INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Acoustic-Prosodic, Lexical and Demographic Cues to Persuasiveness in Competitive Debate Speeches

Huyen Nguyen[1,2], *Ralph Vente*[3], *David Lupea*[4], *Sarah Ita Levitan*[3], *Julia Hirschberg*[5]

[1]Universität Hamburg, Germany
[2]Erasmus University Rotterdam, Netherlands
[3]Hunter College, City University of New York, USA
[4]New York University, USA
[5]Columbia University, USA

huyen.nguyen-1@uni-hamburg.de, ralph.vente09@myhunter.cuny.edu, dal548@nyu.edu, sarah.levitan@hunter.cuny.edu, julia@cs.columbia.edu

Abstract

We analyze the acoustic-prosodic and lexical correlates of persuasiveness, taking into account speaker, judge and debate characteristics in a novel data set of 674 audio profiles, transcripts, evaluation scores and demographic data from professional debate tournament speeches. By conducting 10-fold cross validation experiments with linear, LASSO and random forest regression, we predict how different feature combinations contribute toward speech scores (i.e. persuasiveness) between men and women. Overall, lexical features, i.e. word complexity, nouns, fillers and hedges, are the most predictive features of speech evaluation scores; in addition to the gender composition of judge panels and opponents. In a combined lexical and demographic feature model, we achieve an R^2 of 0.40. Different lexical features predict speech evaluation scores for male vs. female speakers, and further investigation is necessary to understand whether differential evaluation standards applied across genders. This work contributes a larger-scale debate data set in a democratically relevant, competitive format with high external relevance to persuasive speech education in other competitive settings.

Index Terms: persuasiveness, gender, debate tournament, computational paralinguistics

1. Introduction

Across high-ranked, influential positions in business, academia, the law or politics, the under-representation of women remains an empirical fact [1]. Given the prime importance of oral persuasion skills and the literature gap on gender disparities in speech patterns and evaluations, this research collects and analyzes the transcripts, audio recordings and evaluation scores of 674 speeches in the highest-profile university debate tournaments,[1] to answer: How do different spoken tactics affect persuasiveness across genders?

The well-defined competitive rules of debate tournaments and their participants provide an attractive setting to systematically answer these research questions. First, highly talented college-level students across various fields of study represent their academic institutions to compete in a dynamic tournament setting[2], with *exogenously assigned* debate topics, speaking positions (i.e for or against the topic), opponents and judge panels. Second, given the same amount of preparation time, speaking time and topic, participants must make a persuasive 7-minute speech to convince a trained audience (i.e. judges), who are often themselves accomplished debaters. This debate format simulates British Parliamentary (BP) debates. It is the most popular debate format used in debate education across the world, meant to foster critical thinking and productive civil discourse practices. Third, the transparent and comprehensive evaluation rules, which are enforced by trained adjudication panels, value only topic-focused, comparative argument strength. Specifically, they outlaw *ad hominem* fallacious argumentation strategies,[3] and subjective judgements driven by physical appearance or personal characteristics. In political debates [2], judge courts [3], hiring interviews [4] or entrepreneurial pitch contests [5], establishing how gender differences in speech behavior translate to merit-based argument evaluation is inherently difficult to identify, due to: (i) lack of transparent evaluation scale and rules; (ii) rampant use of *ad hominem* strategies; and (iii) unobserved backdoor agreements or personal beliefs. Therefore, speech evaluation scores serve as a reliable measure of persuasiveness in our study. Lexical, acoustic-prosodic and demographic analysis of debate speech content gives us the unique opportunity to corroborate factors that matter for persuasiveness across genders in real-life, high-stake contests.

The rest of this paper proceeds as follows. Section 2 outlines the related literature and Section 3 describes the debate corpus and the key descriptive statistics on feature groups. Section 4 explains the regression and feature analyses. We conclude in Section 5 with a discussion of avenues for future research.

2. Related Work

Our investigation into detecting linguistic cues to persuasiveness relates to the literature on detecting speaker states and traits from speech using acoustic-prosodic and lexical features. Studies close to ours include work on detecting charisma [6], likeability [7], emotion [8], personality [9], and trust [10]. Recent work by [11] on 20s speech clips found that persuasiveness ratings are highly correlated with charisma ratings, and differ along the speaker gender dimension. Relatedly, in a corpus of task-oriented spontaneous speeches, [7] noted that the gender of both speakers and listeners correlates with percep-

[1]i.e. World & European Universities Debate Championship, HWS Round Robin Champion League.

[2]i.e. 5 rounds Hobart William Smith (HWS) Round Robin Invitational Champion League, or 9 rounds in the World and European Universities Debate Championship.

[3]i.e. attacks on characteristics of the speaker instead of the substance of the arguments.

http://dx.doi.org/10.21437/Interspeech.2021-1891

tion of likeability. A recent investigation of vocal pitch in 74 000+/- US Congressional floor speeches by [12] found that higher emotional intensity was correlated with persuasiveness in female pitches when they discussed women's topics. The work closest to our paper are [13] on creating Virtual Debate Coach for young politicians and [14] on 30 Oxford style debates, which employ multi-modal analyses. However, their data was constructed in a lab-based environment to train virtual audiences. Comparatively, our work complements existing debate corpora [15, 16, 17] with a large-scale, real-life debate tournament corpus and establishes the link between expert human-scored speech evaluations and persuasion-relevant lexical, acoustic features and demographic characteristics of speakers and judges.

Our multi-model analysis on this novel debate data set also complements the current persuasion detection literature, which often uses text-only corpora, in either lab-based environments or in contexts with inevitably noisy or non-transparent evaluation criteria. For instance, the latest work by [17] to train the IBM Project Debater analyzes argument quality of transcripts from lab-constructed speeches of professional debaters. For online debates, [18] detects persuasiveness in online social media discussions; [19] measures dynamic persuasiveness of online debaters over time; and [20] identifies the attackability of sentences on the ChangeMyView platform on Reddit. By coupling acoustic-prosodic cues with lexical and demographic features in our debate tournament corpus, we can comprehensively investigate the interplay of these features across genders in real-life, high-stake competitions.

3. Data & Descriptive Statistics

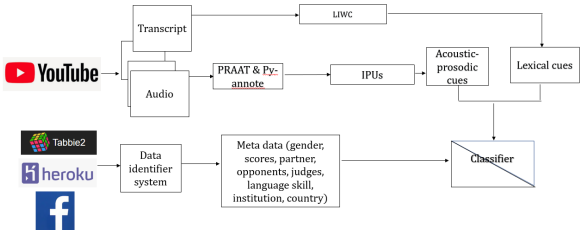

Figure 1: *Data collection & analysis procedure*

Figure 1 summarizes the data collection and analysis procedures we use on this debate data set. From YouTube videos of tournament debates from 2008 to 2018, we extracted over 225 hours of human-transcribed verbatim transcripts ($98 + \%$ accuracy) and audio recordings of these debates. Because of high demand for high-level debate recordings for training purposes, the vast majority of recorded debates are in the upper end of debate tournaments. The scores and descriptive statistics of male and female speakers is summarized in the first line of Table 1. These speeches are in the right tail of the score distribution, since the overall average of 78.40 in this data set is significantly higher than the population mean score of 75. As the recording quality varies across competitions, we ended up with 168+/- hours of usable transcripts and audio files after omitting debates with over 4% of inaudible audios. Since debate tournaments are heavily male-dominated (65 - 70% of participants are male), we currently have 475 speeches given by men but only 199 by women.

To obtain the demographic data, we web-scraped this information from tournament score archival data and confirmed the accuracy with tournament directors. This information is captured in the following categorical variables: `non_native` whether the debater is a native English speaker, `institution_rank` whether the institution the speaker represents is ranked in the top 50 universities worldwide or not), `room_female_dominated` whether the debate room has at least as many male as female speakers (8 speakers in total/debate) `panel_female_dominated` whether the judge panel has at least as many male as female judges (3 to 5 judges/panel/debate), `chair_female` whether the chair judge is a woman. After matching demographic information and evaluation scores from the web-scraped data with the YouTube videos, we triple-checked the videos and transcripts for: (i) transcript completeness and speech comprehensibility; (ii) speaker gender, team identity, speaking position; and (iii) debate round, judges' role and gender composition.

Next, we extracted persuasion-relevant LIWC features using [21]'s program, after removing inaudible/clapping/laughter remarks, punctuation and brackets in the transcripts. These features include: (i) basic lexical features (e.g. `WordCount`, `SentenceCount`, `Character-perWord`, `Fillers`, `Hedges`, etc.), (ii) Parts of Speech, psychometric measures to measure speech style (e.g. `Analytic`, `Authentic`, `Tone`) and (ii) debate-strategic strategies (i.e. `POIReject` number of times the speakers reject offered questions from opponents, `BPWoPhra` proportion of debate-specific filler phrases in speech).

For acoustic-prosodic features, we used Parselmouth [22], a Python interface to Praat [23] and obtained commonly studied acoustic-prosodic features in speech research: (i) `intensity` (the energy in a sound wave), (ii) `pitch` (the fundamental frequency of a voice), (iii) harmonics to noise ratio (`HNR`, a measure of voice quality), (iv) mean percentage jitter (`localJitter`, cycle-to-cycle variation of fundamental frequency) and (v) mean percentage shimmer (`localShimmer`, variability of the peak-to-peak amplitude in decibels) [24]. These features were sampled 100 times per second. Then, all instances of the minimum value of each feature were dropped so our statistics represent voiced regions only. To reduce the dimensionality of time-series data, we extracted 5 quantiles of pitch and intensity. For example, `intensity_20pctl` is the 20th percentile of all samples of intensity for a particular speech, and `intensity_60-80` is the difference between the 60th and 80th percentiles of intensity. In our analysis, acoustic features are z-score normalized by gender.[4]

Table 1 provides the descriptive statistics of all features across all speeches, and those extracted from only speeches given by men or by women. Demographic-wise, most features appear quite balanced between male and female speakers, except for `institution_rank`. While only 28% of male speakers represented the top-50-ranked universities, 42% of women in this sample competed on behalf of top-50-ranked universities. Regarding acoustic-prosodic features, mean harmonics-to-noise-ratio, intensity and pitch differed significantly between male and female speakers (at 95% confidence level). Unsurprisingly, the largest mean and variance difference is pitch, which is in line with existing literature. For the lexical features, speeches given by men are significantly more analytical, with shorter sentences. There are very small difference in fillers (e.g. um, uh), and virtually no difference in the number of times men and women were asked questions during their speeches and the amount of BP-specific words and

[4]Both the raw and normalized features are reported in Table 2.

phrases.

Table 1: *Descriptive statistics of evaluation scores and features: Male vs. Female speakers ($N_{M\cup F} = 674$, $N_M = 475, N_F = 199$)*

	Speeches given by...					
	female speakers		**male speakers**		**all speakers**	
	μ	σ	μ	σ	μ	σ
Speech Scores (50 - 100 scale)	78.63	3.88	77.84	4.58	78.4	4.11
Demographic						
chair_female	0.39	0.49	0.36	0.48	0.37	0.48
female	1.00	0.00	0.00	0.00	0.27	0.44
institution_rank	0.28	0.45	0.42	0.49	0.38	0.49
non_native	0.33	0.47	0.23	0.42	0.25	0.43
panel_female_dominated	0.45	0.50	0.35	0.48	0.38	0.49
panel_female_ratio	0.43	0.50	0.32	0.47	0.35	0.48
room_female_dominated	0.41	0.49	0.13	0.33	0.20	0.40
Acoustic						
mean_hnr	-113.46	34.03	-100.96	34.89	-104.33	35.07
std_hnr	96.57	11.01	97.25	8.84	97.07	9.47
mean_intensity	62.66	4.66	64.63	5.18	64.10	5.12
std_intensity	8.46	1.80	9.30	3.28	9.08	2.98
mean_pitch	180.15	35.94	132.72	29.76	145.50	37.90
std_pitch	141.70	17.85	114.04	13.14	121.49	19.04
localShimmer	0.169	0.017	0.177	0.177	0.175	0.017
localJitter	0.027	0.004	0.030	0.030	0.029	0.004
Lexical						
Adj	9.59	1.77	10.04	2.12	9.92	2.04
Adv	8.40	1.47	7.84	1.50	7.99	1.51
Analytic	43.81	13.56	50.43	15.04	48.64	14.94
ArguIndi	3.38	1.36	3.24	1.33	3.28	1.34
Authentic	30.52	12.83	26.97	12.85	27.93	12.93
BPWoPhra	0.68	0.46	0.71	0.43	0.70	0.44
CertainLIWC	1.70	0.76	1.59	0.66	1.62	0.69
CharCount	6701.32	945.88	6866.03	951.40	6821.63	951.83
CharperWord	4.42	0.50	4.49	0.50	4.47	0.50
Fillers	4.95	4.64	4.71	4.63	4.77	4.63
Hedges	3.62	1.15	3.33	1.17	3.41	1.17
Noun	20.65	2.32	21.48	2.09	21.26	2.18
POIReject	0.07	0.15	0.08	0.15	0.08	0.15
PersonalPron	8.06	1.73	7.38	1.64	7.57	1.69
SentCount	69.49	16.18	76.54	17.85	74.64	17.68
SixLetterPerc	17.80	2.81	18.13	2.96	18.04	2.92
TentaLIWC	2.82	0.91	2.93	0.95	2.90	0.94
Tone	47.48	24.30	47.06	24.04	47.17	24.09
Verb	16.59	1.73	16.36	1.56	16.42	1.61
WordCount	1502.14	206.93	1528.76	197.97	1521.59	200.57
WordperSent	22.47	4.83	20.81	4.26	21.26	4.48
whWordLIWC	2.49	0.73	2.55	0.70	2.53	0.70

4. Regression Experiments

4.1. Overall analysis

Our goal is to predict speech persuasiveness i.e. speech evaluation scores, based on the acoustic-prosodic, lexical and demographic features summarized in Table 1. Our baseline model is a Linear Regression with a bias term, using 10-fold cross-validation. We further ran LASSO, ridge and random forest models to check for the most predictive models. We trained and evaluated these models separately for: (1) acoustic-prosodic, (2) lexical, (3) and demographic features, and (4) a combined model, to robustly check their relevance to speech evaluation scores. For our data set, we found that random forest models with complex decision boundaries over-fit the test set dramatically. Given the tuning parameters we used, we observed no significant improvement in predictive performance between Linear, Lasso, and Ridge regression models. Therefore, we report the most predictive features for each model as the average of 100 trials from linear regression experiments.

To evaluate the results, we use the mean R^2, i.e. the proportion of the variance in judges' scores that are explained by the features, and the standard deviation R^2 across all folds and trials. The linear regression results for these four models are shown in the first column of Table 2, with the top 5 most predictive features and their statistical significance for each group of features, along with their associated weights (i.e. coefficient values learned by the model). Our

Table 2: *Most predictive features for the entire data set ($N_{M\cup F} = 674$)*

Feature group	Feature	Weight
Raw Acoustic-prosodic	intensity_20pctl***	1.9678
mean $R^2 = 0.1812$	intensity_60-80pctl***	1.4226
stdev $R^2 = 0.0906$	pitch_60-80pctl**	0.0596
	intensity_0-20pctl***	-1.8084
	intensity_min***	-1.8301
	Bias	63.4868
Gender Normalized Acoustic-prosodic	intensity_20pctl ***	5.6675
mean $R^2 = 0.1755$	intensity_80pctl***	5.2740
stdev $R^2 = 0.0919$	pitch_60-80pct**l	0.7240
	intensity_min***	-0.7390
	intensity_40pctl***	-9.8571
	Bias	78.72
Lexical	CharperWord***	0.7353
mean $R^2 = 0.3121$	Noun***	0.2293
stdev $R^2 = 0.0782$	CharCount***	0.0015
	Fillers***	-0.1791
	Hedges***	-0.2917
	Bias	62.3433
Demographic	panel_female_dominated***	2.6301
mean $R^2 = 0.2356$	room_female_dominated	1.2266
stdev $R^2 = 0.0929$	non_native***	0.1578
	institution_rank***	-0.1349
	chair_female	-0.2203
	panel_female_ratio**	-2.1284
	Bias	78.12
Combined Model	institution_rank***	0.9851
mean $R^2 = 0.4039$	room_female_dominated	0.9851
stdev $R^2 = 0.0895$	CertainLIWC	0.1481
	Hedges**	0.0880
	Noun***	0.0012
	SixLetterPerc**	-0.2508
	CharCount***	-0.2508
	Fillers***	-0.2917
	Bias	66.6863

*$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: *Most predictive features: Male vs. Female speakers ($N_M = 475, N_F = 199$)*

Feature Group	**Male Speakers**		**Female Speakers**	
	Feature	Weight	Feature	Weight
Acoustic-prosodic				
	mean R^2	0.1823	mean R^2	0.0587
	stdev R^2	0.1179	stdev R^2	0.2311
	intensity_20pctl***	2.2177	hnr_20pctl	2.9465
	intensity_40-60pctl***	1.3137	intensity_60-80pctl***	1.6151
	pitch_80pctl **	0.0211	pitch_80pctl***	0.0210
	intensity_0-20pctl **	-2.098	intensity_min**	-0.0197
	intensity_min***	-2.1210	intensity_40pctl	-3.0505
	Bias	66.032	**Bias**	73.63
Lexical				
	mean R^2	0.2700	mean R^2	0.2156
	stdev R^2	0.0994	stdev R^2	0.1671
	BPWoPhra***	0.9854	CertainLIWC***	1.1895
	whWordLIWC***	0.7308	TentaLIWC	0.5362
	Analytic***	0.0487	ArguIndi	0.2467
	WordperSent	0.0078	Verb	0.0643
	Adj***	-0.3784	Adj	-0.4340
	Bias	77.58	**Bias**	79.37
Demographic				
	mean R^2	0.2244	mean R^2	0.1676
	stdev R^2	0.1200	stdev R^2	0.2061
	institution_rank***	2.5967	institution_rank***	3.0218
	panel_female_dominated***	1.6228	panel_female_dominated	0.0260
	chair_female	-0.0434	chair_female	-0.3264
	panel_female_ratio	-0.3234	room_female_dominated*	-0.3264
	non_native***	-2.0332	non_native***	-2.6522
	Bias	77.80	**Bias**	79.13
Combined				
	mean R^2	0.3002	mean R^2	0.0024
	stdev R^2	0.10649	stdev R^2	0.5676
	CharCount**	0.0032	std_intensity	-0.063
	Fillers***	-0.1494	Adj	0.1898
	Noun***	0.2302	CharCount***	0.0015
	SixLetterPerc	-0.0139	Fillers***	-0.1993
	WordCount	-0.0084	TentaLIWC	0.4019
	Bias	65.6783	**Bias**	67.4882

*$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

acoustic-prosodic model (1) achieved a mean R^2 of .18, with the intensity and pitch explaining some of the variance. Our lexical-only model (2) was the most predictive model, achieving a mean R^2 of 0.31. Our demographic-only model (3) achieved a mean R^2 of 0.24. These individual feature set regression results show that speech verbal content is the most predictive of speech persuasiveness. Overall, our best model is the combined lexical and demographic model, with mean R^2 of 0.40. The most predictive features from this model are `institution_rank`, `room_female_dominated`, `SixLetterPercentage`, `CharCount`, `Hedges`, `Noun`, `Fillers` and `CertainLIWC`.

With respect to the most predictive features of persuasiveness across these models, the best performing acoustic-prosodic features are intensity and pitch related features (not HNR), as shown in Table 2. Specifically, low and high intensity percentiles, i.e. softer or louder speech, correlate positively with persuasiveness. This suggests that varying speech intensity matters for persuasiveness, differing from the finding that intensity variation lowers "likes" on YouTube videos in [25]. Regarding lexical features, the most predictive ones are `CharPerWord`, i.e. number of characters per word, which indicates that longer, more complex word usage is more persuasive. In line with the common belief that polished and confident language is more persuasive, we found that `Fillers` (e.g. um, uh) and `Hedges` (words that express uncertainty, e.g. "sort of") negatively correlate with speech evaluation scores. Finally, for demographic features, the judge panel and debate room gender composition, i.e. `panel_female_dominated` and `room_female_dominated`, have the largest weights. This finding highlights the importance of gender of both speakers and judges in speech evaluation. We also note that whether a speaker represents a top-50-ranked academic institution i.e. `institution_rank` correlates positively with persuasiveness, only in a demographic and lexical feature model.

4.2. Male vs. Female Speakers

To understand which features matter for persuasiveness across genders, we ran the same analysis on the subsets of speeches given by male vs. female speakers, as reported in Table 3. Lexical and demographic features clearly predict persuasiveness better than acoustic-prosodic ones for both genders. Note here that the features that correlate with higher persuasiveness differ for men and women. In the combined model, for both male and female speakers, more character counts (i.e. longer, more complex words) correlate with higher scores; whereas speeches with more fillers receive lower scores. Nonetheless, since our model has a significantly worse fit for female speakers than for male speakers, further investigation is necessary to understand what matters for persuasiveness in female speeches.

The low R^2 of 0.06 and the high standard deviation of 0.23 on acoustic-prosodic models among female speeches suggest two potential reasons. One possibility is that judges could employ differential standards when evaluating the auditory aspects of a speech for male vs. female speakers. For speeches given by men, judge evaluations are more consistent, with lower standard deviation of 0.12 and a better fit of 0.18. This consistent evaluation standard for male speeches holds also for lexical and demographic features. A combined feature regression model (4) applied on male vs. female speech subsets gives us an R^2 of 0.30 for the former, while only 0.24 for the latter. However, while the best performing features differ between male and female speakers, we cannot reach a real conclusion here, given the current high standard deviation.

Another possibility for these results is that the current data set has significantly fewer number of speeches by women, with only $N_F = 199$ female speeches compared to $N_M = 475$ male speeches. Nonetheless, upon running the same regression experiment on a randomly selected subset of similar size on speeches given by men ($N_M = 190$), we observe an R^2 performance of 0.32 and a standard deviation of 0.25. Although the overall fit degrades for this subset of speeches by men, the fit is comparatively much better. This result suggests that sample size is *not* necessarily the only driver of this poor fit.

5. Conclusions

This paper presents a multi-modal analysis to predict persuasiveness in a novel data set of 674 real-life, high-stake debate tournament speeches. Using four 10-fold cross validation regression models, we found that the most effective model was a combination of demographics and lexical features. Linear regression models perform the best among the tested models, with ridge regression as the next best model. Gender composition of judge panels is important in predicting speech scores, thus highlighting the critical role of judge identities. Breaking down the analysis into male vs. female speakers reveals potentially effective debate persuasiveness for male speakers, yet further investigation is necessary on female speakers to conclusively pinpoint the crucial speech elements perceived as persuasive by judges. All in all, this work contributes to the scientific understanding of persuasiveness in competitive, real-life tournaments among highly talented and motivated speakers, thus having important relevance for persuasive speech education in other competitive settings.

The current analysis is qualified by three limitations that offer fruitful avenues for future research. First, the limited number of observations for female speeches and the lack of sentence-level lexical variables relevant for persuasion both need to be addressed. In future work, we plan to collect more female speeches given and produce a balanced overview of varying feature importance for these minority groups. Second, given the wide assortment of recording conditions e.g. echo, audience latent noise, differing response of the microphone, to improve R^2 performance in acoustic-prosodic models, we will also de-noise the files and re-extract the acoustic features. Finally, we want to incorporate visual cues e.g. body language; syntactic complexity; argument components (i.e. claims, premises) and relations (i.e. support/attack) in speeches to complete the multi-modal predictor of persuasiveness, particularly across social groups (e.g. gender, native English speakers, experienced vs. novice debaters).

6. Acknowledgements

We gratefully acknowledge the financial fund from the Diversity & Inclusion Office of Erasmus University Rotterdam, Netherlands to collect and construct the data set. We are also indebted to Columbia Speech Lab for providing us with the necessary platform and resources to conduct the acoustic-prosodic analysis part of the paper.

7. References

[1] C. Goldin, S. P. Kerr, C. Olivetti, and E. Barth, "The expanding gender earnings gap: Evidence from the lehd-2000 census," *American Economic Review*, vol. 107, no. 5, pp. 110–14, 2017.

[2] L. Hargrave and T. Langengen, "The gendered debate: Do men and women communicate differently in the house of commons?" *Politics & Gender*, pp. 1–27, 2020.

[3] E. Ash, D. L. Chen, and A. Ornaghi, "Implicit bias in the judiciary," 2019.

[4] C. Isaac, B. Lee, and M. Carnes, "Interventions that affect gender bias in hiring: A systematic review," *Academic medicine: journal of the Association of American Medical Colleges*, vol. 84, no. 10, p. 1440, 2009.

[5] A. W. Brooks, L. Huang, S. W. Kearney, and F. E. Murray, "Investors prefer entrepreneurial ventures pitched by attractive men," *Proceedings of the National Academy of Sciences*, vol. 111, no. 12, pp. 4427–4431, 2014.

[6] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[7] A. Gravano, R. Levitan, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[9] G. An, S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine, and J. Hirschberg, "Automatically classifying self-rated personality scores from speech." in *INTERSPEECH*, 2016, pp. 1412–1416.

[10] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues." in *INTERSPEECH*, 2018, pp. 416–420.

[11] Z. Yang, J. Huynh, R. Tabata, N. Cestero, T. Aharoni, and J. Hirschberg, "What makes a speaker charismatic? producing and perceiving charismatic speech," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 685–689.

[12] B. J. Dietrich, M. Hayes, and D. Z. O'BRIEN, "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech," *American Political Science Review*, vol. 113, no. 4, pp. 941–962, 2019.

[13] M. T. Petukhova, V., A. Malchanau, and H. Bunt, "Virtual debate coach design: assessing multimodal argumentation performance," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 41–50.

[14] M. Brilman and S. Scherer, "A multimodal predictive model of successful debaters or how i learned to sway votes," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 149–158.

[15] V. Petukhova, A. Malchanau, Y. Oualil, D. Klakow, S. Luz, F. Haider, N. Campbell, D. Koryzis, D. Spiliotopoulos, P. Albert *et al.*, "The metalogue debate trainee corpus: Data collection and annotations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[16] M. Chollet, P. Ghate, C. Neubauer, and S. Scherer, "Influence of individual differences when training public speaking with virtual audiences," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 1–7.

[17] M. Orbach, Y. Bilu, A. Toledo, D. Lahav, M. Jacovi, R. Aharonov, and N. Slonim, "Out of the echo chamber: Detecting countering debate speeches," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, Jul. 2020, pp. 7073–7086.

[18] C. Hidey and K. McKeown, "Persuasive influence detection: The role of argument sequencing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[19] K. Luu, C. Tan, and N. A. Smith, "Measuring online debaters' persuasive skill from text over time," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 537–550, 2019.

[20] Y. Jo, S. Bang, E. Manzoor, E. Hovy, and C. Reed, "Detecting attackable sentences in arguments," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, Nov. 2020.

[21] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc 2015," Tech. Rep., 2015.

[22] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[23] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[24] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth Annual Conference of the International Speech Communication Association (ISCA Speech)*, 2007.

[25] S. Berger, O. Niebuhr, and M. Zellers, "A preliminary study of charismatic speech on youtube: Correlating prosodic variation with counts of subscribers, views and likes." in *INTERSPEECH*, 2019, pp. 1761–1765.