend to focus on short periods of time, such as a single address (Finlayson and Martin, 2008) or a particular election campaign (Rosenberg and Hirschberg, 2009; Hart and Lind, 2010).

In this study, our interest is in longitudinal and environmental variations in speaker appeal. Thus, we choose to study a single speaker, which allows us to focus on sources of variability other than inter-speaker variations. Paralinguistic traits are known to be complex and subjective (Schuller and Batliner, 2014; D'Errico et al., 2013). As such, a number of authors attempt to simplify the problem by studying a single speaker (Pennebaker and Lay, 2002; D'Errico et al., 2013; Touati, 1993; Niebuhr et al., 2016). In this study, we analyse the acoustic behaviour of a single speaker, former Irish Prime Minister Mr. Enda Kenny, using recordings which cover a 7 year period, and a range of recording and speaking environments (e.g. interview, parliamentary debate). In this political context there is an added benefit to the single speaker study, as it reduces the effect of biases caused by the political affiliation of speakers and annotators (Rosenberg and Hirschberg, 2009; Weiss, 2005; Cullen and Harte, 2017). In this study we explore the effect of differing situation and motivation of speech on both the perception of speaker appeal, and our ability to automatically predict speaker traits. This is different to previous studies which focus on a single mode of interaction, either the podium speech (Kaplan and Rosenberg, 2012; Weninger et al., 2012) or the debate (Strangert and Gustafson, 2008; Scherer et al., 2012; Kim et al., 2014). We also discuss the variation in both the perception of speaker appeal, and the acoustic behaviour of the speaker over the 7 years of the database. This covers the speaker's time as leader of the opposition, and his first two years as head of government. While previous studies have examined changes in linguistic behaviour over time (Pennebaker and Lay, 2002), or changes in the voice before and after a significant event (Signorello et al., 2012; Touati, 1993), we are not aware of any other study which attempts to track changes in a speaker's vocal behaviour over such a long period.

We focus on the delivery aspect of speech, building on the existing body of work linking speaker ability and leadership with the vocal behaviour of the speaker (Rosenberg and Hirschberg, 2009; Weninger et al., 2012; Wörtwein et al., 2015; Burkhardt et al., 2011). A variety of acoustic features are extracted, and are used to predict annotations of six attributes related to speaker appeal: boredom; charisma; enthusiasm; inspiration; persuasion; and likeability. We have previously reported results on this database using a standard set of spectral and prosodic features (Cullen and Harte, 2017). However, the wide range of recording environments encountered in the database, leads to significant variation in the acoustic features. Thus, in this paper we explore the use of noise robust features for classification. Using Mel Frequency Cepstral Coefficients (MFCCs) as our baseline we implement two methods of noise compensation. The first is a blind speech de-reverberation, which is intended to normalise the diverse acoustic environments encountered in the database. RASTA filtered MFCCs are also used, as they have shown superior noise robustness for a range of speech processing tasks (Kockmann et al., 2011; Chia-Ping et al., 2005). We find that while RASTA filtering is beneficial, de-reverberation has a negative impact on classifier performance, suggesting that the acoustic environment (i.e. the setting of the speech) influences perception of speaker appeal.

Motivated by these findings we explore the effect of recording situation on the perception and prediction of speaker appeal. Speaking style varies depending on the motivation and setting of the speech (Slatcher et al., 2007; Rosenberg and Hirschberg, 2009; Pennebaker and Lay, 2002), and also with the acoustic environment (Astolfi et al., 2015). For example, election rally speech tends to be more emotive and energetic than parliamentary speech. We find that this variation in speaking style induces a bias in the perception of speaker appeal depending on the situation, with interview speech being consistently rated lass appealing and election rally speech being rated more appealing. We exploit these biases by developing situation specific models of appeal in order to improve overall classification performance.

Finally, we discuss changes in the speaker's prosodic behaviour and the ratings of speaker appeal over the span of the database. The database was annotated at a sentence level, out of chronological order. The majority of raters were American, and therefore unlikely to be familiar with the speaker. We observe a decline in ratings of charisma, enthusiasm, and Overall Speaker Appeal over the span of the database, and find that this downward trend in appeal is accompanied by lower fundamental frequency (F0) of the speaker, and reduced F0 range and variance over time.

The remainder of the paper is structured as follows. Section 2 discusses related work in the prediction of paralinguistic traits and the analysis of political speech in particular. We briefly introduce the Irish Political Speech database used in this study in Section 3. Section 4 outlines the features and classifiers used in our prediction experiments. We compare the performance of the noise-robust features on this database in Section 5. Following this, in Section 6 we explore in more depth the effect of situation on the perception and prediction of speaker appeal. Finally, in Section 7 we explore the variation in both the perception and production of speaker appeal over time. Conclusions are drawn in Section 8.

Please cite this article as: A. Cullen et al., Perception and prediction of speaker appeal − A single speaker study, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.04.004