Automatically Classifying Self-Rated Personality Scores from Speech

Guozhen An[1], *Sarah Ita Levitan*[2], *Rivka Levitan*[4], *Andrew Rosenberg*[3], *Michelle Levine*[2], *Julia Hirschberg*[2]

[1]Department of Computer Science, CUNY Graduate Center, USA
[2]Department of Computer Science, Columbia University, USA
[3]Department of Computer Science, Queens College (CUNY), USA
[4]Department of Computer and Information Science, Brooklyn College (CUNY), USA

gan@gradcenter.cuny.edu, sarahita@cs.columbia.edu, rlevitan@brooklyn.cuny.edu
andrew@cs.qc.cuny.edu, mlevine@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Automatic personality recognition is useful for many computational applications, including recommendation systems, dating websites, and adaptive dialogue systems. There have been numerous successful approaches to classify the "Big Five" personality traits from a speaker's utterance, but these have largely relied on judgments of personality obtained from external raters listening to the utterances in isolation. This work instead classifies personality traits based on self-reported personality tests, which are more valid and more difficult to identify. Our approach, which uses lexical and acoustic-prosodic features, yields predictions that are between 6.4% and 19.2% more accurate than chance. This approach predicts Openness-to-Experience and Neuroticism most successfully, with less accurate recognition of Extroversion. We compare the performance of classification and regression techniques, and also explore predicting personality clusters.

Index Terms: personality recognition, self-reported personality

1. Introduction

Personality refers to individual differences in characteristic patterns of thinking, feeling, and behaving [1]. A commonly used model of personality is the NEO-FFI five factor model of personality traits, also known as the Big Five: Openness to Experience (having wide interests, imaginative, insightful), Conscientiousness (organized, thorough, a planner), Extroversion (talkative, energetic, assertive), Agreeableness (sympathetic, kind, affectionate), and Neuroticism (tense, moody, anxious) [2]. These traits were originally identified by several researchers working independently [3] and the model has been employed to characterize personality in multiple cultures [4].

Big Five traits have been found to predict many life outcomes, including academic [5] and occupational [6] success, interpersonal relationships [7, 8], and health outcomes [9]. Personality recognition also has potential uses in many applications. Recommendation systems (e.g. travel [10], music [11]) can be customized for specific personalities. Personality recognition can also inform matching algorithms for dating applications [12]. Dialogue systems too can adapt to users personalities, and this adaptation has been found to be preferred by users [13]. Automatic methods for natural language generation with personality traits have also been developed and evaluated [14].

In our current work we experiment with ways to automatically identifying the NEO-FFI Big Five personality traits from speech, which will be useful for applications such as dialogue systems. Although there is previous research on this task, most has focused on predicting personality scores labeled by annotators asked to identify personality traits of others, rather than from self-reported personality inventories. Although ratings by observers who know the subject well are considered valid in personality research, ratings by strangers have been shown to correlate only weakly with self reports, and have moderate to weak internal consistency (as measured by Cronbach's alpha) [15].

We focus instead on the prediction of self-reported NEO-FFI scores, which are considered more valid indicators of a persons true personality. This is a much more difficult task than predicting stranger ratings, which are based only on the speech samples, which therefore necessarily contain all the information needed for the prediction. To classify personality from speech using self-reported personality inventories, we compare three approaches: classification of high/medium/low personality score categories, regression against continuous personality scores, and classification of personality score clusters.

In Section 2, we review previous work on linguistic markers of personality. We describe the corpus used for our study in Section 3. In Section 4 we outline the methods for preparing the data and features used in classification. Section 5 presents the results of our machine learning experiments. We conclude in Section 6 and discuss future research directions.

2. Related Work

There have been numerous successful approaches to the automatic identification of personality from text and from speech. Mohammadi et al. [16] used prosodic features to detect personality from short ten second audio clips, labeled by human judges with observer personality scores. They use an SVM classifier to tackle a binary problem whether a clip is above or below the mean score for each of the NEO-FFI five personality traits. They report recognition rates ranging from 64.7% (Agreeableness) to 79.4% (Extraversion).

Lexical features have also been used for personality identification. Argamon et al. [17] extracted four sets of stylistic lexical features from student essays and classified the essays as high or low (top or bottom third) for extraversion and neuroticism, using an SMO classifier. They analyzed the contributions of lexical features for each trait. Another source of lexical features, Linguistic Inquiry and Word Count (LIWC) [18] categories, have been shown to correlate with Big Five personality traits, both in writing samples [19] and in spoken dialogue [20].

In a comprehensive study of personality recognition, Mairesse et al. [21] explored LIWC, psycho-linguistic, and prosodic feature sets; compared classification, regression, and ranking algorithms; and evaluated the predictions of both observer reports and self reports of personality. Their results indicate that observer reports are easier to predict — they achieved good results with models of observed personality but no results above baseline with models of self-reported personality.

We believe that self-reported personality is important to model, as this has been shown to be valid in many experiments (e.g. retest reliability, correlation with life outcomes). Our work can be viewed as a continuation of the groundwork laid by [21]. We similarly explore both prosodic and lexical features, but we use an expanded acoustic-prosodic feature set, and we add the Dictionary of Affect in Language (DAL) feature set [22] which has not been previously used for personality identification. While the previous work mentioned above classified personality traits by binning scores into high and low for each trait, using the median, mean, or top/bottom third as a threshold, we introduce here a new method of binning personality scores using population mean scores. We also explore a new approach in which we cluster personality scores, and then classify speech samples according to their cluster ids.

3. Corpus

The corpus we examine [23, 24] consists of within subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), both speaking English, where native language is defined as language spoken at home until age 5. The corpus includes data from 172 subject pairs – 122.5 hours of speech. To our knowledge, this is by far the largest corpus of cleanly recorded deceptive and non-deceptive speech collected and transcribed, and includes self-identified truth/lie labels.

The corpus was collected using a fake resume paradigm in which pairs of subjects were recorded playing a lying game, alternating between interviewing their partner and being interviewed about answers to a set of 24 biographical questions. Demographic data was also obtained from each subject and subjects also filled out a NEO-FFI (5 factor) personality inventory [2], assessing Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N).

We also collected a 3-4 minute baseline sample of speech from each subject for use in speaker normalization, in which the experimenter asked the subject open-ended questions (e.g., What do you like the best/worst about living in NYC?). Subjects were instructed to be truthful in answering. Once both subjects had completed all the questionnaires and we had collected both baselines, they began the lying game.

Transcripts for the recordings were obtained using Amazon Mechanical Turk (https://www.mturk) (AMT). Three transcripts for each audio segment from different ‘Turkers’ were obtained, and combined using *rover* techniques [25], producing a rover output score measuring the agreement between the initial three transcripts. For clips with a score lower than 70%, transcripts were manually corrected; we needed to hand correct 9.7% of clips. For the experiments in this paper we used the 331 baseline files that have been corrected to date, comprising approximately 23 hours of speech.

4. Method

4.1. NEO Score Binning

Figure 1: *Interviewing subject.*

One challenge of predicting personality is how to set up the machine learning experiments. The NEO scores are calculated on a continuous scale for each of the five dimensions, so it is natural to model this as a regression problem. Another approach is to convert the numeric scores to nominal values using thresholds, and to model this as a three-class classification problem. We compare results of the two approaches in this work, as well as a third approach, predicting personality score clusters, as described in Section 5.3.

We use the thresholds provided in [26] to label the NEO scores as “High” (HI), “Medium” (ME) or “Low” (LO) for each dimension. These thresholds are determined by population norms from a large sample of administered NEO-FFI, and are different for males and females. Table 1 shows the mapping of numeric NEO scores to the three categorical labels.

Table 1: *Personality mapping from a continuous scale to High, Medium, and Low.*

Trait	Gender	LO	ME	HI
O	Male	< 23	$23 =<, <= 30$	> 30
	Female	< 23	$23 =<, <= 30$	> 30
C	Male	< 30	$30 =<, <= 37$	> 37
	Female	< 32	$32 =<, <= 38$	> 38
E	Male	< 24	$24 =<, <= 30$	> 30
	Female	< 25	$25 =<, <= 31$	> 31
A	Male	< 29	$29 =<, <= 35$	> 35
	Female	< 31	$31 =<, <= 36$	> 36
N	Male	< 13	$13 =<, <= 21$	> 21
	Female	< 16	$16 =<, <= 25$	> 25

Table 2 shows the distribution of the three categorical labels in each trait after mapping from NEO-FFI scores. As we might expect, the three classes are highly unbalanced, with the majority of subjects usually falling into the Medium class, and a smaller percentage in either the High class or the Low class.

Table 2: *Distribution of three class after relabeling*

Distribution	O	C	E	A	N
LO	22	147	79	125	74
ME	129	130	125	141	107
HI	180	54	127	65	150

4.2. Low-Level Descriptor Features

Previous research showed that different personality traits can be detected by a variety of speech factors, such as fundamental frequency [27], voice quality, intensity [28], frequency and duration of silence pauses [29]. Motivated by these findings, we used the OpenSMILE library to extract acoustic-prosodic features [30]. The OpenSMILE Low-Level Descriptor (LLD) feature set contains approximately 6,373 acoustic-prosodic features as described in the Interspeech 2013 COMPARE Challenge [31]. These are with extracted using the baseline 2013 Challenge configuration. The Low-Level Descriptor features include pitch (fundamental frequency), intensity (energy), spectral, cepstral (MFCC), duration, voice quality (jitter, shimmer, and harmonics-to-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness.

4.3. Fundamental Frequency Variation Features

It has been mentioned in previous research [32, 33] that there are strong correlations between personality and fundamental frequency. In order to capture this information, we extracted 42 features which come from fundamental frequency variation (FFV) spectrum with 7 components [34]. From each of the 7 spectrum components, we extract 6 features: mean, minimum, maximum, median, standard deviation, and variance. These features have been found to be helpful in characterizing dialogs [34] and also in acoustic modeling for speech recognition [35]. The FFV features capture a frame level spectral representation of f0 dynamics, as opposed to the pitch features in the OpenSmile LLD set, each of which describes the pitch of the entire sound.

4.4. Linguistic Inquiry and Word Count Features

Psycholinguistic studies [36] show that the people choose words not only because of the linguistic meaning, but also because of psychological conditions, such as emotion, personality and relational attitude. Therefore, it is possible to detect personalities through text analyses associated with psycholinguistic techniques. Inspired by [19, 21], we used Linguistic Inquiry and Word Count (LIWC) [18] to extract the lexical features. LIWC is a text analysis program that calculates the degree to which people use different categories of words, and can determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions. LIWC features have been used in many studies to predict outcomes including personality [19], deception [37], and health [38]. We extracted a total of 130 LIWC features based on the 64 LIWC categories: 64 features based upon the ratio of words appearing in each LIWC categories over total word count; 64 features based on the ratio of words appearing in each LIWC categories over the total words appearing in any LIWC category; the total number of words appearing in any LIWC category; and the total word count.

4.5. Dictionary of Affect Features

The psychology literature [39] suggests that arousal highly correlates with some dimensions of personality, especially extraversion. Therefore, we used Whissell's Dictionary of Affect in Language (DAL) [22] to extract additional features. The DAL is a lexical analysis tool which is used for analyzing emotive content of speech especially for *pleasantness*, *activation* and *imagery*. It lists approximately 4500 English words, each with ratings for these three categories in the DAL. These were obtained from multiple human judges. We extract nineteen features derived from the DAL scores for each word in each subject's baseline interview transcript. From all words' pleasantness, activation and imagery scores, we calculated the mean, minimum, maximum, median, standard deviation, and variance. We also added the number of words in the transcript that appear in the DAL.

5. Results

After feature extraction is completed, we train and test our model. There are three different experiments in this paper for personality prediction: predicting continuous personality scores, personality score categories, and personality score clusters. We trained models using each of our four different feature sets, then combined them in different ways to determine how much each feature set was contributing either independently or in combination. We used the Weka [40] SMOreg for continuous scale and the SMO classifier for class classification and cluster classification to generate personality hypotheses. All Weka parameters were kept at their default values.

5.1. Personality scores

Our first model attempted to predict continuous personality scores using regression. We find that different feature sets perform differently on the different personality traits. We use the Spearman correlation to compare the results from different set of features, since the self-reported personality scores are not evenly distributed. Table 3 shows that the LLD feature set performs best on Conscientiousness; the FFV feature set performs best on Openness to Experience; the LIWC feature set rates highest on Agreeableness; the DAL feature set achieved the highest performance on Extraversion; and a combination of four feature sets performs best on Neuroticism. We did not set any baseline performance for this experiment since there were no other comparable experiments and no other plausible simple baselines. Although [21] did use a similar procedure for experimenting on self-reported personality recognition, they used t-tests to report their results rather than correlations, and none of their models showed significant improvement over the baseline on self-reported personality scores.

Table 3: *Regression Performance of Personality Recognition. (Spearman Correlations)*

Feature	O	C	E	A	N
LLD	0.308	**0.233**	0.148	0.238	0.052
FFV	**0.409**	0.028	0.232	0.304	0.302
LIWC	0.267	-0.086	0.109	**0.340**	0.176
DAL	0.287	-0.120	**0.381**	0.118	0.226
Combined	0.238	0.080	0.357	0.199	**0.314**

5.2. Personality score categories (high, medium, low)

For a three-class classification of personality score categories based on population means, we run two sets of experiments. First, we use only the lexical and acoustic-prosodic features to predict High (HI), Medium (ME), and Low (LO) for each personality traits. We then use these features combined with four other ground truth labels to predict the fifth for each trait, providing an upper bound for the performance of a multi-label prediction ensemble.

In contrast to the regression results, the individual feature sets do not perform well on each trait excepting the LLD features. For the first set of experiments, Table 4 shows that the LLD feature set performs best on Openness and Neuroticism; a combination of LLD and FFV performs best on Conscientiousness; a combination of FFV, LLD and DAL achieve the highest performance on Extraversion; and a combination of the LIWC and FFV features performs best on Agreeableness. We use unweighted average recall (UAR) to compare the results from the different sets of features, since the distribution for three classes in each trait are not balanced (Table 2). We set our baseline UAR to 33.3%, since we know of no other work using a three-way classification which might serve as a baseline; a similar experiment by [21] uses a two-class classification.

Table 4: *Classification Performance of Personality Recognition. (UAR)*

Feature	O	C	E	A	N
Baseline	33.3%	33.3%	33.3%	33.3%	33.3%
LLD	**43.8%**	38.2%	31.5%	35.7%	**42.3%**
FFV	38.5%	37.2%	29.1%	31.2%	32.4%
LIWC	42.4%	34.6%	31.9%	39.4%	30.8%
DAL	38.1%	32.1%	32.8%	36.4%	32.0%
LLD + FFV	42.8%	**39.1%**	35.5%	36.1%	41.2%
LIWC + FFV	41.4%	33.7%	30.6%	**40.5%**	32.9%
DAL + LLD + FFV	37.8%	38.7%	**37.4%**	33.0%	41.0%
Combined	38.8%	36.6%	32.3%	36.9%	41.9%
Improvement	+10.5%	+5.8%	+4.1%	+7.2%	+9.0%

For the second classification experiment, we find that the performance shown in Table 5 improves over that shown in Table 4 for every trait by at least 2.3% and at most 8.7% when adding the ground truth label of HI, ME, LOW to the lexical and acoustic-prosodic features. The LIWC features combined with the trait labels performs best on Openness; the FFV feature set combine with the trait label performs best on Extraversion; a combination of LLD and FFV features with the trait label rates highest on Neuroticism; a combination of LIWC and FFV features with trait labels reaches highest performance on Agreeableness; and a combination of LLD and DAL features with trait label performs best on Conscientiousness.

Table 5: *Classification Performance of Personality Recognition with other trait ground truth. (UAR)*

Feature	O	C	E	A	N
Baseline	33.3%	33.3%	33.3%	33.3%	33.3%
LLD	40.0%	41.9%	35.5%	36.2%	47.3%
FFV	39.1%	40.8%	**39.7%**	39.2%	40.7%
LIWC	**52.5%**	38.0%	38.7%	43.0%	36.9%
DAL	41.4%	37.7%	36.8%	40.7%	44.8%
LLD + FFV	40.8%	41.3%	34.5%	36.5%	**47.9%**
LIWC + FFV	51.2%	42.4%	37.6%	**45.3%**	39.5%
DAL + LLD	37.4%	**43.4%**	35.3%	33.4%	45.6%
Combine	40.1%	40.8%	34.5%	36.4%	45.4%
Improve	+19.2%	+10.1%	+6.4%	+12.0%	+14.6%
Table 4	+10.5%	+5.8%	+4.1%	+7.2%	+9.0%
Table 5 - Table 4	+8.7%	+4.3%	+2.3%	+4.8%	+5.6%

5.3. Personality score clusters

Finally, instead of trying to predict individual traits in isolation, here we view them together as comprising a single whole personality. Because of sparsity, we cannot consider every possible combination of high, medium and low for each of the five traits. Instead, we treat each combination in our data as a single instance and cluster them using k-means, and treat each cluster id as representative of a personality *configuration*. This id may be useful as a feature in downstream tasks for which individual personality trait features have been shown to be helpful, since it models an integrated view of the personality that is closer to how it functions in the real world. (High agreeableness, for example, is different for a high-extroversion personality than for a low-extroversion personality.)

The prediction of the five clusters improves over the baseline approximately 5-7% for different feature sets. The best performing feature set is a combination of LIWC and FFV features, and it improves over the baseline by 7.9%. We use accuracy (ACU) to compare the results from different sets of features, since the distribution for five clusters in each trait are equal, which is different from the class distribution. We set our baseline result to chance at 20.0%.

Table 6: *Five Cluster Classification Performance of Personality Recognition. (ACU)*

Feature	ACU	Improve
Baseline	20.0%	
DAL	25.2%	+5.2%
LIWC	25.5%	+5.5%
FFV	25.8%	+5.8%
LLD	27.0%	+7.0%
LIWC + FFV	**27.9%**	+7.9%
Combine	24.5%	+4.5%

6. Discussion and Conclusion

In this work, we present results showing that lexical and acoustic-prosodic features can predict self-reported personality traits identified by the NEO-FFI Five-Factor personality inventory with considerable success. We have experimented with a number of feature sets and a variety of techniques to achieve improvements significantly over our baselines.

While direct comparison to stranger-predicted ratings is unavailable, we note that the ranking of accuracy for the Big Five differs by condition. Stranger ratings of Neuroticism and Extroversion were predicted with the highest accuracy in [21], and Openness was most difficult to predict. Stranger-predicted Extroversion had highest accuracy and Openness lowest in [16]. We predict self-ratings of Openness with *highest* accuracy (Table 4), and Extroversion *lowest*. This disparity is consistent with the idea that stranger-ratings can only be based on lexical and vocal characteristics of speech samples. Big Five traits such as Extroversion and Neuroticism are popularly associated with stereotypical speech behaviors and thus may be easier to classify from isolated speech, whereas Openness to Experience is much less stereotyped in terms of speech behaviors. Thus, Openness may well be more accurately self-reported than stranger identified. So we can conclude that there are plausible explanations for the differences we find between our study of self ratings and other studies of stranger ratings.

In future work, we will experiment with different methods to classify personality scores and different feature set train the model. We will also experiment with different machine learning algorithms such as Neural Nets to train our model.

7. Acknowledgements

This work was partially funded by AFOSR FA9550-11-1-0120 and by NSF DGE-11-44155. We thank Zoe Baker-Peng, Lingshi Huang, Leighanne Hsu, Melissa Kaufman-Gomez, Gideon Mendels, Yvonne Missry, Elizabeth Petitti, Sarah Roth, Molly Scott, Jennifer Senior, Grace Ulinski, Christine Wang, and Mandi Wang for help in running the experiments.

8. References

[1] A. E. Kazdin, "Encyclopedia of psychology," 2000.

[2] P. T. Costa and R. R. MacCrae, *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources, 1992.

[3] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[4] R. R. McCrae, "Trait psychology and culture: Exploring intercultural comparisons," *Journal of personality*, vol. 69, no. 6, pp. 819–846, 2001.

[5] E. E. Noftle and R. W. Robins, "Personality predictors of academic outcomes: big five correlates of gpa and sat scores." *Journal of personality and social psychology*, vol. 93, no. 1, p. 116, 2007.

[6] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[7] M. B. Donnellan, R. D. Conger, and C. M. Bryant, "The big five and enduring marriages," *Journal of Research in Personality*, vol. 38, no. 5, pp. 481–504, 2004.

[8] P. Prinzie, G. J. J. Stams, M. Deković, A. H. Reijntjes, and J. Belsky, "The relations between parents big five personality factors and parenting: A meta-analytic review." *Journal of personality and social psychology*, vol. 97, no. 2, p. 351, 2009.

[9] S. Soldz and G. E. Vaillant, "The big five personality traits and the life course: A 45-year longitudinal study," *Journal of Research in Personality*, vol. 33, no. 2, pp. 208–232, 1999.

[10] U. Gretzel, N. Mitsche, Y.-H. Hwang, and D. R. Fesenmaier, "Tell me who you are and i will tell you where to go: use of travel personalities in destination recommendation systems," *Information Technology & Tourism*, vol. 7, no. 1, pp. 3–12, 2004.

[11] R. Hu and P. Pu, "A study on user perception of personality-based recommender systems," in *User Modeling, Adaptation, and Personalization*. Springer, 2010, pp. 291–302.

[12] E. L. Kelly and J. J. Conley, "Personality and compatibility: a prospective analysis of marital stability and marital satisfaction." *Journal of personality and social psychology*, vol. 52, no. 1, p. 27, 1987.

[13] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

[14] F. Mairesse and M. Walker, "Personage: Personality generation for dialogue," in *Annual Meeting-Association For Computational Linguistics*, vol. 45, no. 1, 2007, p. 496.

[15] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance." *Journal of Personality and Social Psychology*, vol. 62, no. 4, p. 645, 1992.

[16] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of the 2nd international workshop on Social signal processing*. ACM, 2010, pp. 17–20.

[17] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[18] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[19] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[20] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life." *Journal of personality and social psychology*, vol. 90, no. 5, p. 862, 2006.

[21] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, pp. 457–500, 2007.

[22] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: Iv. reliability, validity, and applications," *Perceptual and Motor Skills*, vol. 62, no. 3, pp. 875–888, 1986.

[23] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg, "Individual differences in deception and deception detection," 2015.

[24] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 1–8.

[25] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[26] K. Locke, "Neo scoring," 2015.

[27] H. Giles, K. R. Scherer, and D. M. Taylor, "9. speech markers in social interaction ," 1979.

[28] E. B. Mallory and V. R. Miller, "A possible basis for the association of voice characteristics and personality traits," *Communications Monographs*, vol. 25, no. 4, pp. 255–260, 1958.

[29] A. W. Siegman and B. Pope, "Personality variables associated with productivity and verbal fluency in the initial interview." in *Proceedings of the Annual Convention of the American Psychological Association*. American Psychological Association, 1965.

[30] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[31] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[32] C. D. Aronovitch, "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker," *The Journal of social psychology*, vol. 99, no. 2, pp. 207–220, 1976.

[33] B. L. Brown, "Effects of speech rate on personality attributions and competency evaluations," *Language: Social psychological perspectives*, pp. 293–300, 1980.

[34] K. Laskowski, M. Heldner, and J. Edlund, "The fundamental frequency variation spectrum," *Proceedings of FONETIK 2008*, pp. 29–32, 2008.

[35] X. Cui, B. Kingsbury, J. Cui, B. Ramabhadran, A. Rosenberg, M. S. Rasooli, O. Rambow, N. Habash, and V. Goel, "Improving deep neural network acoustic modeling for audio corpus indexing under the iarpa babel program," in *Interspeech*, 2014.

[36] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," 2014.

[37] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[38] J. W. Pennebaker, T. J. Mayne, and M. E. Francis, "Linguistic predictors of adaptive bereavement." *Journal of personality and social psychology*, vol. 72, no. 4, p. 863, 1997.

[39] W. Heller, "Neuropsychological mechanisms of individual differences in emotion, personality, and arousal." *Neuropsychology*, vol. 7, no. 4, p. 476, 1993.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.