Cross-Cultural Production and Detection of Deception from Speech

Sarah Ita Levitan[1], Guozhen An[2], Mandi Wang[1], Gideon Mendels[1],
Julia Hirschberg[1], Michelle Levine[1], Andrew Rosenberg[2]
[1]Department of Computer Science, Columbia University, New York, USA
[2]Department of Computer Science, Queens College, CUNY, New York, USA
sarahita@cs.columbia.edu, gan@gc.cuny.edu, mw2971@columbia.edu, gm2597@columbia.edu,
julia@cs.columbia.edu, mlevine@cs.columbia.edu, andrew@cs.qc.cuny.edu

ABSTRACT
Detecting deception from different dimensions of human behavior has been a major goal of research in psychology and computational linguistics for some years and is currently of considerable interest to military and law enforcement agencies. However, relatively little work has been done to develop automatic methods to detect deception from spoken language or to compare deception detection and production between different cultures. We present results of experiments on a new corpus of deceptive and non-deceptive speech, collected from native speakers of Standard American English and Mandarin Chinese, all speaking English, to investigate acoustic, prosodic, and lexical cues to deception. We report first on the role of personality factors derived from the NEO-FFI (Neuroticism-Extraversion-Openness Five Factor Inventory) and of gender, ethnicity and confidence ratings on subjects' ability to deceive and to detect deception. We then present classification results discriminating deceptive from non-deceptive speech, using these features as well as acoustic and prosodic cues. We find that combining acoustic and prosodic features with information about the speaker's personality, gender, and language results in a classification accuracy of 65.86%, which represents ~10% relative improvement from baseline accuracy.

Categories and Subject Descriptors
[**Computing Methodologies**]: Artificial Intelligence → Natural Language Processing – *Speech Analysis*

General Terms
Experimentation, Human Factors

Keywords
Deception detection, speech, cross-cultural, American English, Mandarin Chinese

1. INTRODUCTION
Finding new methods for detecting deception is a major goal of researchers in psychology and computational linguistics as well as commercial organizations, law enforcement, military, and intelligence agencies. While many new techniques and technologies have been proposed and a few have been tested in the field, there have been few real successes. The lack of large, cleanly recorded corpora; the difficulty of acquiring 'ground truth' for the truth/lie distinction; and major differences in incentives for lying in the laboratory vs. lying in real life situations are all obstacles to this work. Another well-recognized issue is the strong belief that there are individual and cross-cultural differences in deception detection and production, although little has been done to identify these.

The goal of our research is to develop techniques to identify deceptive communication in spoken dialogue. As part of this effort, we are studying how acoustic, prosodic, and lexical features of an individual's speech can be used, together with knowledge of gender, ethnicity and personality factors, to distinguish deceptive from non-deceptive behavior. Although our studies are done in the laboratory in order to provide cleanly-recorded, comparable sessions, we provide an effective monetary incentive to subjects for both detecting and producing effective deceptive behavior. While subjects are asked to lie in answer to certain questions, they are allowed to construct their own lies and to indicate as they talk which utterances are true and which are false. We also study deception production and perception within and across cultures and genders, with male and female subjects who are native speakers of Standard American English (SAE) and Mandarin Chinese (MC).

In this paper, we describe new results of experiments on cross-cultural cues to deception, correlating gender, ethnicity, and personality characteristics from the NEO-FFI Five Factor Analysis [11] with subjects' ability to deceive and to judge deception in others' speech. We also describe our first classification experiments using these data as features, and including acoustic-prosodic features to automatically distinguish deceptive from non-deceptive speech using machine learning algorithms. In Section 2, we describe previous work on cues to deception and deception detection. In Section 3 we describe our experimental design and corpus collection. In Section 4 we explain the methods used in corpus annotation, segmentation, and the alignment of speech and orthographic transcriptions. Section 5 presents new results of correlations between deceptive behavior and gender, ethnicity, and personality traits. Section 6 describes results of our classification experiments. We conclude in Section 7 and discuss future research.

2. RELATED WORK
Previous research on deception has included examination of facial expressions, body gestures, brain imaging, body odor, as well as linguistic information, to supplement the use of standard biometric indicators commonly measured in polygraphy which

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
WMDD'15, November 13, 2015, Seattle, WA, USA.
© 2015 ACM. ISBN 978-1-4503-3987-2/15/11…$15.00.
DOI: http://dx.doi.org/10.1145/2823465.2823468

alone have been shown to perform no better than chance [15]. Attempts to distinguish truth from lie using facial expressions have been controversial [13][10][19] and are also difficult to automate, requiring expensive video capture technology, laborious human annotation, and subsequent alignment with transcribed and semantically interpreted language to identify mismatches between "micro-expressions" and language. There have been promising results using automatic capture of body gestures as cues to deception [39][34], but this method again requires multiple, high-caliber cameras to capture movements reliably and match them with speech. The use of brain imaging techniques for deception detection is still in its infancy [30] and requires the use of MRI techniques not practical for general use. Body odor as an indicator of deception is in a very early stage [47].

Language cues to deception include text-based studies such as Statement Analysis [2], SCAN [42][44], and lexical signals taught by John Reid and Associates in their training on interview and interrogation of suspects [41]. With the exception of work by Bachenko et al. on lexical cues proposed for Statement Analysis [3] few of these cues have been scientifically validated, although other lexical cues to deception have been found useful in distinguishing truth from lie by Pennebaker and colleagues [36][37] and by Hancock et al. [25]. However, little work has been done on cues to deception drawn from the speech signal. Simple features such as intensity and hypothesized vocal tremors have performed poorly in objective tests [23][28][27][15], although other features examined by Harnsberger et al. [26] and Torres et al. [45] have proven more useful. In previous work on deception in American speech, Hirschberg et al. [27] developed automatic deception detection procedures trained on acoustic, prosodic and lexical cues and tested on unseen data which achieved accuracy of 70% and an F1 measure of 75.78% (predicting truth) --- compared to 58.7% accuracy of human judges on the same data. In the process of identifying common characteristics of deceivers, they also noticed a range of individual differences in deceptive behavior, e.g., some subjects raised their pitch when lying, while some lowered it significantly; some tended to laugh when deceiving, while others laughed more while telling the truth. They also discovered that human judges' accuracy in judging deception could be predicted from their scores on the NEO-FFI, suggesting that such simple personality tests might also provide useful information in predicting individual differences in deceptive behavior itself [14].

Differences in verbal deceptive behavior in different cultures have been identified by several researchers [9][16]. Studies of deceptive behavior in non-Western cultures have primarily focused on understanding how culture affects *when* people deceive and *what* they consider deception [31][43]. Studies investigating the universality of deceptive behavior have found that, while stereotypes may exist [5] these may not correlate with actual deceptive behavior [46][49] and that culture-specific deception cues do exist [9][16] .

In the work presented here, we investigate both the ability to deceive and to detect deception considering gender and ethnicity and examining additional cues to deception: acoustic and prosodic information, as well as features extracted from the NEO-FFI personality inventory [11] and subjects' gender and native language; note that the analysis of personality features here extends our initial identification of correlations between gender, ethnicity, and personality described in [32], while adding new classification experiments which include these as well as new acoustic and prosodic features.

3. CORPUS

To investigate questions of individual and cross-cultural similarities and differences in deception perception and production, we have collected a large corpus of within-subject deceptive and non-deceptive speech from native speakers of SAE and MC, both speaking in English. The corpus is balanced for ethnicity (native language) and gender. Our balanced corpus includes data from 126 (previously unacquainted) subject pairs, constituting 93.8 hours of speech. This balanced corpus was used for our statistical analysis (described in section 5), while a larger dataset, unbalanced for gender and ethnicity, was used for our classification experiment (described in section 6). The larger dataset is a superset of the balanced corpus and consists of conversations between 139 pairs of subjects comprising 100.5 hours of speech. To our knowledge, this is by far the largest corpus of cleanly recorded deceptive and non-deceptive speech collected and transcribed, with known truth/lie distinctions.

We employ a form of the 'fake resume' paradigm in which we elicit true and false biographical information from subjects to serve as ground truth. Subjects are separately informed that they will play a lying game with another subject, in which they will alternate between interviewing their partner and being interviewed themselves about answers to a set of 24 biographical questions. As interviewees, they should try to convince their interviewer that everything they say is true. As interviewers, they should try to identify when the interviewee is lying and when they are telling the truth. To motivate them, they are told that their compensation depends on their ability to deceive while being interviewed, and to judge truth and lie correctly while interviewing. As interviewer, they receive $1 each time they correctly identify an interviewee's answer as either lie or truth and lose $1 for each incorrect judgment. As interviewee, they earn $1 each time their lie is judged to be true, and lose $1 each time their lie is correctly judged to be a lie by the interviewer.

Subjects are then asked to complete a 24-item biographical questionnaire truthfully. In addition to their true answers, they are told to create a false answer for half of the questions as indicated on their answer sheet. They are given guidelines in preparing false answers that differ sufficiently from truth, to ensure that lying will not be too easy. For example, for the question "Where were you born," the false answer must be a place that the subject has never visited, a false answer to "What is your father's occupation" must be different from their mother's true occupation, and so on. An experimenter checks the false answers to make sure subjects follow the guidelines. Next, each subject completes the NEO-FFI personality inventory [11], which is described below. While one subject is completing their NEO-FFI inventory, we collect a 3-4 minute baseline sample of speech from the other subject for use in speaker normalization. The experimenter elicits natural speech by asking the subject open-ended questions (e.g., "What do you like best/worst about living in NYC?"). Subjects are instructed to be truthful during this part of the experiment. Once both subjects have completed all the questionnaires and we have collected baseline samples of speech, the lying game begins.

For recording purposes, subjects are seated across from each other in a double-walled sound-proof booth, separated by a curtain so that there is no visual contact (Figure 1); this is necessary since our focus is on spoken and not visual cues. There are two parts to each session. During the first half, one subject acts as the interviewer while the other answers the biographical questions, lying for half and telling the truth for the other half, based on the modified questionnaire. In the second part of the session, roles

are reversed. Each subject is recorded on a separate channel using Crown CM311A Differoid head-worn close-talking microphones and a TASCAM HD-P2 High Resolution stereo recorder.

Figure 1. Setup of experiment in sound-proof booth

The interviewer is able to ask the biographical questions in any order s/he chooses, and is encouraged to ask follow-up questions to help determine the truth of the interviewee's answers. For each question, the interviewer records his/her judgment, along with a confidence score from 1-5. As the interviewee answers the questions, s/he presses a T or F key on a keyboard (which the interviewer cannot see) for each phrase, logging each segment of speech as true or false. Thus, while the biographical questionnaire provides the 'global' truth value for the answer to the question asked, the key log provides the 'local truth' value for each phrase, which is automatically aligned with each speech segment. At the end of the experiment, subjects complete a brief questionnaire, which includes additional confidence questions.

The NEO-FFI personality assessment given to subjects [11] is based on the five-factor model of personality, an empirically-derived and comprehensive taxonomy of personality traits. It was developed by applying factor analysis to thousands of descriptive terms found in a standard English dictionary. It is used to assess the five personality dimensions of the following factors:

- *Openness to Experience.* Designed to capture imagination, aesthetic sensitivity, and intellectual curiosity. It is "related to aspects of intelligence, such as divergent thinking, that contribute to creativity" [11] . Those who score low on this dimension prefer the familiar and tend to behave more conventionally. People high in Openness are "willing to entertain novel ideas and unconventional values" [11] .

- *Conscientiousness*. Addresses individual differences in self-control, such as the ability to control impulses, but also to plan and carry out tasks. It measures contrasts between determination, organization, and self-discipline and laxness, disorganization, and carelessness.

- *Extraversion*. Meant to capture proclivity for interpersonal interactions, and variation in sociability. It reflects contrasts between those who are reserved vs. outgoing, quiet vs. talkative, and active vs. retiring.

- *Agreeableness*. Measures interpersonal tendencies and is intended to assess an individual's fundamental altruism. Individuals high in Agreeableness are sympathetic to others and expect that others feel similarly.

- *Neuroticism*. Contrasts emotional stability with maladjustment. It is intended to capture differences between those prone to worry vs. calm, emotional vs. unemotional behavior, and vulnerable vs. hardy.

4. ANNOTATION AND SEGMENTATION

For purposes of analysis the corpus has been manually transcribed using a crowdsourcing approach. Multiple hand transcriptions are then compared and sometimes corrected in the lab to produce a gold standard transcript. This transcript is then aligned with the truth/lie information collected from subjects and with the speech signal. The resulting aligned material is then segmented into prosodic phrase units so that true/false labels can be classified according to the acoustic-prosodic and lexical information they contain.

4.1 Transcription and alignment

The speech data was collected as two channels of a single audio file with each session running 30-60 minutes. To facilitate transcription, we segmented the audio to smaller utterance-like units. First, we separated the two channels, so a transcriber is only responsible for transcribing a single speaker. Then, based on intensity thresholding, we identified silent regions, splicing the audio file in the middle of each silent region. We tuned the segmentation parameters (intensity threshold, minimum silence length and minimum non-silence length) to obtain speech segments that are 10 seconds long, though some are up to 30 seconds or more. We initially hoped to omit all silence from the segments to simplify the transcription task, but we all strategies that involved removing silence were prone to errors, omitting quiet speech or cutting off speech too early or in the middle of a longer phrase. While this approach presents transcribers with a good deal of silence, the risk of omitting the speech from one speaker was too great.

We collected the transcripts through Amazon Mechanical Turk (AMT). Amazon Mechanical Turk is a large scale crowdsourcing service that is used to perform "Human Intelligence Tasks" (HITs), generally small tasks that require some human effort. We designed our HITs to contain 21 audio clips, one 'quality control' clip, and 20 others. The biggest challenge with using AMT for a task like this is quality control. A research assistant correctly transcribed each quality control audio clip. Comparison with this correct transcript allows us to estimate how reliably the transcriber is performing overall. We posted three assignments for each HIT, thus obtaining three transcripts for each audio segment from three different transcribers or 'Turkers'.

After collecting three transcripts for the audio clips, we used the *Rover* tool [18] to combine them to one single transcript. Rover is a tool which combines hypothesized transcripts to generate a more reliable "consensus" result. This results in a single best transcript, $w = w_1,\ldots, w_N$ where w_i is the *i*-th word in the transcript, from *K* candidate transcripts ($w^{(k)} = w_1^{(k)},\ldots, w_N^{(k)}$). Note that, in order to generate w and $w^{(k)}$ such that all transcripts are the same length *N*, rover can insert null words $w_i = \varepsilon$ in any transcript sequence. In addition to generating the consensus transcript w, we also calculated a rover output score, $s(w)$, measuring the agreement between the initial transcripts as:

$$s(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\delta(w_i = w_i^{(k)}),$$

where δ is a Kronecker delta function, equaling 1 when the condition is true, and 0 otherwise.

In some instances, there was substantial disagreement among the Turkers. For those clips with $s(w)$ lower than 70%, we corrected the transcript manually. Ultimately, we needed to hand correct 9.7% of our transcribed clips.

Once we obtained a high quality transcript for each clip, we force aligned the transcript with the audio. We evaluated three different aligners for this task: 1) Prosody-lab Aligner [21] 2) Penn Aligner [48] and 3) the Kaldi Speech Recognition Toolkit [38]. We found Kaldi to generate the most reliable performance. Kaldi is a toolkit for speech recognition written in C++, and freely available under the Apache license. The acoustic model we used for forced alignment is a triphone Gaussian Mixture Model (GMM) with Feature space Maximum Likelihood Linear Regression (fMLLR) adaptation and trained using the standard Kaldi recipe and the Wall Street Journal corpus [20][33]. We initially used the CMUDict [7] as a lexicon. However, this resulted in approximately 4,006 out-of-vocabulary (OOV) tokens from all our transcriptions. These OOVs included 2,202 true OOV terms (words such as proper names which did not appear in CMUDict, as well as word fragments such as false starts, e.g. "st- stairs" (for which we constructed pronunciations), and 1,804 transcriber spelling mistakes which required hand correction of the transcripts.

4.2 IPU segmentation and labeling

The unit of analysis in this work is an *inter-pausal unit* (IPU), defined as a pause-free segment of speech from a speaker [27]. The minimal pause length between segments used here is 50ms. We segment all speech into IPUs using Praat [4]. The silence detection is done using a simple intensity measure: first the intensity is determined, and then speech and silence intervals are calculated using intensity thresholding. We use Praat's default value for the intensity threshold: segments with a maximum intensity that falls 25dB below the maximum intensity of the speaker are labeled as silent.

We observed IPU segmentation errors resulting from this method of silence detection, in which areas of speech were mistakenly identified as silence. We therefore decided to obtain IPU segmentation using another method, in which we used the transcribed and aligned data to identify IPU boundaries as silences identified by the aligner as it aligns the spoken words. By parsing the aligned transcription files, we use the word boundaries to infer silence and speech labels, and then extract IPUs. For this work, we use the original, noisy IPU segmentation obtained using Praat [4]. However, we plan to repeat our experiments with our new word-alignment defined IPUs and hypothesize that this cleaner data will improve our results.

The next step after IPU segmentation is assigning true and false labels to the IPUs. As previously mentioned, subjects labeled each utterance as true or false using key presses during the interview. We converted these discrete time-stamped points to intervals, with the assumption that each key press labels the preceding speech up to the previous true/false label – as the subjects were instructed to do. For each IPU, we checked which interval it overlapped with and labeled it with the corresponding true or false label. If an IPU overlapped two contradicting labels, we first checked the distance between the conflicting key presses. Subjects sometimes mistakenly pressed the wrong key during the interview, and were instructed to immediately correct any mistaken key presses by quickly pressing the correct one. If the conflicting key presses were less than 10ms apart, we treated that as an error and used the second key press as the correct label. Otherwise, we chose the label that had the longer coverage of the IPU. There is an inherent difficulty in this labeling task because we do not know the "true" intention of the subject when labeling one of these ambiguous local lies, so we cannot objectively evaluate our method of interpreting them. For the classification experiments reported in this work, we did not include IPUs that had conflicting labels since they were a small percentage of our IPUs – about .5%. In total, we used 145,621 IPUs in our classification experiments.

4.3 Crosstalk identification

Although our data was collected using separate channels for each subject in the interview, in some cases *crosstalk* from the other speaker was quite audible. Although transcribers were instructed to transcribe only the loudest speaker, in some cases crosstalk was mistakenly transcribed, aligned and segmented into IPUs. Since this crosstalk was generally of somewhat lower intensity than the speech the transcriber should have recorded, we used the lower intensity levels to filter out the crosstalk.

For each recording session IPUs we calculated the mean intensity and its standard deviation. Using these values, we processed each IPU and compared its mean intensity with the mean intensity of all the IPUs for that speaker. If it was less than two standard deviations from the mean, we labeled it as crosstalk. We chose this decision boundary experimentally.

The formula used for identifying crosstalk is as follows. Label an IPU as crosstalk if:

$$\mu(x_i) < \mu(x) - 2\sigma(\mu(x))$$

where x_i is the current IPU, and mean represents the mean intensity.

Crosstalk identification is an important step for using the IPU segmentation obtained from the aligned transcription files. Because we use the Praat segmentation in this work, the crosstalk identification step was not necessary. However, we plan to use this preprocessing step when we repeat our experiments with the alignment defined IPUs.

5. NEO-FFI CORRELATIONS WITH GENDER, ETHNICITY AND DECEPTIVE BEHAVIOR

In our analysis of the relationship between personality factors, gender, ethnicity and the ability of subjects to deceive and to detect deception, we examined whether subjects' ability to detect deception correlates with their ability to deceive. We had previously studied these correlations on an unbalanced corpus of 126 conversations consisting of ~87 hours of speech [32]. We now present results on a balanced (by gender and ethnicity) corpus including 126 conversations consisting of ~94 hours of speech. Similar to our previous finding, our current data indicates that subjects who are better at detecting deceptive answers are also better at deceiving, $r(252) = 0.13$, $p = 0.04$. When we considered the effect of gender and native language, we originally saw that the correlation was strongest for females, specifically SAE females. Now however, with our larger balanced corpus, we see that the correlation is strongest across *all* females ($r(126) = 0.26$, $p = 0.003$), both SAE females and MC females.

As we saw earlier, those who are better at detecting deception are also more likely to label their partners' answers as untrue, whether or not their partner in fact had lied ($r(252) = 0.74$, $p < 0.001$). We also see that across all subjects, those who are more likely to label their partners' answers as lies are also better at deceiving, $r(252) = 0.13$, $p = 0.04$. This clarifies our earlier finding in which we had only found the later effect for female subjects.

Next, we examined how individual differences in gender, culture, and personality affect subjects' ability to deceive and detect

deception. In line with our previous results [32] there is no effect of subjects' gender, native language, or personality factors on their ability to detect deception. In addition, there is no effect of gender and native language on people's ability to deceive. However, the personality factor of Extraversion *negatively* correlates with subjects' ability to deceive for SAE males, r(68) = -0.25, p = 0.04. We no longer find an effect of Extraversion and MC females' success or Conscientiousness and SAE females' success. In our classification experiments we are investigating how types of *individual* differences in speech behavior may interact with personality, gender, and ethnicity to predict deception and ability to detect it.

Finally, we examined how confidence affects subjects' ability to deceive and to detect deception, and how personality factors affect confidence. Previously we had found a gender difference in regard to confidence ratings – that is, female subjects' ability to detect deception negatively correlated with their average confidence in judgments [32] . In our current corpus, we find this effect across all subjects (r(250) = -0.14, p = 0.03), and, once again, specifically for females (r(126) = -0.24, p = 0.01). In addition, we now find that male subjects', and specifically MC male subjects', ability to deceive negatively correlates with their average confidence ratings (r(124) = -0.185, p = 0.04 and r(58) = -0.35, p = 0.007). Our original hypothesis remains – interviewers who are less confident in their judgments may ask more follow-up questions and thus obtain more evidence to determine deception. Perhaps males who are better liars are less confident in judging other people's lies and therefore may also tend to ask more follow-up questions and obtain more evidence. Further analyses that include number of follow-up questions and answer length are needed to determine whether this holds true. Previously we had found that, across females, average confidence in detecting deception *negatively* correlated with Neuroticism. Now we see this same effect but specifically for MC females, r(68) = -0.27, p = 0.02. In addition, we now see that, across subjects, average confidence in detecting deception *negatively* correlates with Openness to Experience, r(249) = -0.14, p = 0.03. This holds true specifically for females, and MC females (r(126) = -0.21, p = 0.02 and r(68) = -0.29, p = 0.02). It thus appears that women who are less "neurotic" are more confident in their deception judgments and women who are less "open" are also more confident in their deception judgments. However, when we look only at male subjects, we do not find an effect of personality factors on their average confidence in detecting deception. We have now begun to explore how personality factors may combine with aspects of spoken behavior, however, as predictors of deception.

6. CLASSIFICATION EXPERIMENTS

We use a superset of our balanced corpus for the classification experiments. This data is unbalanced and consists of conversations between 139 pairs of subjects comprising 100.5 hours of speech. For statistical analysis, it is important to use data that is balanced for pair types in order to make meaningful comparisons between the groups. For our machine learning experiments, balancing for gender and ethnicity is less of a concern, since we are not making claims about differences between subject pairs but rather we are exploring whether information about a person's gender and native language can help with the deception classification. Therefore, we chose to leverage the additional data (14 sessions) despite the fact that it is not balanced.

We first describe results of our deception detection classification experiments in which we compare three classification models, implemented in the Weka machine learning library [24]. We treat this problem as a binary classification problem, and aim to distinguish between the two classes: is the speaker telling the truth or lying --- true vs. false. We chose to explore two ensemble methods, *random forests* [6] and *bagging* [7], since ensemble learners run efficiently on large datasets. Random forests generate multiple decision trees, each trained on a random subset of features, and classifying by majority vote. Bagging, or *bootstrap aggregating*, generates new training sets by uniformly sampling from the original training set with replacement, and then training multiple models and classifying by majority vote. This method has the advantage of reducing variance, which helps avoid overfitting. We compare the performance of the ensemble methods to a decision tree method, J48, which is Weka's implementation of the *Iterative Dichotomiser 3* (ID3) algorithm [40]. This is a commonly used algorithm to generate a decision tree from a dataset. All experiments are evaluated using stratified 10-fold cross-validation.

The main feature set consists of acoustic and prosodic features extracted at the IPU level using Praat [4]. We used the following 14 acoustic-prosodic features for our classification experiments: f0 minimum, f0 maximum, f0 mean, f0 median, f0 standard deviation, f0 mean absolute slope; intensity minimum, intensity maximum, intensity mean, intensity standard deviation; *jitter*, *shimmer*, *noise to harmonics ratio*. The first six features are different measures of the *fundamental frequency*, the physical correlate of pitch. The next four are measures of a correlate of perceived loudness. The last three features are measures of *voice quality*, variation in vocal fold behavior which leads to listeners' perception of the harshness or creakiness or breathiness of the voice. We also estimate *speaking rate* by calculated the ratio of voiced to total frames and include this as a feature in classification. All these features have been proposed in the literature on deception as possible indicators of deception [12].

We also explored different methods of feature normalization. In our first set of experiments we used raw acoustic-prosodic features only, and in our next set we explored two normalization methods. In one method, we normalized a given speaker's features using the mean and standard deviation of the speaker's features throughout both parts of the session. We refer to this as "session normalization."

In the second normalization method, we employed the baseline data collection part of the experiment, in which we collected a 3-4 minute sample of each participant's (truthful) speech before they had met their partner. To capture the speaker's deviation from his or her truthful way of speaking, we normalized their speech during the lying game using features extracted from the baseline. We hypothesized that this normalization, which we call "baseline normalization," will be useful for deception detection. We used the *z-score normalization*[1] method for both normalizations.

The baseline for these experiments is 59.9% (assigning the label of the majority class – true – to each IPU, or, assuming that speakers are always telling the truth.

[1] The formula for z-score normalization is

$$Normalized\ (e_i) = \frac{e_i - \bar{E}}{std(E)}$$

Table 1. Accuracy of 3 models, using raw acoustic-prosodic features and 2 methods of feature normalization

Model	Raw	SessionNorm	BaselineNorm
J48	59.89	62.09	62.19
Bagging	58.65	61.19	61.01
RandomForest	61.23	**63.03**	62.79

Table 1 compares the classification results of 3 models and raw vs. normalized features. We find that the Random Forest model yields the best accuracy for all normalization methods, with "Session normalization" yielding the highest accuracy, 63.03%. This is a 5.2% increase over the baseline. We observe that both speaker- and session-normalized features perform better for this task than do raw features, and they result in very similar accuracy.

In addition to these acoustic-prosodic features, we explored the following nine features that help capture the broader gender and cultural differences between speakers: the five NEO-FFI scores, speaker gender, speaker native language, partner gender, and partner native language. We built models using gender, language, and personality information as well as acoustic-prosodic features. We hypothesized that including this information would help to capture individual differences in the way that people exhibit deceptive behavior. Table 2 shows the results of these experiments using session-normalized features.

Table 2. Results using session normalized features and personality scores, gender and language

Model	SessionNorm + NEO, gender, lang
J48	64.86
Bagging	63.9
RandomForest	**65.86**

Again, Random forest is the best model; achieving 65.86% accuracy with session normalized features. This is a 4.5% increase over the best acoustic-prosodic model, and a 9.95% increase over the baseline. We also experimented with combining raw and normalized features, but we did not observe an increase in performance. Our next step will be to add lexical features to the feature set to see how *what* is said contributes to deception detection as well.

7. CONCLUSIONS AND FUTURE WORK

We have described ongoing work in the production and detection (the latter by humans and by machine learning classifiers) of deceptive speech. We discussed our experimental paradigm, a variant of the "fake resume" paradigm and how we implemented it to collect the largest cleanly recorded, transcribed corpus of within-subject deceptive and non-deceptive speech with known truth-lie annotations. Our 120h+ corpus includes a subset balanced for gender and ethnicity and a larger corpus, in each case recorded by pairing males and females, native speakers of Standard American English and Mandarin Chinese, interviewing each other in turn, and motivated to lie or to detect lies by financial gain and indicating themselves when they lied and when they did not. We also describe the automaton techniques we used for corpus transcription, segmentation, and data 'cleaning'.

We present new results of the role of personality factors in the ability of subjects of different genders and ethnicities to deceive and to detect deception, confirming our previous finding that the ability to deceive is significantly correlated with ability to detect deception, finding new evidence that, while this holds for all our subjects, it is strongest for all females, both SAE and MC. While again we find no effect of gender, ethnic background, or personality traits on subjects' ability to detect deception, we do find correlations of certain personality scores with ability to deceive for some subsets of our population. We also again find that ability to detect deception negatively correlates with subjects' confidence in their judgments of deception – now for our entire balanced corpus.

We have also presented initial results of classification experiments using acoustic-prosodic features as well as culture, gender, and personality features to detect deceptive speech. Our best classifier achieves an accuracy of 65.86%, representing an almost 10% increase over the majority class baseline. These results are very promising, as we have not yet introduced lexical features. As a comparison, Graciarena et al. [22] present the results of similar experiments using prosodic features, and their best prosodic model resulted in 62.7% accuracy, a 3.8% increase over their majority baseline. When they combined this system with a cepstral system however, they observe a 6.6% increase over the baseline, suggesting that additional acoustic features may also improve our results. The contributions of the NEO-FFI scores and the speaker's and partner's gender and native language to our classification results are also quite promising; they suggest that these are important factors to consider when building tools to automatically distinguish between truth and deception.

In future, we plan to explore the addition of lexical features once our alignment process is complete. We also plan to experiment with our second method of IPU segmentation using the aligned word boundaries. Another idea for future work is to experiment with different size units for classification, which has been experimented with in previous literature [27]. The IPU is a linguistically meaningful unit corresponding to the prosodic phrase; however it may be difficult to make a decision about the truth-value of such a small segment of speech without using additional context, both lexical and acoustic. We plan to map our IPU level features to a larger speech segment, such as an utterance or turn, and to use a combination of our features to classify these as truth or lie.

Another area for future work is to build gender- and culture-specific deception detection models, instead of simply using this information as features. Different genders and cultures are thought to exhibit deceptive behavior in different ways, and more gender- and culture-specific models may improve the state of the art in automatic deception detection.

8. ACKNOWLEDGMENTS

We thank the following students for their contributions to this study: Zoe Baker-Peng, Lingshi Huang, Melissa Kaufman-Gomez, Yvonne Missry, Elizabeth Petitti, Sarah Roth, Molly Scott, Jennifer Senior, Grace Ulinski, and Christine Wang. This work was partially funded by AFOSR FA9550-11-1- 0120.

9. REFERENCES

[1] M. Aamondt and H. Custer. 2006. "Who can best catch a liar?" *Forensic Examiner* 15, 1 (2006), 6-11.

[2] S. Adams. 1996. “Statement analysis: What do suspects' words really reveal?” *FBI Law Enforcement Bulletin*, October.

[3] J. Bachenko, E. Fitzpatrick, and M. Schonwetter. 2008. “Verification and implementation of language-based deception indicators in civil and criminal,” International Conference on Computational Linguistics (Manchester), 1, 41-48.

[4] P.Boersma and D.Weenink,“Praat: doing phonetics by computer [computer program].” version 5.3.23, http://www.praat.org.

[5] C. Bond and The Global Deception Research Team. 2006. “A World of lies.” *Journal of Cross-Cultural Psychology*, 37, 1, 60-74.

[6] Leo Breiman. 2001. “Random Forests.” *Machine Learning.* 45(1), 5-32.

[7] Leo Breiman. 1996. “Bagging predictors.” *Machine Learning.* 24(2),123-140.

[8] Carnegie Mellon Pronouncing Dictionary https://github.com/cmusphinx/cmudict

[9] M. Cody, W. Lee, and E. Y. Chao. 1989. “Telling lies: Correlates of deception among Chinese,” in J.P. Forgas and J. M. Innes, eds., *Recent Advances in Social Psychology: An International Perspective*, 359-568.

[10] J. Cohn. 2009. http://abcnews.go.com/GMA/HealthyLiving/autism-research-benefit-studying-babies-facial-recognition-experts/story?id=9244817&page=3, 12/04/2009.

[11] P. T. Costa and R. R. McCrae. 1992. “Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI).” Odessa, FL: Psychological Assessment Resources.

[12] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, *129*(1), 74.

[13] P. Ekman, M. Sullivan, W. Friesen, and K. Scherer. 1991. “Face, voice, and body in detecting deception,” *Journal of Nonverbal Behaviour,* 15, 2, 125-135.

[14] F. Enos, S. Benus, R. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg. 2006. “Personality factors in human deception detection: Comparing human to machine performance.” Interspeech 2006 (Pittsburgh).

[15] A. Eriksson and F. Lacerda. 2007. “Charlatanry in forensic speech science: A problem to be taken seriously.” *The International Journal of Speech, Language and the Law*, 14, 2, 169-193, http://www.scribd.com/doc/9673590/Eriksson-Lacerda-2007.

[16] F. Feldman. 1979. “Nonverbal disclosure of deception in urban Koreans,” *Journal of Cross-Cultural Psychology*, 10, 215-221.

[17] R. Feldman, L. Jenkins, and O. Popoola. 1979. “Detection of deception in adults and children via facial expressions,” *Child Development*, 350-355.

[18] Fiscus, J. G. 1997. “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 1997*)* 347-354.

[19] M. Frank, M. O’Sullivan, and M. Menasco. 2008. “Human behavior and deception detection,” in J. G. Voeller, ed., *Wiley Handbook of Science and Technology for Homeland Security*, New York: John Wiley & Sons.

[20] Garofolo, John, et al. 1993. CSR-I (WSJ0) Complete LDC93S6A. Web Download. Philadelphia: Linguistic Data Consortium.

[21] Gorman, Kyle, Jonathan Howell and Michael Wagner. 2011. “Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech,” *Canadian Acoustics.* 39, 3, 192–193.

[22] Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., & Kajarekar, S. 2006. “Combining prosodic lexical and cepstral systems for deceptive speech detection,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), 1, 1033-1036.

[23] D. Haddad and R. Ratley. 2002. “Investigation and evaluation of voice stress analysis technology,” http://www.ncjrs.org/pdffiles1/nij/193832.pdf, March.

[24] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations,* 11, 1.

[25] J. Hancock, L. Curry, S. Goorha, and M. Woodworth. 2008. “On lying and being lied to: A linguistic analysis of deception.” *Discourse Processes*, 45, 1-23,.

[26] J. Harnsberger, H. Hollien, C. Martin, and K. Hollien. 2009. “Stress and deception in speech: evaluating layered voice analysis,” *Journal of Forensic Sciences*, vol. 54, 3, 642-50.

[27] J. Hirschberg et al. 2005. “Distinguishing deceptive from non-deceptive speech.” Interspeech 2005 (Lisbon).

[28] H. Hollien, J. Harnsberger, C. Martin, and K Hollien. 2006. “Evaluation of the NITV CVSA.” *Journal of Forensic Sciences*, 53, 183-193.

[29] H. Hollien and J. Harnsberger. 2006. “Voice stress analyzer instrumentation evaluation.” Technical report, Counterintelligence Field Activity.

[30] D. Langleben et al.. 2005. “Telling truth from lie in individual subjects with fast event-related fMRI.” Human Brain Mapping, 26, 4, 262-72.

[31] M. Lapinski and T. Levine. 2000. “Culture and information manipulation theory: The effects of self-construal and locus of benefit on information manipulation.” Communication studies, vol. 51, 1, 55-73.

[32] S.I. Levitan et al. 2015. “Individual differences in deception and deception detection.” Cognitive 2015 (Nice).

[33] Linguistic Data Consortium. 1994. CSR-II (WSJ1) Complete LDC94S13A, DVD. Philadelphia.

[34] T. Meservy et al. 2005. “Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior.” *IEEE Intelligent Systems*, 20, 5, 36-43, September.

[35] National Research Council of the National Academies: Committee to Review the Scientific Evidence of the

Polygraph. 2003. *The Polygraph and Lie Detection*. National Academies Press: Washington, D. C.

[36] M. Newman, J. Pennebaker, D. Berry, and J. Richards. 2003. "Lying words: Predicting deception from linguistic style." *Personality and Social Psychology Bulletin*, 29. 665-675.

[37] Pennebaker, M. Francis, and R. Booth. 2001. "Linguistic Inquiry and Word Count." Erlbaum Publishers, Mahwah, NJ.

[38] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesel, K. 2011, "The Kaldi speech recognition toolkit," IEEE Automatic Speech Recongition and Understanding (ASRU 2011).

[39] T. Qin, J. Burgoon, and J. Nunamaker. 2004. "An exploratory study on promising cues in deception detection and application of decision tree." Hawaii International Conference on System Sciences, 223-32.

[40] R. Quinlan (1986). Induction of decision trees. Machine Learning. 1(1):81-106.

[41] J. E. Reid and Associates. 2000. "The Reid Technique of Interviewing and Interrogation," Reid, John E. and Associates, Inc.

[42] A. Sapir. 1987. "Scientific Content Analysis (SCAN)." Laboratory of Scientific Interrogation. Phoenix, AZ.

[43] J. Seiter, J. Bruschke, and C. Bai. 2002. "The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship." *Western Journal of Communication* 66, 2, 158-180.

[44] N. Smith. 2001. "Reading between the lines: An evaluation of the scientific content analysis technique (SCAN)," Police Research Series. London, UK.

[45] J. Torres, E. Moore, and E. Bryant. 2008. "A Study of Glottal Waveform Features for Deceptive Speech Classification." ICASSP 2008 (Las Vegas).

[46] A. Vrij and G. Semin. 1996. "Lie experts' beliefs about nonverbal indicators of deception," 20, 1, 65-80, 1996.

[47] S. Waterman. 2009. "DHS wants to use human body odor as biometric identifier, clue to deception," United Press International (UPI). http://www.upi.com/Top_News/Special/2009/03/09/DHS-wants-to-use-human-body-odor-as-biometric-identifier-clue-to-deception/UPI-20121236627329/, 3/9/2009.

[48] Yuan, J., & Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America,* 123, 5.

[49] M. Zuckerman, B. DePaulo, and R. Rosenthal. 1981. "Verbal and non-verbal communication of deception." In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Academic Press, New York, 1-59.