5-2018

Intergroup Variability in Personality Recognition

Arundhati Sengupta
The Graduate Center, City University of New York

INTERGROUP VARIABILITY IN PERSONALITY RECOGNITION

by

ARUNDHATI SENGUPTA

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2018

Intergroup Variability in Personality Recognition

by

Arundhati Sengupta

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Rivka Levitan Thesis Advisor
Date	Gita Mortohardjono Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Intergroup Variability in Personality Recognition

by

Arundhati Sengupta

Advisor: Rivka Levitan

Automatic Identification of personality in conversational speech has many applications in natural language processing such as leader identification in a meeting, adaptive dialogue systems, and dating websites. However, the widespread acceptance of automatic personality recognition through lexical and vocal characteristics is limited by the variability of error rate in a general purpose model among speakers from different demographic groups. While other work reports accuracy, we explored error rates of automatic personality recognition task using classification models for different genders and native language groups (L1). We also present a statistical experiment showing the influence of gender and L1 on the relation between acoustic-prosodic features and NEO- FFI self-reported personality traits. Our results show the impact of demographic differences on error rate varies considerably while predicting "Big Five" personality traits from speaker's utterances. This impact can also be observed through differences in the statistical relationship of voice characteristics with each personality inventory. These findings can be used to calibrate existing personality recognition models or to develop new models that are robust to intergroup variability.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Rivka Levitan for her continuous guidance and support throughout the alignment project. I want to thank Guozhen An who has provided a lot of help for the project during the early stages. I am also thankful to Christen N. Madsen II for his guidance during statistical analysis of my data. I would also like to thank my honorary cohort Boram Kim and Armando Tapia for their invaluable support and advice. My deepest thanks to my parents for their constant support for my education and career; and lastly, to my best friend and husband Sarthak Banerjee, who has been my biggest advocate throughout my graduate study, I definitely could never have finished it without you.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Personality can be defined as "consistent behavior patterns and intrapersonal processes originating within the individual" (Burger, 2015). From a psychological perspective, personality is the reflection of a state of mind that varies between individuals because of biological and environmental factors (Corr et al., 2009) such as individual's political belief, career choice, physical, and mental health (Ozer et al., 2006). The NEO-FFI questionnaire, measuring the Costa & McCrae five-factor model, also known as Big Five: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, the most commonly used resource for measuring personality in nonclinical populations model for personality (Costa et al., 1992, McCrae & Costa, 2003). Several researchers have identified these traits independently (Digman, 1990) and used this model to characterize personality in multiple cultures (McCrae, 2001; Bond et al., 1975).

Some personality inventories such as the CPI, the Comrey CPS, the GZTS have always considered gender as an important feature for evaluating personality (Digman, 1990). Shafran et al. (2003) describes that the female and male speakers have different voice characteristics and they use language differently. The same is true for speakers with different native languages (L1) (Joel et al., 2013). It has been observed that different cultures and demographic groups have an influence on personality and how it is expressed in speech and language (Scherer, 1979).

In this work the main research objectives are:

1. To create a machine learning model using acoustic, prosodic, and lexical feature to compare error rates for different demographic groups (gender and L1).

2. To perform a statistical analysis to investigate the relationship between the acoustic-prosodic features and five personality traits as well as their variability across different demographic groups.
3. To demonstrate the same approach is applicable to other affective labels such as deception.

We have defined the scope of our work in several ways. The personality features for this work has been collected through NEO-FFI personality inventory which measures personality based on five traits:

- Extraversion (active, talkative, energetic)
- Agreeableness (sympathetic, generous, appreciating)
- Conscientiousness (organized, reliable, thorough)
- Openness (curious, imaginative, original)
- Neuroticism (unstable, worrying, anxious)

The demographic groups are categories based on the gender (male and female) and native languages (Standard American English and Mandarin Chinese) of the participants.

The purpose of this work is not to emphasize the accuracy of the machine learning model or improve it but to understand how a simple general purpose model behaves differently when trained and tested on data belonging to different demographic groups.

We hypothesize that there is a significant intergroup variability in the personality recognition task among different gender and L1 groups, and that this variability can be better understood by investigating the differences in the error rates of models trained and tested on various proportions of in-group and out-group data. These findings can be used to motivate and inform further development of personality recognition models that are robust to intergroup variability.

This study is based on the Columbia Cross-Cultural Deception (CXD) Corpus, a collection of task-

oriented conversational speech between Standard American English and Mandarin Chinese speakers, which is described in Chapter 3, where a subject produces multiple instances having same value for gender, L1, and five personality traits. We used the threshold provided in Mohammadi et al. (2010) to represent the self-rated NEO scores on a two-level scale (lower and upper 50 percentile).

In this work, we have experimented with two approaches to understand how gender and L1 group impacts the error rate in personality detection task.

To fulfill our objectives, we have engaged in the following research activities:

1. Trained and tested different machine learning models for various subgroups of data as mentioned below, and compared the prediction error rate (F1) as an indicator of intergroup variability in the expression of each personality trait.
    a. A machine learning model trained and tested on gender with varying percentage of in-group and out-group instances.
    b. Identified the variability in error rate for different L1 groups while trained on gender-specific data.
2. Performed a statistical analysis to see how the correlation between the prosodic features extracted from the corpus and different personality traits varies for different gender and L1 group.

Chapter 2

Related Work

There have been a number of studies, mostly conducted by psychologists, and more recently by computer scientists, which try to identify personality automatically from text as well as from speech. Lexical features are considered a reliable indicator for personality identification. Pennebaker et al. (1999) showed the correlation between lexical features represented as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) and five personality factors. The same study reported that the personality trait "openness" is negatively correlated with the words in the "immediacy" category, and the prevalence of making distinctions in writing is negatively correlated with the "extraverted" and "conscientiousness". They found that "neuroticism" is positively correlated with the use of negative emotion words and negatively correlated with positive emotion words, and vice versa for "agreeableness" and "extraversion".

Argamon et al. (2005) argues that function (non-content) words, because of their high frequency and high importance for grammatically correct sentences, are not under conscious control by the author. Therefore, their frequency and interdependence represent a style which can be used to recognize personality. In their work, they have extracted four lexical stylistic features from essays written by the students and classified the authors as high or low for extraversion and neuroticism based on those features using an SMO classifier.

In another study, Mehl et al. (2006) described that the LIWC categories related to spoken dialogue was found to have a correlation with agreeableness, extraversion, and conscientiousness. Conscientiousness was negatively correlated with the use of swear words.

In one of the earliest study on relationship between voice characteristics and personality traits

Mallory and Miller (1958) found a significant positive association between dominance (extraversion) and loudness, intensity, and pitch; and submissiveness has a positive association with speech rate. Smith et al. (1975) performed a statistical study on the relationship between speech rate and personality traits. They found that perceived competence (conscientiousness) has a positive correlation with speech rate. Whereas, benevolence (agreeableness) has a non-linear correlation with speech rate.

In a personality recognition study, Mairesse et al. (2007) used LIWC, psycholinguistics, utterance type, and prosodic feature sets in classification, regression, and ranking algorithms to predict self and observer personality scores from essay and conversational data. The study revealed that the observer reported personality scores are predicted more accurately than self-reported personality scores. Oh et al. (2011) analyzed the FFM (Five Factor Model) traits based on self and observer rating, posit that observer rating is significantly more valid who know the subject well. Borkenau & Liebler (1992) asserted that there is a weak correlation and moderate to weak internal consistency (as measured by Cronbach's alpha) between the strangers rating and self-reports.

In another study Mohammadi et al. (2010) used prosodic features to study personality from short ten-second audio clips labeled by human judges with observer personality scores. They performed a SVC binary classification splitting each trait into high and low based on the average of the NEO-FFI scores and achieved good accuracy for extraversion.

Levitan et al. (2016a) used acoustic-prosodic and language features to predict self -reports of personality scores. They had used a novel psychologically motivated approach of labeling each personality trait with three different levels: High (HI), Medium (ME) or Low(LO). They compared the performances of models using different combinations of feature sets, and showed that different models performed best for each personality trait. In a recent study, An and Levitan (2018)

experimented with two different methods for mitigating intergroup variability (partitioning the data into homogeneous models and normalization) for three-level classification of each personality trait. They found when the data is sub-grouped into homogeneous models, the performance was better in predicting openness and conscientiousness.

Studies have shown that manifestations of personality are influenced by gender and culture differences. Scherer (1979) showed that the use of nasal and louder voice is perceived as extraverted, and American extraverts tend to make fewer pauses while speaking. In contrast, German extraverts produce more pauses than introverts. Thus personality markers are culture-dependent, even among western societies. Stewart (1998, 2004) argued that the "big five" personality traits are associated with masculinity and femininity. For example, extraversion which is related to dominance is more commonly seen in males and agreeableness in females. Further, he suggested that the underlying facets of these big five traits are aligned to certain gender -e.g. "openness to experience" (an indicator of tolerance) could be allied with gender female, whereas openness to risk-taking related with gender male. In a study between a group of male and female student (Mehl et al., 2006) found that some linguistic cues vary significantly across gender. For example, males who are marked high in conscientiousness use more filler words, while females don't.

In this thesis we tried to find the intergroup variability of the error rate for different demographic groups (gender/L1) using self-reports. Additionally, we perform a classification where the scores of each trait are split into High and Low subsets. The latter includes the samples that have scores lower than the median, while the former includes the samples with the scores higher than the median. The median of personality traits is separately evaluated for male and female participants.

Chapter 3

Data and Features

3.1 Columbia Cross-Cultural Deception Corpus:

To investigate the individual and cross-cultural differences and similarity in personality recognition, we have used the Columbia Cross-Cultural Deception (CXD) Corpus (Levitan et al., 2015a) which consists of within-subject deceptive and non-deceptive English speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC). Native language is defined as the language spoken at home until age 5. The corpus includes 122.5 hours of speech data from 134 subject pairs and 268 individual speakers, who are not previously familiarized with one another. This is currently the largest corpus of deceptive speech.

The data was collected using a fake resume paradigm in which pairs of participants played a lying game. Subjects took turns interviewing their partner and being interviewed from a set of 24 biographical questions such as “What is your mother’s job?”, the false answer must be different from their mother’s or father’s true occupation, and the answer for “where were you born?” must be a place where the subject has never visited. Subjects were asked to tell a lie as an answer to a predetermined subset of questions. Before the recording of the interview, the false answers were checked by the experimenter to ensure that the participants followed the guidelines. The interview was recorded in a soundproof booth where subjects were seated crosswise from each other and separated by a curtain to avoid visual contact. Close-talk headsets were used for recording the conversation (Levitan et al., 2016a; An et al., 2016). Also, a 3-4 minute baseline sample from each subject was collected. The participants were asked some open-ended question e.g. ‘what do you like the best/worst about living in NY?’ and the participants were asked to be truthful in answering

(Levitan et al., 2015a; An et al., 2016).

3.2 Transcription of the data

Transcripts of the recorded data were acquired using Amazon Mechanical Turk[1] (AMT). For each audio segment, three transcripts were obtained from different 'Turkers'. After that, the transcripts were merged using rover techniques (Fiscus, 1997) which produce a rover output score measuring the inter-annotator agreement. Transcripts were manually corrected if the score was lower than 70% (9.7% of the clips).

3.3 Unit of Analysis

The unit of analysis in our experiment is the *turn*. A *turn* is defined as a maximal sequence of inter-pausal units (IPU) from a single speaker. Inter-pausal units (IPU) are defined as a maximal sequence of words surrounded by silence by at least 50ms (Gravano and Hirschberg, 2011). Turn boundaries were extracted in the following manner: the audio was force-aligned with the manual orthographic transcription and the speech was segmented if there was more than a 0.5 seconds of silence (An and Levitan, 2018). There were a total 29175 turn-level instances and the average duration of each instance is 9.03s, though, there were quite a few outliers.

3.4 Personality Assessment

Before the interview begins, demographic data from each subject was collected and subjects filled out the NEO-FFI (Five Factor) personality inventory (Costa et al., 1992, McCrae & Costa, 2003). Factor analysis was performed on thousands of descriptive term in a standard English dictionary to develop NEO-FFI inventories. It is used to assess the five personality dimensions, namely:

[1] https://www.mturk.com

· Openness to Experience (O). Captures imagination, creativity, intellectual curiosity. It is "related to aspects of intelligence, such as divergent thinking, that contribute to creativity" (Costa et al., 1992). Individuals with low openness are known as closed-minded and behave orthodoxly. Conversely, those who score high are "willing to entertain novel ideas and unconventional values" (Costa et al., 1992).

· Conscientiousness (C). Aim to capture human traits such as plan and carry out task and controlling impulsive behavior. It measures the contrasts between determination vs. weakness, organization vs. clumsiness, and self-discipline vs. carelessness.

· Extraversion (E). Designed to assess the tendency for interpersonal interactions, and variation in sociability. It is intended to capture the contrasts between those who are self-absorbed vs. those who are outgoing, quiet vs. talkative, and disinterested vs. active.

· Agreeableness (A). Designed to measure one's cooperative and trusting nature. Those who score high on this dimension are empathetic and altruistic and expect that others feel similarly.

· Neuroticism (N). Quantifies the mental stability of a person. Individuals having high score on neuroticism are expected to be more moody than average person and more likely to experience feelings like anxiety, worry, and fear.

The questionnaire contains 60 questions[2] and for each question, the subject responded on a five-point Likert scale: strongly agree, agree, neutral, disagree, and strongly disagree.

Table 1.1 shows the Pearson Correlation (95 percent confidence interval) between the personality traits performed on the raw scores and showing that there are no strong correlations between personality traits.

[2] http://www.cs.columbia.edu/speech/cxd/forms/NEO-FFI_English&Chinese.pdf

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Openness	-	- 0.1*	0.1*	0.18***	0.24***
Conscientiousness	-	-	0.26***	0.22***	-0.30***
Extraversion	-	-	-	0.27***	-0.32***
Agreeableness	-	-	-	-	-0.17**
Neuroticism	-	-	-	-	-

Table 1.1 Pearson Correlation (with 95% CI) between the personality traits

Signif. codes: '***' p<0.001 '**' p<0.01 '*' p<0.05 '.' P<0.1

3.5 Feature extraction

3.5.1 Acoustic and Prosodic Features

Previous researches (Scherer et al., 1979; Mallory & Miller, 1958; Siegman & Pope, 1965; An et al., 2016) showed that different speech factors such as fundamental frequency, voice quality, intensity, duration of silence can be used to predict different personality traits. Inspired by these findings, for this study we used the OpenSMILE library to extract acoustic-prosodic features (Eyben et al., 2013). The OpenSMILE Low-Level Descriptor (LLD) feature set contains 11 acoustic-prosodic features. These include pitch (fundamental frequency), intensity (root mean square energy), voice quality (jitter, shimmer, and harmonics-to-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness (Schuller et al., 2009). Various statistics (mean, minimum, maximum, skew etc.) are applied to each of the low-level features, for a total of 384 acoustic/prosodic features.

3.5.2 Linguistic Inquiry and Word Count Features

People choose different words not only for linguistic meaning but also to convey different psychological /emotional conditions (Mohammadi and Vinciarelli, 2014). Therefore, by using different psycholinguistic methods we can predict personality through text or speech analysis.

Motivated by Pennebaker et al. (1999) and Mairesse et al. (2007), we used Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) to extract the lexical features of the spoken data. LIWC is a text analysis program that classifies words in psychologically meaningful categories. Each of these categories represents various emotional, cognitive, and structural traits which the underline word represents e.g. 'love', 'nice' and 'sweet' belongs to the positive emotion category and 'hurt', 'ugly', and 'nasty' belongs to negative emotional categories. Previously, in many studies LIWC features have been used to detect deception (Newman et al., 2003; Levitan et al., 2016a, 2016b), personality (Pennebaker et al., 1999; Mairesse & Walker 2006), and health (Pennebaker et al., 1997; Huh et al., 2013). We extracted 130 features based on the 64 LIWC categories.

3.5.3 Dictionary of Affect Features

Heller (1993) stated that there is a high correlation between the arousal of some dimensions of personality, specially extraversion. We used Whissell's Dictionary of Affect in Language (DAL) (Whissell et al., 1986) to extract additional features for detecting the personality. The DAL is a lexical analysis program which is originally designed to quantify undertones (connotations and associations) of emotional words especially pleasantness, activation, and imagery. It contains approximately 8742 English words, each with ratings for these three categories obtained from several human judges (Whissell, 2009). From each participant's baseline interview transcript, we extract nineteen features and calculated the mean, minimum, maximum, median, standard deviation, and variance of the pleasantness, activation, and imagery scores for all words found in the DAL.

Chapter 4

Methodology

In this chapter, we will outline the classification experiments and statistical analysis processes. We performed these analyses with respect to personality labeling of the data using turn as our unit of analysis.

4.1 Preprocessing of the NEO-FFI score

For this experiment, we used two levels of personality scores which is determined by checking if a score is greater than median value across all participants or not. While reviewing the personality scores for gender groups, we observed that the median value is not same for Agreeableness, Extraversion, and Neuroticism in case of male and female. Therefore, we preprocessed the data to come up with alternate personality scores for these features depending on the gender group the speaker belongs to.

4.2 Classification Experiments

Columbia Cross-Cultural Deception (CXD) Corpus includes English speech from approximately equal amount of female and male Standard American English (SAE) and Mandarin Chinese (MC) speakers. It has been observed the acoustic-prosodic feature values have different ranges for male and female gender group e.g. the pitch of a voice changes according to the length of the speaker's vocal tract, that's why female voice has higher pitch value than male. In the CXD Corpus, the mean and standard deviation for female pitches are 73.89 and 46.43, while male pitches are 18.28

and 24.97 respectively. In order to perform meaningful comparisons across gender groups, we used *z-score*[3] normalization on a given speaker's features using the mean and standard deviation derived from gender groups that the speaker belongs to.

In another approach to experiment with normalization, we normalized each feature value across all the samples for the same speaker. Our motivation was to remove any biases introduced by the speaker in the data. However, this doesn't contribute significantly toward the performance of the model and therefore, was not included in final experiment set up.

To investigate the intergroup variability in the expression of personality, we created machine learning models trained and tested on a different group of data, partitioned by gender / L1 and compared the F1-value[4] for the personality prediction task among them. We used two primary approaches to create the train and test groups. In our first approach (Figure 4.1), we tried to explore the variances in the F1 value of models while trained on gender specific data and tested on various proportions of in-group and out-group data. We partitioned the data into training and test set with the ratio of 70:30. We split the training data into two groups based on speaker's gender. The female and male specific train set has further been balanced with same number of instances and approximately same number of speaker's data. The test set is transformed to create five versions with varying percentages (0, 25, 50, 75, and 100) of in-group and out-group instances. We refer in-group instances as instances belonging to the speaker having same gender as the training data and out-group as different. Further, we tested the five version of test data on the machine learning

[3] The formula for z-score normalization is
$Normalized\ (e_i) = \frac{e_i - \bar{\mathrm{E}}}{std(\mathrm{E})}$

[4] $F1 = 2.\frac{precision.recall}{precision+recall}$ $precision = \frac{tp}{tp+fp}$ $recall = \frac{tp}{tp+fn}$
$tp = True\ Positive, fp = False\ Positive, fn = False\ Negative$

models build using SVM (Support Vector Machine), trained on gender specific data for each personality trait and reported the F1 values for personality prediction task.

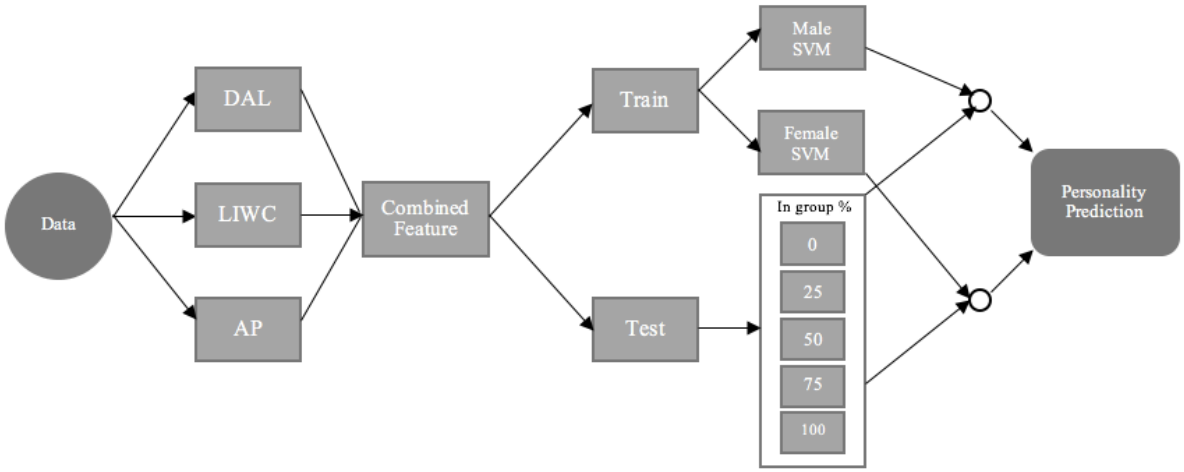

Figure 4. 1 Model 1: Trained on gender specific data and tested on various proportions of in-group and out-group data

In second approach, we trained the model on the gender specific data and tested on in-group data split by L1 (Figure 4.2). To perform this experiment, we divided the data into male and female. Following this, we partitioned each gender specific data into train and test set split in a ratio of 70:30. Then we split the data base on L1 of the speaker (SAE and MC). At this point, we created two machine learning models using SVM trained on gender specific data for each personality group. Further, we tested this model separately for SAE and MC speakers' data to report the F1 value for predicting personality.

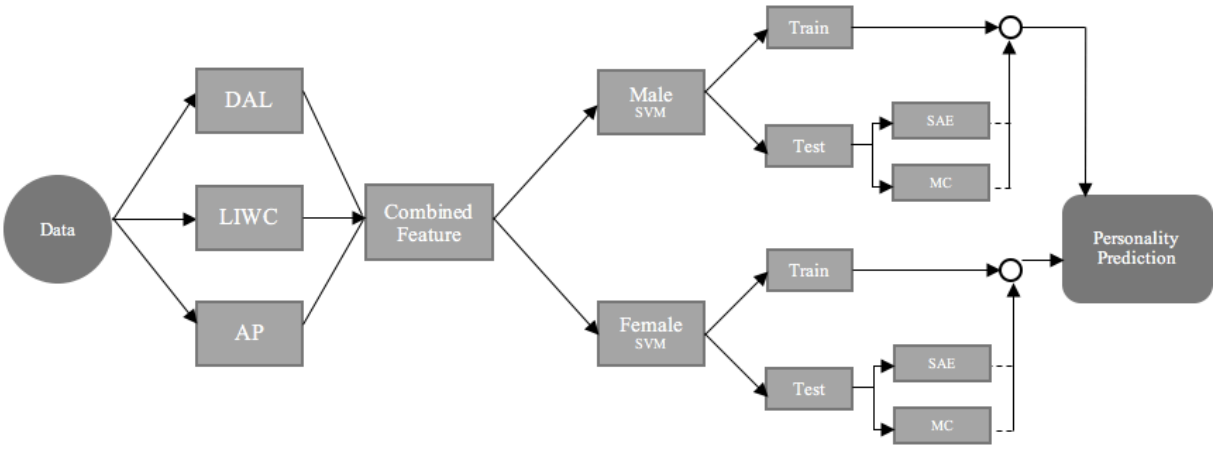

Figure 4. 2 Model 2: Trained the model on the gender specific data and tested on in-group data split by L1

We have used sklearn SVC[5] module for our machine learning experiments. The SVC kernel used for our experiment is 'rbf' with C value of 1 and cache size = 2000. One of the problem with SVC classifier is that it takes longer time to train when number of instances exceeds 10,000 as mentioned in sklearn documentation. As we have much larger volume of instances, we used sklearn's BaggingClassifier which partitions the instances and train the model in parallel on a multiprocessor system.

[5] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

4.2.1 Baseline

Since we know of no other work using error rate to see the intergroup variability in two-way classification for personality recognition task, we train our baseline model for these experiments by combining AP, LIWC, and DAL feature sets for all genders and L1 groups.

4.3 Statistical Experiment

Motivated by earlier study on prosodic features and personality (Scherer, 1978), we wanted to explore the relationship between prosodic features (pitch and intensity) and the personality traits for different gender / L1 groups. The prosodic features we used are as follows.

Pitch. The fundamental frequency of a voice represents pitch which describes how often the sound wave repeats itself. The maximum, mean, and range of pitch was extracted using openSMILE. For statistical analysis, it is important to use data that is balance in order to make meaningful comparisons between groups we normalized all pitch features using z-score normalization for speakers of the appropriate gender.

Intensity, also known as amplitude or energy, describes the degree of energy in a sound wave and is perceived as the loudness of a sound. We look at the mean, maximum, and range of the intensity of the speech segment, extracted using openSMILE. As with pitch, we scaled the values using z-score normalization.

To analyze the intergroup variability, we performed two kind of statistical analysis. In the first condition, we built separate models for male, female, and all speakers with six acoustic-prosodic features to understand how gender influences the relationship between personality traits and vocal features. In the second condition, we enhanced the first model to examine the interaction of L1 with each of the six acoustic-prosodic features mentioned above. In this way, we will be able to

understand how prosodic features and different personality traits vary for different gender and L1 groups.

Before building the model we checked whether or not the features are collinear using variance inflation factor score (VIF) and found that the VIF score for all features are less than 10 (Myers, 1990). In our data, we have both fixed and random effects, but the mixed-effects models couldn't converge due to split in the data. Therefore, we used multiple regression (lme4 by Maechler & Bates 2010) using R (R Development Core Team, 2017) to understand the intergroup variability in prosodic features and different personality traits. We have added gender and language marker fields in the acoustic-prosodic data to perform the regression test. The gender marker field is set to 1 for all instances of a male and 0 for female speaker. Similarly, the language marker field (L1) is set to 1 for Standard American English (SAE) and 0 for Mandarin Chinese (MC) speakers. The gender field was used to partition the data into male and female groups, whereas, L1 field was included in regression experiment as an interaction on intensity and pitch features.

Chapter 5

Results and Discussion

The analysis presented here is intended to quantify the influence of intergroup variability on personality prediction, rather than the accuracy of the personality prediction itself.

5.1 Machine Learning based Classification Result

In the machine learning based approach we evaluate the F1 scores for binary classification models based on the CXD corpus with self-reports of different personality traits for the subjects. We hypothesize that models trained on greater percentages of in-group data will perform better.

5.1.1 Baseline

These scores represent the F1 for the baseline model when trained on 70% of baseline train data and 30% of baseline test data (shown in Table 5.1). A high F1 value indicates that the model performance is good and the error rate is low, whereas the lower value implies the opposite.

	O	C	E	A	N
All	0.54	0.47	0.51	0.47	0.47

Table 5. 1 Baseline F1 for each trait

5.1.2 Train on Male/ Female, test on different proportion of the data

Figure 5.1 shows the result of our first experiment on machine learning approach when the model is trained male instances and Figure 5.2 shows the result when the model is trained on female instances. The result supports our hypothesis that the error rate for personality recognition task varies across

different gender based on the percentage of in-group and out-group instances.

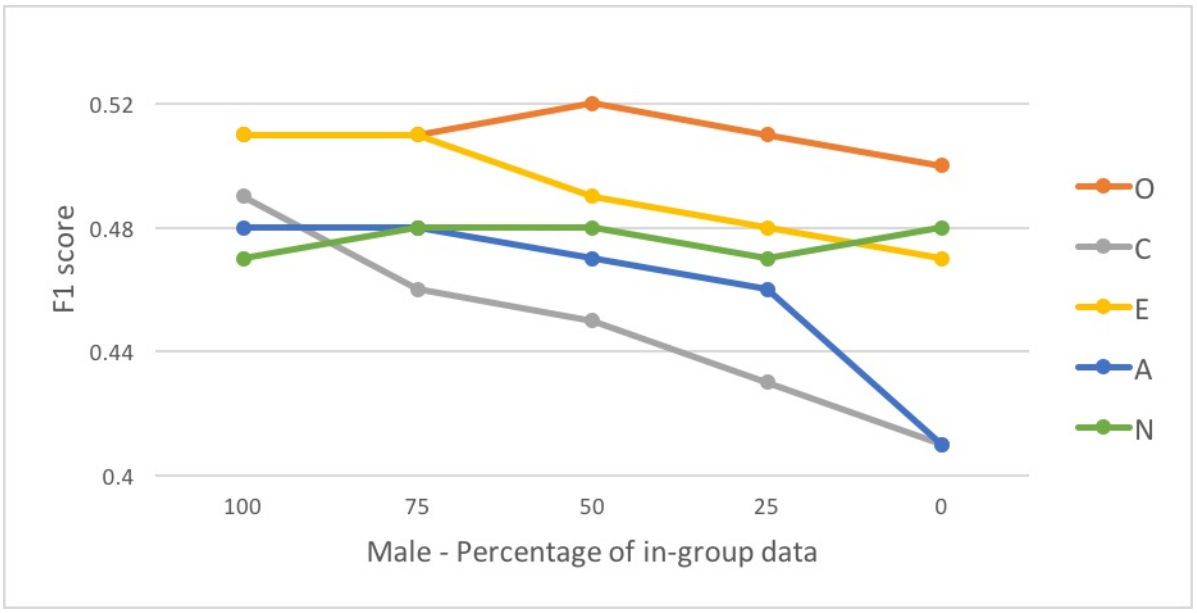

Figure 5. 1The relationship between in-group test data percentage and F1 score for each personality trait when train on Male instances.

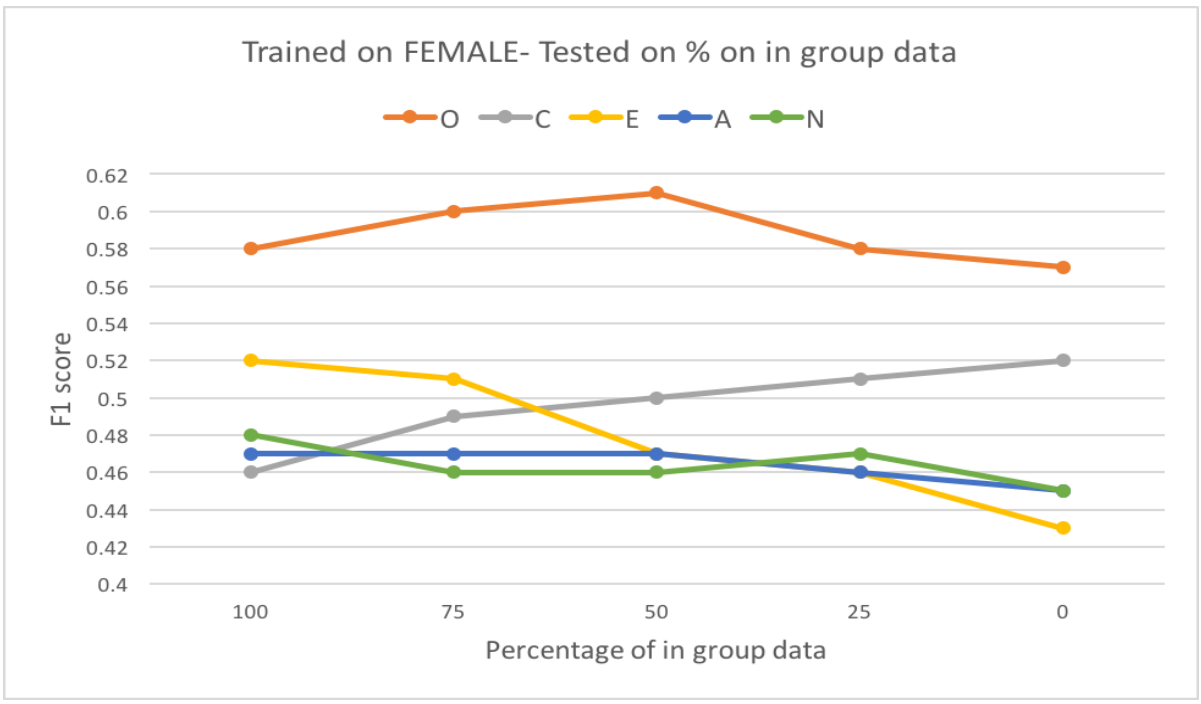

Figure 5. 2The relationship between in-group test data percentage and F1 score for each personality trait when train on Male instances.

For all five personality inventories, the models tested with 100 percent in-group instances performed better than the baseline in either male or female group. Similarly, the model trained on in-group

instances and tested on out-group instances performed worse than the baseline.

The impact of out-group test instances for test data is important for Conscientiousness and Agreeableness measures when trained on all male instances. For Extraversion when tested on 100 percent in-group instances for both gender group the model performed all most similar to the baseline. However, testing the same model with all (100%) out-group variables, it performed worse (about 8 percentage point for female group and 4 percentage point for male group).

Another important way to understand the variability is to evaluate the standard deviation for each personality scores across the five observations. A higher score indicates the measure are more prone to error where the test data and the train does not belong to same gender group. Based on the result, we found that when trained on the female group Extraversion have a higher standard deviation (0.024 and 0.037) and similarly, for male group Conscientiousness and Agreeableness reports high standard deviation (0.030 and 0.029). Conversely, while predicting Openness for male group across various ratios of out-group instances we observed the minimum standard deviation.

Based on the on the observation above, in summary, we suggest that the mixture of in-group and out-group instances does have an impact on the error rate and the models tend to perform better when trained and tested on the same group. However, this effect is not pronounced for some personality traits, which may be subject to less inter-group variability in their expression.

5.1.3 Train on gender and test on one L1

For the second classification experiment, we found that (the result is shown in Table 5.2) the model performance for predicting personality varies from baseline while trained on gender and tested on different L1 groups.

	Train on Gender	Test on L1	
		English	Mandarin
O	Male	0.38	0.49
	Female	0.50	0.53
C			
	Male	0.42	**0.56**
	Female	**0.53**	0.44
E			
	Male	0.52	0.48
	Female	0.52	0.54
A			
	Male	0.51	0.46
	Female	0.52	0.48
N			
	Male	**0.64**	0.47
	Female	0.51	0.47

Table 5. 2 F1 scores when train on gender and test on L1 groups

From the result, we note that for Neuroticism, the model train on male and tested on male SAE performed much better (0.64) than the baseline whereas the same group performed worst (0.38) while predicting Openness. For both genders the result of SAE outperformed MC group in case of Agreeableness and Neuroticism. SAE group and MC group have a large variance (0.42 and 0.56 respectively) compared to the baseline when tested for male speakers for Conscientiousness. We also observed the similar variance, but on opposite direction (0.53 and 0.44 respectively) when trained on female instances in comparison with baseline, although, the difference is less compared to the male group. The variance in predicting personality between male SAE and MC group is high for Openness and Neuroticism where they vary about 13 and 11 percentage point respectively. The result for Extraversion and Agreeableness across gender and L1 groups varies the least (SD = 0.03) and closest to the baseline which suggests that the task for predicting Extraversion and Agreeableness is least affected by cultural or gender differences. Conversely, the results suggest that the expression of

Neuroticism is highly impacted by cultural differences. This difference varies largely between different gender group only for SAE (0.64 vs. 0.51) speakers, although, it is unaffected by gender groups for MC speakers (0.47).

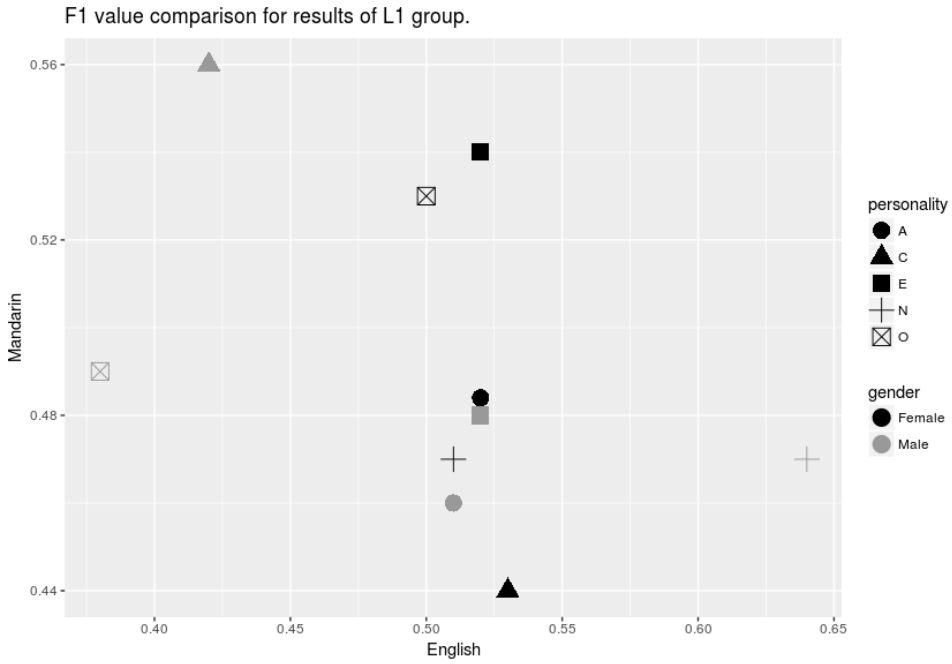

Figure 5. 3 F1 value comparison for L1 groups

The Figure 5.3 plots the result shown in Table 3 where the results for SAE and MC test instances are drawn in x and y axis respectively. The chart uses black and gray colors to distinguish between model train on male and female data and five different shapes representing different personality traits as mentioned in the legend. The interesting point in the chart are the ones which are away from the center. These points represent an error rate which is high for one of the language and low for the other. For example, the gray triangle at the top left corner of the chart represents the F1 value of Conscientiousness in male group which was high for Mandarin Chinese (0.56) but low for SAE (0.42). It implies the variability for predicting 'Conscientiousness' for different language groups.

The analysis shows that models trained on specific gender groups have high variability across L1 groups and therefore, the performance of predicting personality based on acoustic-prosodic and

lexical features, will vary for different cultural groups even when we have gender-specific models.

5.2 Statistical Analysis Result

We performed statistical analysis to understand the relationship between prosodic features (pitch and intensity) and the personality traits for different gender / L1 groups. Our first experiment shows the relationship between these features and personality traits for subsets of the data that are homogeneous with respect to gender (all-female or all-male). We also included the result for group which includes instances for male and female (marked as All) speakers. This could act as a baseline when reviewing the result for gender specific homogeneous groups. We further enhanced this model to include language effect on these six prosodic features to understand whether the relationship varies significantly for different language groups.

5.2.1 Openness

Table 5.3 shows the multiple regression results of our first experiment for Openness which suggests almost all intensity features and their interaction with L1 have a significant relationship with this personality trait, but pitch is only evident to be significant for female groups. The effect of range in intensity was found to be stronger for female group than male in a positive direction which suggest an increased value in this feature result in higher probability of the subject having a high score for Openness. L1 interaction with intensity range suggests that the SAE female speakers are more likely to score high in Openness than MC female speakers. The max value of intensity is also an effective predictor, similar to the range, but in the opposite direction. We found that the reduction in max value for intensity increases the chances of a subject being open, especially for female or SAE groups, in almost the same magnitude as observed for intensity range.

Pitch is found to be more significant ($p < 0.001$) for female speakers than male speakers while predicting Openness through prosodic feature. However, when we include the interaction with L1 it

became significant for male group as well. The range attribute of pitch is found to be an important positive indicator of Openness specially for male SAE group but not for others.

	Openness		
	Male	Female	All
Features			
Intensity range	4.04***	8.31***	6.05 ***
Intensity max	-3.99 ***	-8.38 ***	-6.05 ***
Intensity mean	0.09***	-0.02	0.03 ***
Pitch mean	0.001	-0.3 ***	-0.14 ***
Pitch range	-0.08	-0.88 **	-0.32
Pitch max	0.22	1.00 ***	0.44
Language Interaction with Prosodic Features			
Intensity range	3.29	5.32 ***	4.41 ***
Intensity max	-3.33 ***	-5.43 ***	-4.46 ***
Intensity mean	0.04 ***	-0.12 ***	-0.03 *
Pitch mean	-0.03 **	-0.40 ***	-0.22 ***
Pitch range	1.01 **	-0.93 *	-0.27
Pitch max	-0.80	1.02 **	0.40
Intensity mean: L1	0.11	0.18 ***	0.13 ***
Pitch mean: L1	0.09 ***	0.22 ***	0.16 ***
Pitch range: L1	-3.38 ***	0.07	-0.19
Pitch max: L1	3.23	0.00	0.18
Intensity range: L1	1.64	7.17 ***	3.60 ***
Intensity max: L1	-1.31	-7.08 ***	-3.46 ***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Table 5. 3 Regression result for Openness. L1 is language feature. The L1 interaction value indicates the amount by which SAE speaker's magnitude differs from MC speakers. Green cells show significantly positive β value and red cells indicates significantly negative β value. Pale color cells indicate low magnitude and brighter indicates high magnitude relationship.

5.2.2 Conscientiousness

We observed that the intensity is a significant ($p < 0.001$) indicator for Conscientiousness (shown in Table 5.4) in case of male speakers which suggest that male speaker with shorter intensity range and high maximum intensity value is likely to be more Conscientious than others. From the language interaction with the intensity features, we found a similar relationship with female speakers, but in opposite direction. L1 also plays an important role in the above mentioned observation where the result for SAE and MC groups changes significantly ($p < .001$) with high magnitude. These changes suggest that SAE for both gender group are more likely to be Conscientious with a higher value of intensity max and a lower value of intensity range.

Mean pitch value has a small but significant ($p<0.001$) relationship with Conscientiousness across gender groups. The L1 interaction with pitch mean also suggests this observation is more evident for MC group than SAE across gender.

	Conscientiousness		
	Male	Female	All
Features			
Intensity range	-12.37 ***	0.39	5.27 ***
Intensity max	12.59 ***	-0.33	-0.09 ***
Intensity mean	-0.19 ***	-0.01	-5.14 ***
Pitch mean	0.16 ***	0.08 ***	0.12 ***
Pitch range	0.73	-0.52†	-0.19†
Pitch max	-0.81	0.49†	0.13
Language Interaction with Prosodic Features			
Intensity range	-7.49 ***	8.14 ***	0.63
Intensity max	7.73 ***	-8.11 ***	-0.49
Intensity mean	-0.19 ***	0.03†	-0.08 ***
Pitch mean	0.37 ***	0.13 ***	0.23 ***
Pitch range	1.85	-0.71†	-0.07
Pitch max	-1.97	0.64†	-0.03
Intensity mean: L1	-0.02	-0.09 ***	-0.05 **
Pitch mean: L1	-0.38 ***	-0.09 ***	-0.22 ***
Pitch range: L1	-0.31	0.49	-0.08
Pitch max: L1	0.37	-0.40	0.16
Intensity range: L1	-10.85 ***	-18.52 ***	-15.03 ***
Intensity max: L1	10.81 ***	18.57 ***	15.02 ***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Table 5. 4 Regression result for Conscientiousness. L1 is language feature. The L1 interaction value indicates the amount by which SAE speaker's magnitude differs from MC speakers. Green cells show significantly positive β value and red cells indicates significantly negative β value. Pale color cells indicate low magnitude and brighter indicates high magnitude relationship.

5.2.3 Extraversion

For Extraversion, pitch and intensity mean found to have a small but significant relationship for both gender groups. The mean pitch value has small magnitude relationship with Extraversion which does not change much when we consider the influence of gender and L1 groups. However, intensity with language interaction has a significantly high influence on Extraversion (shown in Table 5.5). From the results, we observed that male SAE subjects with high intensity range are more likely to be Extrovert.

	Extraversion		
	Male	Female	All
Features			
Intensity range	0.71	0.39	-0.48
Intensity max	-0.58	-0.34	0.16
Intensity mean	0.17 ***	0.15 ***	0.57 ***
Pitch mean	-0.05 ***	-0.03 ***	-0.04 ***
Pitch range	0.62	-0.45	-0.44
Pitch max	-0.66	0.40	0.40
Language Interaction with Prosodic Features			
Intensity range	-2.76 **	-1.41†	-1.73 **
Intensity max	2.91 **	1.44†	1.84 **
Intensity mean	0.25 ***	0.04†	0.13 ***
Pitch mean	0.09 ***	-0.11 ***	-0.01
Pitch range	1.49	-0.14	0.22
Pitch max	-1.54	0.10	-0.28
Intensity mean: L1	-0.14 ***	0.27 ***	0.07 ***
Pitch mean: L1	-0.31 ***	0.18 ***	-0.07 ***
Pitch range: L1	-1.11	-0.86	-1.66 **
Pitch max: L1	1.15	0.84	1.69 **
Intensity range: L1	7.74 ***	4.11 **	5.40 ***
Intensity max: L1	-7.84 ***	-4.07 **	-5.44 ***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Table 5. 5 Regression result for Extraversion. L1 is language feature. The L1 interaction value indicates the amount by which SAE speaker's magnitude differs from MC speakers. Green cells show significantly positive β value and red cells indicates significantly negative β value. Pale color cells indicate low magnitude and brighter indicates high magnitude relationship.

5.2.4 Agreeableness

Subject with higher Agreeableness tends to show lower intensity range and higher intensity max value in their prosodic features (shown in Table 5.6). These observations vary significantly with high magnitude for different gender groups. An SAE male speaker with higher max intensity value with and short intensity range are likely to be more Agreeable. Conversely, the pitch doesn't seem to have high impact in determining Agreeableness personality of a subject across gender/ L1 groups.

	Agreeableness		
	Male	Female	All
Features			
Intensity range	-2.16 ***	-1.94 ***	2.02 ***
Intensity max	2.16 ***	1.92 ***	0.04 ***
Intensity mean	0.08 ***	0.01	-2.03 ***
Pitch mean	-0.16 ***	-0.17 ***	-0.16 ***
Pitch range	-0.49	-0.52†	-0.52†
Pitch max	0.52	0.62	0.58 ***
Language Interaction with Prosodic Features			
Intensity range	3.84 ***	-6.49 ***	-1.28 *
Intensity max	-3.74 ***	6.48 ***	1.33 *
Intensity mean	-0.01	-0.19 ***	-0.09 ***
Pitch mean	-0.07 ***	-0.10 ***	-0.08 ***
Pitch range	1.87†	0.23	0.14
Pitch max	-1.83†	-0.12	-0.07
Intensity mean: L1	0.20 ***	0.40 ***	0.29 ***
Pitch mean: L1	-0.19 ***	-0.16 ***	-0.18 ***
Pitch range: L1	-4.80	-1.80 **	-1.49 **
Pitch max: L1	4.79	1.80 **	1.48 **
Intensity range: L1	-14.43 ***	10.12 ***	-2.14 *
Intensity max: L1	14.21 ***	-10.16 ***	2.01 *

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Table 5. 6 Regression result for Agreeableness. L1 is language feature. The L1 interaction value indicates the amount by which SAE speaker's magnitude differs from MC speakers. Green cells show significantly positive β value and red cells indicates significantly negative β value. Pale color cells indicate low magnitude and brighter indicates high magnitude relationship.

5.2.5 Neuroticism

For both gender intensity range and max values are significant ($p < 0.001$) and high magnitude indicator (shown in Table 5.7) for Neuroticism. For male subjects, L1 variation doesn't play an important role while deciding a subject's Neuroticism level based on intensity features. However, for female subjects, it does play a significant role (female SAE with high-intensity range and low max intensity value likely to be more Neurotic). Comparatively, pitch plays a less important role in this scenario. Although, L1 interaction with pitch features suggests that the male SAE subjects with higher mean pitch value and SAE female subjects with lower mean pitch value are more likely to be Neurotic.

	Neuroticism		
	Male	Female	All
Features			
Intensity range	10.69 ***	6.67 ***	-7.89***
Intensity max	-10.64 ***	-6.66 ***	0.03**
Intensity mean	-0.11 ***	0.18 ***	7.92***
Pitch mean	-0.06 ***	0.04 ***	-0.01*
Pitch range	-0.64	-0.22	-0.44
Pitch max	0.67	0.15	0.43
Language Interaction with Prosodic Features			
Intensity range	10.35 ***	3.80 ***	5.02 ***
Intensity max	-10.42 ***	-3.79 ***	-5.06 ***
Intensity mean	-0.05 **	0.30 ***	0.09 ***
Pitch mean	-0.33 ***	0.18 ***	-0.06 ***
Pitch range	-2.78 *	0.08	-0.72 **
Pitch max	2.85 *	-0.21	0.71 **
Intensity mean: L1	-0.14 ***	-0.21 ***	-0.14 ***
Pitch mean: L1	0.53 ***	-0.32 ***	0.11 ***
Pitch range: L1	2.92	-0.77	0.65
Pitch max: L1	-3.01	0.91	-0.63
Intensity range: L1	-0.06	10.55 ***	7.55 ***
Intensity max: L1	0.42	-10.55 ***	-7.38 ***

Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '†' 0.1

Table 5. 7 Regression result for Neuroticism. L1 is language feature. The L1 interaction value indicates the amount by which SAE speaker's magnitude differs from MC speakers. Green cells show significantly positive β value and red cells indicates significantly negative β value. Pale color cells indicate low magnitude and brighter indicates high magnitude relationship.

The statistical analysis implies that each acoustic-prosodic features relate differently with each personality traits across L1 and gender groups. In most of the cases, these relationship is statistically significant ($p < 0.001$). These noteworthy differences in the relationship could not be reflected or exploited in a general purpose model that does not have access to demographic information.

Chapter 6

Deception Detection

In this chapter, we demonstrate that the intergroup variability is not only applicable for personality prediction but also for other affective labels. For this experiment, we are trying to perform deception detection using the same machine learning approaches which were used earlier for personality recognition task.

In recent years, automatic detection of deception became one of the major research areas not only in psychology but also in computational linguistics, military, and law enforcement agencies. Previous researches on deception detection used standard biometric indicators measured using the polygraph, body gesture, facial expression, brain imaging, lexical and acoustic-prosodic features etc.

Language specific features are gaining popularity as being inexpensive and easier to collect. A number of studies have been carried out in the area of deception detection using linguistic cues for both speech and text. Psycholinguistic categorization tool such LIWC was found to be helpful in detecting deception (Newman et al., 2003). Hirschberg et al. (2005) developed automatic deception detection trained on spoken cues and tested on unseen data for American English and achieved accuracies better than human judges. Verbal deceptive behavior across culture was found to be different by several researchers (Feldman, 1975; Cody et al., 1989). Levitan et al. (2015a) used the demographic and NEO-FFI along with acoustic-prosodic features for deception detection from spoken dialogue and achieved 10 percentage point improvement over baseline accuracy. In another work Levitan et al. (2016a), included DAL, LIWC with acoustic-prosodic and phonotactic features

for automatic detection of deception and showed improvement over baseline accuracy on Interspeech ComParE challenge corpus.

In our experiment, we want to explore how error rate for the deception detection task varies among different genders and L1 groups using the same procedure we used for personality detection.

6.1 Material and Methods

Columbia Cross-Cultural Deception (CXD) Corpus has deception labels for each within-subject deceptive and non-deceptive speech (Levitan et al., 2015a). We used DAL, LIWC, and acoustic-prosodic features (as described in Chapter 3) for this experiment. Although the corpus structure remains the same, motivated by earlier studies (Levitan et al. 2015a, 2015b, 2016a; Mendels et al., 2017) on deception detection, we used inter-pausal unit (IPU) segment instead of turn segment as unit of analysis.

To demonstrate that the procedure to identify the intergroup variability for personality recognition can also be applied in deception detection, we have used the same experiment design described in Chapter 4. Deception detection is a two-way classification experiment where higher F1 value indicates lower error rate.

6.2 Result and Discussion

Figure 6.1 shows the result of our first experiment on machine learning approach for deception detection. Based on this result, we can identify that the model trained on male instances performed best when tested on 100% in-group instances (0.42). However, the performance drops gradually as we increase the out-group instance (for 0 % in-group instances F1 = .37). A similar observation can be made for a model trained on female instances where the F1 value starts from 0.47 and steadily

reduces to 0.42 as we increased the out-group instances.

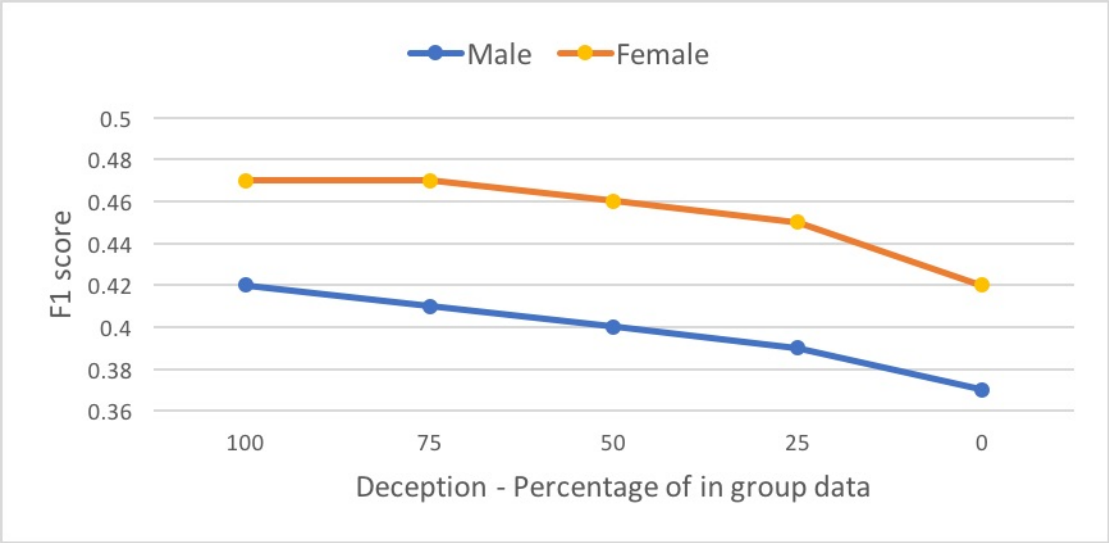

Figure 6. 1 F1 measures for different proportion of test data for deception detection for two models trained separately on Male and Female data

For the second machine learning experiment (result shown in Table 6.1) where we trained the model on different gender and tested on SAE and MC demonstrate that there is an intergroup variability among the male and female speaker of both SAE and MC. When the gender is constant, we do not see much intergroup variability for different L1 groups.

	Train on Gender	Test on L1	
		English	Mandarin
Deception	Male	0.40	0.39
	Female	0.48	0.49

Table 6. 1 F1 scores when train on gender and test on L1 groups.

Chapter 7

Conclusions

We have presented in this thesis a series of analysis and experiments to explore the intergroup variability across various demographic (gender and L1) groups for personality recognition task. To our knowledge, this is the first report of experiment to identify the error rate between models trained and tested across in-group and out-group categories. We hypothesized that there is a significant intergroup variability which can be better understood by investigating the differences in error rate and statistical relationship with personality traits between in-group and out-group data. Our results do suggest that the expression of personality through vocal characteristics or language use varies among speakers from different gender and cultural background. From these empirical results, we can conclude that both gender and L1 play an important role in personality recognition task. We can also infer that the influence of demographic differences on error rate varies considerably while predicting "Big Five" personality traits from speaker's utterances. We also observed that the prosodic features relate differently with each personality traits across L1 and gender groups. This study helps to understand how personality recognition task could be impacted when a model is trained and tested on data from cross-demographic groups. These findings can be used to motivate further development of personality recognition models that are robust to intergroup variability.

More specific trends that we observed, include:

- Extraversion is more likely to be impacted and Openness is most resilient to error while predicting personality for out-of-group speakers.

- The error rate for the native speakers of SAE is high while identifying Openness. Conversely, Neuroticism was found to be least among the same group of speakers.

- Native MC speakers across gender groups showed a higher error rate for predicting Agreeableness and Neuroticism than their SAE counterparts.

- Machine learning models trained on gender-specific data behaves differently while predicting personality for different L1 groups.

- Statistical analysis result replicates the earlier finding that extraversion can be predicted from voice quality for male SAE speaker. In addition, there is slight but a significant positive association between Extraversion and the voice characteristics of loudness and lower pitch.

Using a supplementary experiment for deception detection we have shown that the approach for identifying intergroup variability for personality recognition task can be generalized in other classification tasks. From experiment results, we noticed that the percentage of out-group instances inversely correlates with the F1 value across gender groups for deception detection task.

In future work, we can implement the findings of this study in advanced machine learning models to improve the accuracy for different demographic groups. We also believe that future work should investigate the subset of relevant prosodic and lexical features for each demographic groups in a trait-dependent way.

Bibliography

An, G., Levitan, S. I., Levitan, R., Rosenberg, A., Levine, M., & Hirschberg, J. (2016). Automatically Classifying Self-Rated Personality Scores from Speech. In INTERSPEECH (pp. 1412-1416).

An, G., & Levitan, R. (2018). Comparing approaches for mitigating intergroup variability in personality recognition. arXiv preprint arXiv:1802.01405.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. (2005). Lexical predictors of personality type.

Bond, M. H., Nakazato, H., & Shiraishi, D. (1975). Universality and distinctiveness in dimensions of Japanese person perception. Journal of Cross-Cultural Psychology, 6(3), 346-357.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. Journal of personality and social psychology, 62(4), 645.

Burger, J. M. (2011). Introduction to personality. Wadsworth/Cengage Learning.

Cody, M. J., Lee, W. S., & Chao, E. Y. (1989). Telling lies: Correlates of deception among Chinese. Recent advances in social psychology: An international perspective, 359-368.

Corr, Philip J.; Matthews, Gerald (2009). The Cambridge handbook of personality psychology (1. publ. ed.). Cambridge, U.K.: Cambridge University Press. ISBN 978-0-521-86218-9.

Costa, P. T., & McCrae, R. R. (1985). The NEO personality inventory. Odessa. FL: Psychological Assessment Resources

Costa, P. T., & McCrea, R. R. (1992). Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi). Psychological Assessment Resources.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual review of psychology, 41(1), 417-440.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013, October). Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838). ACM.

Feldman, R. S. (1979). Nonverbal disclosure of deception in urban Koreans. Journal of Cross-Cultural Psychology, 10(1), 73-83.

Fiscus, J. G. (1997, December). 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 (pp. 347-354). IEEE.

Gravano, A., Levitan, R., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2011). Acoustic and prosodic correlates of social behavior. In Twelfth Annual Conference of the International Speech Communication Association.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. Computer Speech & Language, 25(3), 601-634.

Heller, W. (1993). Neuropsychological mechanisms of individual differences in emotion, personality, and arousal. Neuropsychology, 7(4), 476.

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., ... & Pellom, B. L. (2005). Distinguishing deceptive from non-deceptive speech. In Ninth European Conference on Speech Communication and Technology.

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. Journal of biomedical informatics, 46(6), 998-1005.

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., & Rosenberg, A. (2015a, November). Cross-cultural production and detection of deception from speech. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (pp. 1-8). ACM.

Levitan, S. I., Levine, M., Hirschberg, J., Cestero, N., An, G., & Rosenberg, A. (2015b). Individual differences in deception and deception detection. Proceedings of Cognitive.

Levitan, S. I., An, G., Ma, M., Levitan, R., Rosenberg, A., & Hirschberg, J. (2016a). Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection. In INTERSPEECH (pp. 2006-2010).

Levitan, S. I., Levitan, Y., An, G., Levine, M., Levitan, R., Rosenberg, A., & Hirschberg, J. (2016b). Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection (pp. 40-44).

Maechler, M., & Bates, D. (2010). lme4: Linear mixed-effects models using S4 classes. R package version, 099937-099935.

Mairesse, F., & Walker, M. (2006, June). Automatic recognition of personality in conversation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (pp. 85-88). Association for Computational Linguistics.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research, 30, 457-500.

Mallory, E. B., & Miller, V. R. (1958). A possible basis for the association of voice characteristics and personality traits. Communications Monographs, 25(4), 255-260.

McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. Journal of personality, 69(6), 819-846.

McCrae, R. R., & Costa, P. T. (2003). Personality in adulthood: A five-factor theory perspective. Guilford Press.

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. Journal of personality and social psychology, 90(5), 862.

Mendels, G., Levitan, S. I., Lee, K. Z., & Hirschberg, J. (2017). Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. Proc. Interspeech 2017, 1472-1476.

Mohammadi, G., Vinciarelli, A., & Mortillaro, M. (2010, October). The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In Proceedings of the 2nd international workshop on Social signal processing (pp. 17-20). ACM.

Myers, R. H. (1990). Classical and modern regression with applications (No. 04; QA278. 2, M8 1990.).

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5), 665-675.

Oh, I. S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. Journal of Applied Psychology, 96(4), 762.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. Annu. Rev. Psychol., 57, 401-421.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. Journal of personality and social psychology, 72(4), 863.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. Journal of personality and social psychology, 77(6), 1296.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

Scherer, K. R. (1979). Personality markers in speech. Cambridge University Press.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. In Tenth Annual Conference of the International Speech Communication Association.

Siegman, A. W., & Pope, B. (1965). 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.

Shafran, I., Riley, M., & Mohri, M. (2003, December). Voice signatures. In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on (pp. 31-36). IEEE.

Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on personality perception. Language and speech, 18(2), 145-152.

Stewart, A. J. (1998). Doing Personality Research. The gender and psychology reader, 54.

Stewart, A. J., & McDermott, C. (2004). Gender in psychology. Annu. Rev. Psychol., 55, 519-544.

Team, R. C. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. In Proceedings of the eighth workshop on innovative use of NLP for building educational applications (pp. 48-57).

Whissell, C., Fournier, M., Pelland, R., Weir, D., & Makarec, K. (1986). A dictionary of affect in language: IV. Reliability, validity, and applications. Perceptual and Motor Skills, 62(3), 875-888.

Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychological reports, 105(2), 509-521.