Building a Database of Political Speech

Does culture matter in charisma annotations?

Ailbhe Cullen
Dept. of Electronic and
Electrical Engineering
Trinity College Dublin
Ireland
cullena3@tcd.ie

Andrew Hines
Dept. of Electronic and
Electrical Engineering
Trinity College Dublin
Ireland
hinesa@tcd.ie

Naomi Harte
Dept. of Electronic and
Electrical Engineering
Trinity College Dublin
Ireland
nharte@tcd.ie

ABSTRACT

For both individual politicians and political parties the internet has become a vital tool for self-promotion and the distribution of ideas. The rise of streaming has enabled political debates and speeches to reach global audiences. In this paper, we explore the nature of charisma in political speech, with a view to automatic detection. To this end, we have collected a new database of political speech from YouTube and other on-line resources. Annotation is performed both by native listeners, and Amazon Mechanical Turk (AMT) workers. Detailed analysis shows that both label sets are equally reliable. The results support the use of crowd-sourced labels for speaker traits such as charisma in political speech, even where cultural subtleties are present. The impact of these different annotations on charisma prediction from political speech is also investigated.

Categories and Subject Descriptors

H.2.m [**Database Management**]: Miscellaneous—*paralinguistic database design*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language parsing and understanding*

General Terms

Data Collection; Speech Processing

Keywords

Computational Paralinguistics; Charisma Detection

1. INTRODUCTION

In recent years there has been growing interest in both charismatic and political speech. It has been shown that voting behaviour can be reliably predicted both from the candidates' voice [7], and from their appearance [12]. A number of studies have explored prosodic patterns in political speech [14, 5, 17]. Rosenberg et al. [14] find significant correlations between charisma and a number of prosodic features. Weninger et al. [18] attempt to predict charisma from the speech of proven leaders, while Kim et al. [11] detect conflict in political debates.

There are few existing annotated databases of political speech. The most relevant is that of Rosenberg et al. [14], which consists of speech from the nine Democratic candidates for the 2004 American presidential election. This contains 45 speech samples, five from each speaker. A larger database is the Canal9 corpus [11], which consists of televised debates between politicians. This is a more dynamic database, intended for the study of social dynamics, but is annotated for conflict, not charisma. The CORPS database, collected by Strapparava et al. [16], is a text only database containing transcripts of political speeches, and tags for audience reactions.

This paper presents a new audio-visual database of political speech. This database has been collated from a variety of on-line sources, with no control over the recording procedure, and as such is more representative of conditions which must be tolerated by a real-world content retrieval system. This database is novel in that it spans seven years (2006 to 2012), enabling future longitudinal studies of how the perception of charisma changes over time. Furthermore, it is much larger than the database collected by Rosenberg et al. [14], thus making it more suitable for machine learning tasks. The recordings contain speech from a wide range of settings, from parliamentary addresses to election rallies. It is expected that both speaking style and perceived charisma will vary depending on the motivation and setting of the speech [13].

Traditionally, paralinguistic databases have been annotated by groups of trusted, or "expert" labellers, as in [18]. Recently it has become popular to crowd source annotations, for example via Amazon Mechanical Turk (AMT) [6, 11]. Previous work has shown consistency between AMT and expert labellers [15]. In this paper we question how appropriate AMT annotation is when the content of the database may be perceived as culturally nuanced. There is conflicting evidence for the influence of culture on charisma perception. Biadsy et al. [3] report no significant difference in the perception of charisma by raters from America, Palestine, and Sweden. However, D'Errico et al. [5] demonstrate a difference in the perception of French and Italian listeners. Though the majority of raters in this work are native

AVEC'14, November 7, 2014, Orlando, FL, USA.

ACM 978-1-4503-3119-7/14/11...$15.00.
http://dx.doi.org/10.1145/2661806.2661808.

English speakers, a cultural divide still exists between the American and Irish raters.

The audio portion of the database is labelled for four attributes: charisma, likeability, inspiration, and enthusiasm. Annotation is conducted twice, once by a cohort of Irish residents who are recruited from within our university, and once using Amazon Mechanical Turk (AMT). Our comparison of labels provided by Irish or "native" residents and American AMT workers shows a strong agreement in the perception of charisma, enthusiasm, and inspiration, but differing perception of likeability. We suggest that this is due to the native raters' familiarity with the speaker, rather than a systematic cultural difference.

Finally, we report results of some initial charisma detection experiments, using only the audio modality. Regression is performed using a multi layer perceptron (MLP), and results are compared for prediction of native and AMT labels. These results have important implications, supporting the use of cross cultural ratings of charisma in the future.

The remainder of the paper is structured as follows. Section 2 introduces the database. Section 3 discusses the annotation procedure. Section 4 outlines the experimental set up. Section 5 discusses the results. Conclusions are discussed in section 6.

2. IRISH POLITICAL SPEECH DATABASE

This is an English language database, containing videos of a single speaker, an Taoiseach[1] Enda Kenny, in six settings: parliamentary addresses; public messages; interviews; internal party conferences; election rallies; and other (containing mostly opening speeches from an assortment of public functions). It is expected that some degree of charisma is required in order to rise to the highest position in government. Furthermore, Mr. Kenny is known to have undergone media training over the time span of the database. A total of 1456 sentences, from 47 recordings are included in the database. These range in date from 2006 to 2012, and have been chosen to obtain a balance between recordings from Mr Kenny's time in opposition (2006 - Mar 2011) and government (Mar 2011 - present), and between the six aforementioned settings.

The database was collected from the on-line archives of the Irish parliament [2], YouTube uploads from and/or featuring Irish politicians (e.g. [1]), and television interviews. A key factor when using data from such diverse sources is sound quality. The audio sampling rate varies from 22kHz to 48kHz. While the standard of some studio interviews is quite high, many recordings contain some noise. Preprocessing of the audio was carried out using Adobe Audition CS6 (Adobe Systems, CA, USA). The default voice activity detection was used to segment the recordings to a phrase level, this was then manually corrected to whole sentences. Isolated noises (coughs, bangs, interruptions, etc.) and Irish language utterances were manually segmented. Finally, the loudness was equalised for all clips.

3. ANNOTATION

We report here results from the first two phases of annotation. First a subset of 60 sentences were labelled by Irish residents, which are chosen to be balanced across the six settings (parliamentary address, public message, interview, internal party conference, election rally and other). In the second phase the same 60 sentences were labelled using Amazon's Mechanical Turk[2] service (AMT). The aim was to explore cultural differences in the perception of charisma from Irish politicians by native and American listeners. Provided that this small study results in similar rater reliability as the initial Irish survey, then it is proposed to annotate a much larger portion of the database using AMT.

Table 1: Synonyms provided to annotators

charismatic	fascinating, captivating, winsome
enthusiastic	energetic, active, positive
inspiring	dynamic, confidence-building, stimulating
likeable	pleasant, agreeable, appealing

In both studies, participants were asked to listen to sentences and rate each of four attributes (charisma, likeability, enthusiasm, and inspiration) on a five point Likert scale, with one representing "not at all" and 5 representing "very" (e.g. for charisma we have "not at all charismatic" to "very charismatic"). To aid discrimination a list of synonyms was provided for each attribute. These are given in Table 1.

Previous studies have shown enthusiasm and inspiration to correlate significantly with charisma [14, 18], thus we can use the correlation between attributes as a measure of the reliability of the annotations, as in [6]. Likeability is more subjective than inspiration or enthusiasm. Raters may agree that a speaker is enthusiastic, but whether they find this enthusiasm likeable varies from rater to rater. We include this attribute in order to see how the varying levels of subjectivity affect inter-rater agreement and regression performance.

3.1 Labelling by Native Listeners

Participants were recruited from the college population via email. Advanced students of psychology, linguistics, and speech processing were targeted. There were a total of 29 respondents (13 male, 16 female), of which 11 had studied psychology, 3 had studied linguistics, and 9 had studied speech processing. Respondents ranged from 19 to 50 years in age. The median age was 29, with a standard deviation of 8. The majority of respondents (24) were native English speakers, with the rest being fluent. All respondents had been resident in Ireland for a number of years.

Each participant was presented with a random selection of 30 out of the 60 sentences. The median number of raters per clip is 11, with a minimum of 6, and a maximum of 19. On completing the annotation, raters could complete an optional questionnaire on their personal politics. This is intended to uncover pre-existing biases, as it has been suggested that prior knowledge of or agreement with the speaker may affect ratings [14]. 22 participants completed this form, of which 16 correctly identified the speaker, and 3 reported being a member or strong supporter of a political party. We will refer to these annotators as native raters.

3.2 AMT labelling

Before conducting a full scale AMT labelling, we wanted to check whether the perception of charisma or inter-rater agreement would be different for the AMT listeners, who would not be as familiar with Irish politicians or colloquialisms. Thus, the previous 60 sentences were re-labelled using AMT. These were divided into Human Intelligence Tasks

[1] Taoiseach /ˈtiːʃəx/ = Irish Prime Minister

[2] `https://www.mturk.com/mturk/welcome`

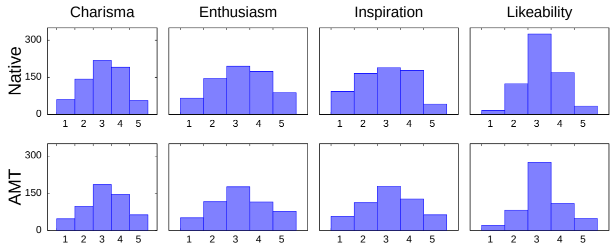

Figure 1: Distribution of raw labels from native and AMT labellers

Table 2: Inter-Rater agreement and comparison with the literature. r = number or raters per clip, N = number of clips in database.

	r	N	corr	Kripp α	Cohen κ
Native Raters	11	60	0.63	0.27	0.48
AMT Raters	9	60	0.64	0.24	0.44
AMT Raters (US)	9	60	0.68	0.29	0.49
AMT (Dáil[3] only)	9	490	0.62	0.24	0.43
Rosenberg [14]	8	45	0.58*	0.22*	0.21
Weninger [18]	10	409	0.57	0.31	0.46

* Not reported in [14]. Calculated from the original data, kindly provided by A. Rosenberg.

(HITs) containing 12 sentences plus one test sentence. The test sentence contains spoken instructions to the worker, and is designed to ensure that they are listening to the sentences. Each HIT was completed by 9 workers.

A total of 31 workers (22 male, 9 female) completed the 45 HITs. Of these 6 completed more than one HIT (12 sentences), with 1 completing all 5 HITs (60 sentences). Annotators were aged from 20 to 59, with a median of 30, and a standard deviation of 9 years. 90% of annotators were American and native English speakers. The remaining 10% were Indian. Raters were not questioned on their personal politics as it is assumed that they are not familiar with Irish politics.

3.3 Assessing annotation reliability

The distribution of labels provided by the native and AMT labellers is shown in Figure 1. At a first glance these appear similar for both sets of raters. Table 2 compares the inter-rater agreement for the charisma labels with that obtained by Rosenberg et al. [14] and Weninger et al. [18]. We compare with with Weninger's raw charismatic ratings as this is more similar to our annotation. Inter-rater agreement is calculated via three metrics: Pearson's correlation coefficient; Cohen's κ; and Krippendorff's α. Pearson's correlation coefficient and Cohen's κ measure pair-wise agreement, thus we report the average agreement between individual raters and the mean rating. For both Cohen's κ and Krippendorf's α we use a quadratic weighting which penalises disagreements according to the magnitude of the difference (as in [14, 18]). The correlation coeffecients and Cohen's κ for the AMT and native labels are comparable with those obtained by Weninger et. al. [18]. The Krippendorf α values for both native and AMT raters are lower than those reported by Weninger et al. [18]. However, they still surpass the agreement obtained by Rosenberg et al. [14]. A slight improvement in inter-rater agreement is obtained by discarding the Indian AMT raters ($\alpha = 0.29$). With only three Indian labellers, it is difficult to say whether this is due to the raters individual differences, or a more systematic cultural bias. Overall, the inter-rater agreement demonstrates the usefulness of this database for the analysis of traits such as charisma.

Table 3: Cross-correlation between mean attribute labels

	enth	insp	like
char	0.93	0.94	0.83
enth		0.93	0.72
insp			0.83

(a) Native Labels

	enth	insp	like
char	0.90	0.91	0.86
enth		0.93	0.87
insp			0.87

(b) AMT Labels

Table 4: Agreement between native and AMT labels

	mean		EWE	
Attribute	ρ	Cohen κ	ρ	Cohen κ
Charisma	0.76	0.65	0.72	0.62
Enthusiasm	0.84	0.78	0.82	0.77
Inspiration	0.81	0.69	0.82	0.69
Likeability	0.49	0.31	0.52	0.37

Gravano et al. [6] assess the reliability of annotation by examining the inter-attribute correlation. Table 3 gives the correlation between the mean labels obtained for each attribute. The high correlation between charisma and enthusiasm and between charisma and inspiration is in line with with Rosenberg et al. [14] and Weninger et al. [18]. The broader term "likeable" was not used in previous studies, however we expected that this too would correlate strongly with charisma. The lower correlation scores obtained for likeability are attributed to the lower inter-rater agreement on this attribute.

Having established that the native and AMT labels are similarly reliable, the next question is how similar are the ratings themselves? To answer this we compute the correlation and Cohen's κ agreement between the aggregate native and AMT labels. Labels are aggregated in two ways: via a simple mean; and using the Evaluator Weighted Estimator (EWE) [8], which weights individual raters according to their agreement with the mean ratings. Table 4 shows that there is strong agreement between the native and AMT raters on charisma, enthusiasm, and inspiration. However, agreement is much lower on likeability. This may be in part due to a cultural difference in the perception of likeability. It may also be due to the fact that the native labellers are more familiar with the speaker. 72% of those who chose to complete the personal politics form recognised the speaker, and so their likeability judgement may be influenced by prior knowledge of his personality, and their agreement or disagreement with his policies. This bias is unlikely to exist in AMT labellers.

4. CHARISMA DETECTION

Two feature sets are explored for the prediction of charisma. The first is a set of prosodic features, which includes pitch, pitch direction, jitter, delta jitter, shimmer, and intensity. These features are motivated largely by the correlations found in Rosenberg et al. [14]. The second set consists of 20 MFCCs, since they have proven effective for a wide range of speech processing applications. All features were extracted

Table 5: Maximum magnitude of correlation ($|\rho|$) between features and charisma ratings (averaged over three folds), and correlation at the 50% cut-off point.

Features	$\|\rho\|$ max	$\|\rho\|$ cut-off
Pros	0.47	0.15
MFCC	0.33	0.10
Pros + MFCC	0.47	0.11

Table 6: Correlation (ρ) and mean absolute error (ϵ) for MLP using EWE labels.

Train	Native				AMT			
Test	Native		AMT		AMT		Native	
Features	ρ	ϵ	ρ	ϵ	ρ	ϵ	ρ	ϵ
Pros	0.60	0.75	0.60	0.81	0.57	0.72	0.46	0.77
MFCC	0.46	0.81	0.37	0.91	0.38	0.95	0.29	1.04
Pros + MFCC	0.66	0.81	0.65	0.65	0.55	0.80	0.49	0.79

over 25 ms frames, spaced 10 ms apart. Since labelling was carried out at a sentence level, statistical functionals were extracted for each feature. This resulted in single feature vector for each clip, of length 109 for the prosodic features, and 160 for the MFCC features.

Since the features sets used are large relative to the size of the database (up to 269 features for 40 training instances), they were pruned before training to avoid over fitting. For each feature set and attribute, features are ranked according to the magnitude of the correlation between the features and ratings, and only the top 50% are kept (80 MFCCs or 55 prosodic features). The magnitude of correlation between the most correlated feature and the charisma ratings, and the magnitude of correlation at the 50% cut-off point are reported in table 5. In general, stronger correlations are obtained with the prosodic features.

A multi-layered perceptron (MLP), implemented via the Weka Toolkit [9], is used to predict charisma. The MLP hidden layer contains half as many nodes as there are input features (40 nodes for MFCC, or 28 for prosodic features). Experiments were carried out using the prosodic and MFCC feature sets individually and using a concatenation of the prosodic and MFCC features.

A three-fold cross-validation was performed. The database is split into three approximately equal partitions. These are chosen so that no one recording appears in more than one partition, and so that there is a balanced distribution of settings (Dáil Éireann[3]; public message; interview; internal party conference; election rally; and other) and recording dates in each partition.

5. RESULTS AND DISCUSSION

Table 6 compares the results of the charisma prediction experiments using native and AMT labels. Prediction performance is measured by the correlation between predicted labels and the true labels. For both label sets, the prosodic features perform better than the MFCCs. Two factors contribute to this. Table 5 shows that prosodic features correlate more strongly with charisma. Similarly, Weninger et al. [18] find that of the 8 most relevant features for charisma, 5 are prosody-based while only 3 are spectral measures. This implies that prosodic features are more useful for charisma prediction. Furthermore, our Irish Political Speech database was collected from a wide range of sources, and so contain a wide range of noise and recording conditions. Automatic Speech Recognition (ASR) tasks [4] have demonstrated MFCCs poor robustness to noise. Future work will explore methods such as RASTA filtering [10] to alleviate this issue.

The best overall feature set for the native labels is the combination of prosodic and MFCC features, while for the AMT labels the prosodic features perform best. The best over all performance ($\rho = 0.66$) compares favourably with the performance achieved by Weninger et al. [18] on their charismatic cover class. However, we must take into account that the experiments performed in [18] are speaker independent, while ours are speaker dependent.

Table 6 also reports results for the prediction of native labels given AMT training labels and visa versa. Not surprisingly the best performance is achieved when the training and test labels come from the same labellers. However, it is promising to note that there is very little difference between the prediction of AMT and native labels when the native labels are used in training.

6. CONCLUSION

In this paper, we have presented a new database of Irish Political Speech, which contains a wide range of recording conditions and noise conditions. Speech has been collected across a span of 7 years, and from a diverse range of settings (parliamentary speech, interviews, etc.). Initially, a small portion of the database was annotated. This annotation has demonstrated that the labels provided by the AMT workers are as reliable as our more trusted native annotators.

With this assurance, a much larger larger set of parliamentary speech was annotated using AMT. Inter rater agreement for this section is given in Table 2 (labelled "AMT (Dáil only)"), and is similar to the agreement obtained from the initial AMT labels. It is intended to annotate large portions of interview, election rally, and inter-party conference speech, in order to compare speaking styles between these different scenarios.

Future studies will explore whether the regression performance reported above can be replicated on these larger datasets. An important issue will be the ability to cope with the wide range of noise conditions present in this database. To do this we intend to draw from the large body of noise compensation literature already existing in the ASR community.

7. ACKNOWLEDGMENTS

The first author is supported by the Irish Research Foundation (IRC), under the Embark initiative. The second author is supported by Google, Inc. and Enterprise Ireland, Innovation Partnership Project IP-2013 0232.

8. REFERENCES

[1] Irish government news channel. https://www.youtube.com/user/MerrionStreetNews.

[2] Oireachtas debates archive. http://oireachtas.heanet.ie/FullArchive/.

[3] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert. A cross-cultural comparison of

[3] Dáil Éireann /dɔɪl ˈɜərən/ = Parliament of Ireland

american, palestinian, and swedish perception of charismatic speech. In *Speech Prosody*, 2008.

[4] C. Chia-Ping, J. Bilmes, and D. P. W. Ellis. Speech feature smoothing for robust asr. In *Acoustics, Speech, and Signal Processing (ICASSP)*, pages 525–528, 2005.

[5] F. D'Errico, R. Signorello, D. Demolin, and I. Poggi. The perception of charisma from voice: A cross-cultural study. In *Affective Comp. and Intelligent Interaction (ACII)*, pages 552–557, 2013.

[6] A. Gravano, R. Levitan, L. Willson, S. Beňuš, J. Hirschberg, and A. Nenkova. Acoustic and prosodic correlates of social behaviour. In *Interspeech*, pages 97 – 100. ISCA, 2011.

[7] S. W. Gregory Jr and T. J. Gallagher. Spectral analysis of candidates' nonverbal vocal communication: Predicting us presidential election outcomes. *Social Psych. Quarterly*, pages 298–308, 2002.

[8] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Automatic Speech Recognition and Understanding*, pages 381–385.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[10] H. Hermansky and N. Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Trans.*, 2(4):578–589, 1994.

[11] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli. Predicting continuous conflict perception with bayesian gaussian processes. *Affective Comp., IEEE Trans.*, 5(2):187–200, 2014.

[12] C. Y. Olivola and A. Todorov. Elected in 100 milliseconds: Appearance-based trait inferences and voting. *J. Nonverbal Behavior*, 34(2):83–110, 2010.

[13] J. W. Pennebaker and T. C. Lay. Language use and personality during crises: Analyses of mayor rudolph giuliani's press conferences. *J. Research in Personality*, 36(3):271–282, 2002.

[14] A. Rosenberg and J. Hirschberg. Charisma perception from text and speech. *Sp. Comm.*, 51(7):640–655, 2009.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[16] C. Strapparava, M. Guerini, and O. Stock. Predicting persuasiveness in political discourses. In *LREC*, pages 1342 – 1345, 2010.

[17] P. Touati. Prosodic aspects of political rhetoric. In *ESCA Workshop on Prosody*, pages 168 – 171, 1993.

[18] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller. The voice of leadership: Models and performances of automatic analysis in online speeches. *Affective Comp., IEEE Trans.*, 3(4):496–508, 2012.