Literature review for Deception detection

by

Guozhen An

A Second Exam submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2015

Literature review for Deception detection

by

Guozhen An

Advisor: Pr. Andrew Rosenberg

Abstract:

Deception is a very common phenomenon, especially human communication, and people have been interested in how to accurately detect deception for much of human history. In recent years, detecting deception has became a huge point of interest for many different fields of research, such as business, jurisprudence, law enforcement, and national security. Deception and its detection is a complicated psychological behavior which is related to cognitive processes and mental activity.

There are several major problems in detecting deception. Firstly, there is huge individual difference among the liars (interpersonal difference). Secondly, the differences between truth telling and lying are typically small for same person (intra-personal difference). A third difficulty for detecting deception is embedding lies in the truth, where true statements are used to support a lie. For the difficulty perspective of computer science and other broad research on this topic, there are very few well designed deception corpora available for analysis and for developing deception detection systems.

In this survey, we will introduce some analysis of verbal cues for detecting deception through lexical analysis, acoustic and prosodic analysis, and speech event analysis. We will also present an "indirect" approach. We think emotion recognition and personality recognition might be informative in addressing inter-personal and intra-personal differences in detecting deception. Motivated by this purpose, we included some successful and interesting research for emotion and personality recognition in this survey.

The overall objective of this survey is to help readers to understand the challenge of deception detection, the state-of-art research for deception detection, and the future research direction for this topic.

Contents

Chapter 1

Introduction

Deception is a very common phenomenon, especially human communication, and people have been interested in how to accurately detect deception for much of human history. Deception is an act or statement intended to make people believe something that the speaker does not believe to be true, or not the whole truth. Of course, unintentional behavior that leads to an untrue belief, such as honest mistakes, or mis-remembering, is not considered to be deception. In recent years, detecting deception has became a huge point of interest for many different fields of research, such as business, jurisprudence, law enforcement, and national security. Deception and its detection is a complicated psychological behavior which is related to cognitive processes and mental activity. Therefore, this topic also holds a significant interest in psychology [1, 2, 3, 4]. Most of the studies in the literature on deceptive behavior have looked at human perception and analyzed deceptive cues including facial [5], gestural [6, 7, 8], and biometric indicators [9].

People lie for different reasons. From a psychological perspective, There are two types of deception, face saving deception and malicious deception. How people feel about deception depends on the reason for telling the lie [1]. Face saving deception is, generally, socially acceptable, and possibly not necessary to detect, but malicious deception is more important, and this is where deception detection is focused. People tell the lies for a number of reasons, for example, some people lie to protect themselves, some people lie to avoid tension and conflict in social interaction, some people lie to minimize hurt feelings and ill will [10]. There are a number of situations where knowing if someone is being deceptive is critically important. For example, a viewer may want to know whether a politicians public speech is really the truth; teachers want to know if the students have cheated during an exam; the customer wants to know whether the product is as good as the salesperson says; an interviewer wants to know the candidate is capable of doing the job as he or she claims; a security officer wants to know whether the visitor really has no harmful intent when entering the building; and the police wants to know whether a suspect is guilty or not. Successfully detecting deception in these situations would benefit not only individuals but also the society as a whole [11].

There is always an interest for whether there are discernible differences between truth teller and liars [12]. In psychology, researchers focus on an individuals non-verbal cues and verbal cues that human perceiver can use to detect deception, for example, non-verbal cues includes eye gaze, body movement, facial gestures, posture, and so on; and verbal cues encompass speech content (pitch, accent, and lexical meaning). Researchers assume that there is distinct psychological activity accompanied with deception, and people think some of these cues may leak through when people lie. Thus people usually intend to detect deception based on non-verbal cues, such as body movement and eye gaze, and ignore or pay less attention to the verbal cues. [13] states that police usually pay more attention to the non-verbal cues than verbal cues, and the result of this is less accurate recognition of deception. Furthermore, Meta-analysis of verbal and nonverbal cues for deception shows that speech related cues are more accurate than nonverbal cues [14, 15]. Finally, some other papers [16, 17, 18] shows that paying sole attention to speech content is more accurate in discriminating truths and lies than paying attention to nonverbal cues. In this paper, we will mainly discuss the importance of verbal cues for deception detection, because deception research has shown that many speech-related cues are more predictive of deception than non-verbal cues, and there are more informative signals in the verbal cues.

There are few studies in computer science about deception, and still fewer studies for automatic deception detection. Some previous research worked on facial expression [5], gesture [19, 20], body movement [6, 7, 8] and other nonverbal cues [21], but there is a small amount of work focused on verbal cues [22, 23, 24, 25], or speech analysis [26, 27, 28, 29, 30]. The present work on speech analysis is more focused on affective computing, such as emotion recognition, personality perception. Many of these techniques can also be applied to deception detection, identifying acoustic, lexical, prosodic and paralinguistic correlates (emotion recognition and personality computing) of deception.

There are several major problems in detecting deception. Firstly, there is huge individual difference among the liars (interpersonal difference). For example, some people raise the pitch of the voice when lying, while some lower it significantly; some tended to laugh when deceiving, while others laughed more while telling the truth. Secondly, the differences between truth tellers and liars are typically small for not only different people but also for same person (intra-personal difference). For example, both liars and truth tellers exhibit emotional and cognitive changes; recognizing which of these are indicative of deception can be challenging. A third difficulty for detecting deception is embedding lies in the truth. When deceiving, people usually tell a lie within some truthful context, in order to convince other people to trust the lie. The embedded lie is extremely difficult to catch. For example, police want to know a mens activity on Tuesday evening, and he tells the details of what he did on Wednesday evening instead of Tuesday, then police can not really distinguish the deception because of quality and quantity of the details in his statement is so realistic except the date. Usually lies that are embedded in truthful statements has very high quality details associated with credible statements, which lead a lie detector decide to trust the deceptive statement [11]. For the difficulty perspective of computer science and other broad research on this topic, there are very few well designed deception corpora available for analysis and for developing deception detection systems.

Automatic emotion recognition may provide helpful techniques for deception detection because of the intra-personal difference between deceptive and truthful speech. The emotions of a person change when telling a lie. [11] states that, compared with truth tellers, liars may experience stronger emotions, may experience higher levels of cognitive load, and are inclined to use more and different strategies to make a convincing impression on others. So it shows a possibility that computer scientists can apply automatic emotion recognition techniques to identify changes of emotion state for helping deception detection. In recent years, automatic emotion recognition has been a hot topic in speech processing, and many approaches successfully recognize emotion by using wide range of techniques. These approaches may also useful to detect deception by recognizing emotion state variation. We will present some approaches in the following section introducing the most useful techniques that can be applied to deception detection.

We know that people lie differently [11], one possible explanation for these differences may be their personality. Based on several decades of research and experiment, the Big-Five or Five-Factor Model was introduced by [31], which is the dominant paradigm in personality research, and one of the most influential models in all of psychology [32]. [33] found that different people lie in different ways, some people raise their pitch when lying, while some lower it significantly; some tend to laugh when deceiving, while others laugh more while telling the truth. [29] also found that judges with different personalities perform at different levels of accuracy when they detect deception. We can hypothesize that personality may also provide useful information in predicting individual differences in deceptive behavior of the speakers they judged. We hypothesize that personality analysis will be a useful cue in analyzing interpersonal differences for different lying styles. For conducting the personality analysis, we need to first detect the personality accurately in order to do further processing. In recent years, there have been many successful approaches for personality detection. We will introduce some of most useful approaches that can be applied to deception detection.

The rest of the paper is organized as follows: Chapter 2 introduces some current existing deception datasets and deception detection approaches. Chapter 3 introduces some emotion datasets and automatic emotion recognition approaches. Chapter 4 describes the state of the art in automatic personality recognition. Chapter 5 draws conclusions and highlights promising future directions in the automatic detection of deception.

Chapter 2

Deception

In recent years, there has been increasing interest in automatically detecting deceptive behavior, particular from law enforcement and government agencies. There have been many techniques to detect deception developed. Some of them have demonstrated success at detecting deception using either verbal and non-verbal cues, such as body movement [6, 7, 8], eye gaze [21], cognitive load [34, 35], facial expression [5], and brain imaging [36] to measure non-verbal cues [14]; pitch, accent [37], energy, lexical, statement analysis[38] for verbal cues. The polygraph is a typical example for detecting deception by measuring non-verbal cues with specific equipment. The polygraph measures and records several physical properties including blood pressure, pulse, respiration, and skin conductivity during the subject ask and answer the questions [39]. The theory for the polygraph is that people believe there will be some kinds differences of physical responses between truth teller and liar, the deceptive answers can be identified from non deceptive answers by measuring these physical properties. In 2003, National Academies of Science state that polygraph testing can detect deception above chance, though below perfection [40].

Deception has been studied by psychology, psychiatry, linguistics, and philosophy for decades, and there are many different definitions for deception. We think the one from [2] is the clearest and most applicable to the research in detecting deception. ”Deception is a successful or unsuccessful attempt, without forewarning, to create in another a belief which the communicator considers to be untrue.” Deception includes several types of communications or omissions that serve to distort or omit the complete truth. It is important to note that lying is an intentional act and that unintentional misremembering is not considered deception under this definition [11].

The table 2.1 shows the results of a recent meta-analysis [41] including 108 studies that represent the deception detection ability of deception detection from different types of individuals. The result shows that even performance of professionals is not very accurate. Surprisingly, the highest performance is from the group of teachers. It maybe that teachers have a lot experience of deception detection during the school life, and it improves the skill of detecting deception for them. The average performance for normal people is about chance, and it explains how difficult is the deception detection for people.

There is a lot of research on non-verbal cues, such as facial expression [5], eye gaze [21], body gestures [6, 7, 8], brain imaging [36], and the standard biometric indicators commonly measured by polygraph [42]. However, eye gaze, facial expression and body gesture analyses are not reliable methods for identifying deception, and polygraphs are available only at significant expense, and they are fairly invasive. [13] deception research states that many verbal cues are more predictive of deceit than non-verbal cues. This paper gives an example of police officers who typically pays more attention to non-verbal behavior than verbal behavior, because they believe that suspects are less able to control their non-verbal than verbal

Table 2.1: *Experimental Results on Deception Detection.*

Group	Subjects	Accuracy
Teachers	20	70 %
Social workers	20	66.25 %
Criminals	52	65.40 %
Secret service agents	34	64.12 %
Psychologists	508	61.56 %
Judges	194	59.01 %
Police officers	511	55.16 %
Customs officers	123	55.30 %
Federal officers	341	54.54 %
Students	8876	54.20 %
Detectives	341	51.16 %
Parole officers	32	40.42 %
Total	11052	54.50 %

behavior, and non-verbal cues to deception are more likely to leak through. However, the experiment detailed in the paper shows that paying attention to non-verbal cues leads to less accurate results than verbal cues, especially when only visual non-verbal cues are considered. [13] even states that there is no nonverbal behavior that is uniquely associated with deception. As previously stated, a specific behavioral indicator of deception does not exist. There are, however, some nonverbal behaviors that have been found to be correlated with deception. [13] find that examining a ”cluster” of these cues yields a significantly more reliable indicator of deception than examining a single cue. Therefore, [11] recommended that the combination of non-verbal and verbal cues would reach a higher accuracy for detecting deception.

In this survey paper, we will mainly discuss verbal cues for deception detection. There are several reasons to focus on verbal cues. Firstly, it is convenient and more easily portable than detecting deception via non-verbal cues. In addition, prior research has found that verbal cues are more predictive and reliable than non verbal cues for deception detection

[14, 15, 16, 17, 18].

We will introduce some verbal cues for deception detection research recently. There are a few papers about verbal cues for detecting deception, and usually they fall into two different categories: text-based and speech-based. These approaches treat deception detection as a classification problem, where data is collected and annotated for deceptive status (i.e. truth vs. lie). Then, these labels are used to train a classifier via supervised learning. Therefore, these approaches have two main stages: feature extraction and learning method. For feature extraction, there are several major signals: lexical features [22, 23, 24], acoustic and prosodic features [26, 27], and speech event features [28].

Following feature extraction, a system needs some method of determining the label of an utterance – truth vs. lie – depending on the number and type of features. Support vector machines (SVMs) are the most common, since SVMs can deal with large number of features efficiently, which also can handle the overfitting and predicting error very well. Therefore, We default to SVMs as the learning algorithm, but we also look at other algorithms such as Naive Bayes and decision trees. These will be discussed in section 5.

2.1 Deception Data

One of biggest limitations of deception research is that there are few carefully annotated data sets available which are accurate and well designed. Generally, it is difficult to obtain deceptive data with annotation in real life, thus many researchers design their own deceptive datasets by hiring subjects to act in a deceptive manner. The Columbia SRI Colorado corpus is a useful speech-based corpus built in this manner, and the Tripadviser dataset [22] is another deceptive dataset which is text based.

The Columbia SRI Colorado corpus was introduced by [26]. [27, 28, 29, 30] used this corpus in their research. The motivation of this data collection is that existing corpora are difficult to analyze due to the varying recording conditions. They hired 32 native American English speakers with a balanced gender from the Columbia university community. Subjects answered questions and performed activities in six areas: music, interactive, survival skills, food and wine knowledge, NYC geography, and civics, and the subjects received financial incentive for both deceptive and non-deceptive speech.

In the experiments, subjects were told to perform a series of tasks in six areas. They were told to tell the truth in two areas, and attempt to deceive the interviewers in four others. Interviewers had to determine the whether subjects were lying or telling the truth, and were also allowed to ask additional questions about these tasks except the ones they had already performed. During the interviews, subjects provided true or false label for each utterance they made by pressing one of two pedals hidden from the interviewer under the table.

The recording for each subject lasted between 25 and 50 minutes, and the entire corpus contains 15.2 hours of dialogue, approximately 7 hours of deceptive and non-deceptive speech. This corpus is the first audio corpus that can permit sophisticated speech analyses to be performed, which also includes not only ground truth information for big topics indicated by subjects, but also deception labels on a per-turn basis annotation. They divided the data into word, slash unit, breath group (phrasal units determined automatically from pause, intensity and subsequently hand-corrected), and turn units, derived by combining automatic procedures and hand transcriptions. Breath group and speaker turn units were derived semi-automatically.

While subjects attempted to deceive the interviewers on topics, not every sentence in a topic was a lie. As a result, [26] introduces two new terms, 'little lie' and 'big lie'. A little lie is defined as a deception attempt on a per-turn set of utterances, and a big lie indicates a per-topic deception attempt. Based on this corpus, the researchers are able to extract not only acoustic prosodic features and lexical syntactic features, but also some other features like filled pause features and critical segment features.

Tripadviser-gold dataset was introduced by [22], while [23, 24] used this data for their research papers. This dataset was gathered 400 truthful reviews from www.tripadviser.com, and 400 deceptive reviews were gathered by using Amazon Mechanical Turk, evenly distributed across 20 most popular Chicago hotels. Amazon Mechanical Turk is a source crowding service provided Amazon, and it made large-scale data annotation and collection efforts financially affordable by anyone who has basic programming skills (basic HTML, XML, JAVAScript).

For deceptive reviews, they created 400 Human Intelligent tasks (HIT) evenly distributed across 20 chosen hotels from the Chicago area. When they create the HIT, they restricted to only US turker with approval rate more than 90% only, and the maximum time for finishing single HIT is 30 minutes with one dollar incentive. The HIT instruction represented a name and website of the hotel, and asked the turker to assume work at hotel market department, and write a fake review because their boss asked them to do it. Finally, they request the review should sound realistic and positive. They rejected any HIT if the review was not reaching certain requirement (e.g, written for wrong hotel, unintelligent, unreasonably short).

For truthful reviews, they gathered from TripAdvisor, the 20 most popular hotels from the Chicago area, for which there were total 6977 reviews available. They eliminated all non 5 star reviews, non English reviews, reviews with less than 150 characters, and any reviews written by first time author. Then, they select 400 reviews from 2124 reviews which qualify these requirements, and these reviews were evenly distributed cross the same 20 hotels. Finally, they got 20 deceptive reviews and 20 truthful reviews for each hotel with minimum 150 characters length reviews. These deceptive and truthful reviews are text based data set for deception research, and researchers have performed many different experiments with them.

In order to develop an effective deception detection system, we first need to obtain high quality deceptive data. Usually, researchers obtain these deceptive data by hiring people to deceive in an experiment, but the more realistic the data is, the more useful system we can expect in real life. Compared to the experimental data, in real life, we may not get the correct annotation for the spontaneous speech at most time, because there is no way to get the ground truth for spontaneous speech for most of the time. This means we also can not obtain the performance of our system accurately, and it is difficult to improve the system for future use. So if our experiment data is more close to the real life data, then we can make a better system which is more suitable for real life situations.

2.2 Features For Deception

Once annotated data has been identified, the second step for deception detection is the extraction of relevant features. Many approaches to verbal detection deception use three types of features: lexical features, acoustic and prosodic features, and speech event features. In this section, we will introduce all major techniques and approaches about these three types of features and relevant feature extraction techniques.

It has been mentioned in previous research [25, 19, 43] that deceivers usually use different patterns of word usage when they are lying, therefore lexical analysis is useful to detect deception. We will introduce several kinds of major techniques for the lexical analysis: Linguistic Inquiry and Word Count (LIWC), Dictionary of Affect in Language (DAL), Part-of-speech (POS), and N-gram.

Linguistic Inquiry and Word Count (LIWC) is a text analysis software program designed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis [44]. LIWC calculates the degree to which people use different categories of words, and can determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions. LIWC is good at detecting the psychological properties. A lot of research [25, 26, 22] has used LIWC to study the various emotional, cognitive, structural, and process components present in individuals' verbal and written communication. Thus, for deception detection, LIWC is often used for detecting emotional and cognitive changes from verbal content.

[25] applied this technique in their research and reported 67 percent accuracy in detecting deceptive speech using logistic regression; [26] also used this method to detect deception

combine with prosodic/acoustic and lexical features, and the accuracy was 66.4 percent of their data. They found that the presence of positive emotion words is the best indicator of deception; deceptive speech has a greater proportion of positive emotion words than truthful speech. [22] is another example of using LIWC to detect deception. They used LIWC in combination with POS, Unigram, Bigram and Trigram features to train Naive Bayes and SVM classifiers. They found that truthful opinions tend to include more sensorial and concrete language than deceptive opinions, in particular, truthful opinions are more specific about spatial configurations. The result of their experiment is surprisingly high, which is 91.2 percent accuracy. The result here is much higher than others or even human performance, I think the main reasons behind this is the data. First, the data set used here is very different from others. The Tripadviser-gold dataset is very narrow and specific dataset containing only reviews of Chicago area hotels. Next, they gathered the deceptive reviews through Amazon mechanical turk, and the quality of these deceptive reviews is not guaranteed compare to other datasets. Therefore, this demonstrates that while people can develop a deception detection system which perform better for the narrow and specific area, it remains difficult to make a system perform well in general.

Another tool for lexical analysis is the Dictionary of Affect in Language (DAL) [45] which is used for analyzing emotive content of speech. The Dictionary of Affect in Language lists approximately 4500 English words, a rating for Pleasantness (Evaluation) and rating for Activation (Arousal) is associated with each word in the Dictionary. Arousal is a state of heightened activity in both our mind and body that makes us more alert [46]. The difference between DAL and LIWC is that the focus of DAL is narrower than LIWC, it only addresses the emotional meaning of words. [26] used DAL to distinguish the different emotional state between deceptive and truthful speech. They found that higher average pleasantness score is more likely to be deceptive, and higher pleasantness standard deviation is less likely to be deceptive.

People have investigated the syntax in deceptive and truthful speech for cues to deception. Part-of-speech (POS) tags are a shallow representation of the syntax of an utterance. Part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its contexti.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. [22, 23] used the POS tag technique to extract some features for detecting deception, and the performance is also quite good.(2.2) [22] identified deceptive opinion spam based on frequency of POS tag by comparing truthful and deceptive reviews, because truthful and deceptive opinions might be classified into informative and imaginative genres [47]. They found that deceptive opinions contain more superlatives indicating that deceptive writing may contain exaggerated language.

Another popular language model used by many people for deception detection is inspection of N-gram. An N-gram is a contiguous sequence of n items from a given sequence of text or speech. Usually many approaches combine the N-gram with other features to detect deception, [22] combined LIWC with N-gram to train the Naive Bayes and SVM classifiers, both of which have performed efficiently on lexical analysis for deception detection.

The best result from [22] is 89.8% accuracy, as well as the [23] reaches 91.2% at highest performance. These result are much better than most of other deception research, and the reason behind this is that computer can do extremely better than human on catching lie in specific and narrow topic.

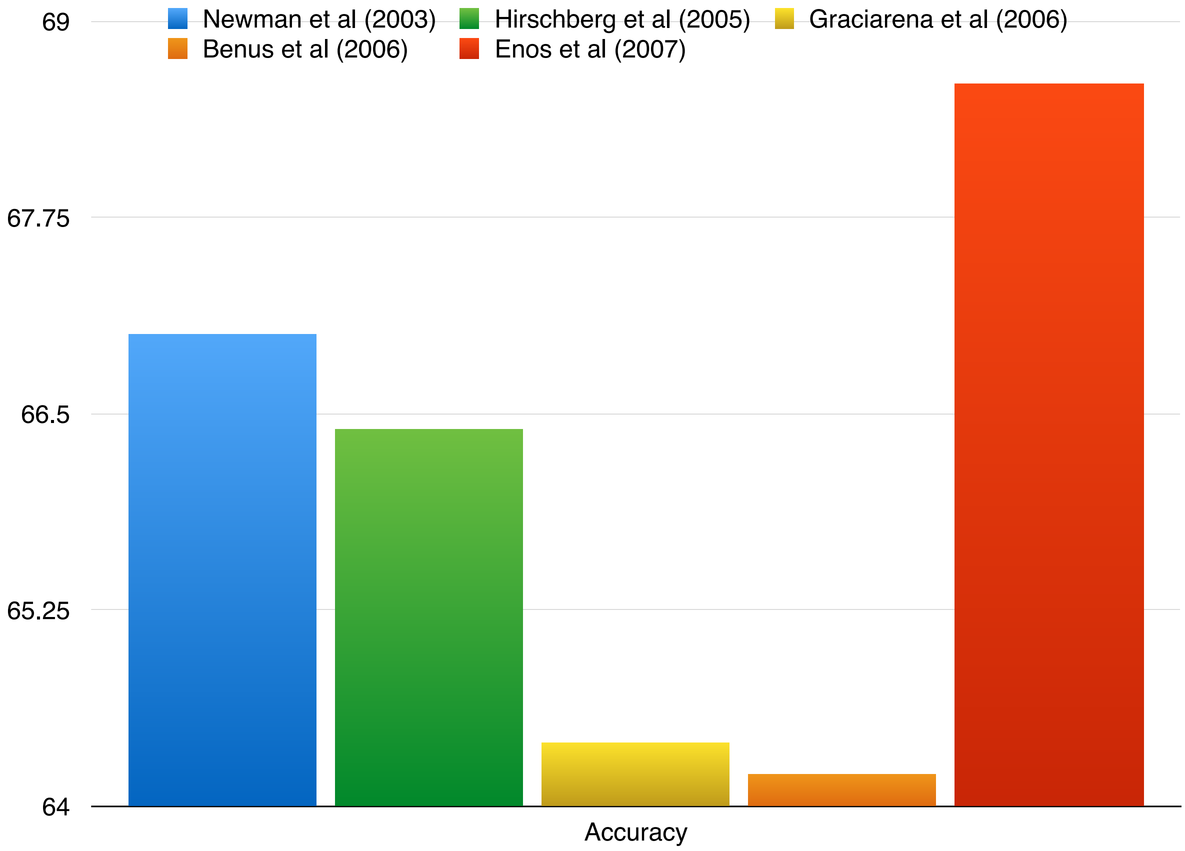
For the speech based deception detection, acoustic and prosodic features are used very often to identify the differences between deceptive and truthful speech, because pitch, energy, speaking rate and other stylistic factors may vary when speakers deceive [14]. [26, 27] used the same acoustic and prosodic features for their experiment in different training methods.

[26] used a wide range of potential acoustic and prosodic features. They extracted and modeled features including duration, pausing, intonation, and loudness, associated with multiple time scales from a few milliseconds to an entire speaker turn. Features were automatically normalized, taking into account long term speaker-specific habits as well as segmental context. Pitch, energy and durational features were computed by using tools from an automatic speech recognition system. A large set of statistical features was computed, including maximum/minimum/mean pitch, range of pitch number of frames that are rising/falling/double/halved/voiced, length of the first/last slope, number of changes from fall to rise. For energy features, raw energy for sentence unit, raw energy for voice unit, maximum/minimum/mean energy were extracted. Finally they extracted several durational features, which include maximum and average phone duration in the sentence unit. At the end, they examined the usefulness of these acoustic prosodic features in distinguishing deceptive speech, and the result shows that energy and f0 features play positive roles for deception detection among the wide range of acoustic and prosodic features.

In the literature, speech events such as filled pauses and laughter are important to distinguish the interpersonal differences on communication behavior [48, 49]. Therefore, speech event features are also a useful predictor of deception in speech. In [28], they investigated the relationship between filled pauses and deception. They found several interesting results. Firstly, they looked into the presence and absence of filled pauses in deceptive speech, and they found that subjects used the filled pauses significantly more frequently in locally truthful than locally deceptive statements; turn-internal silent pauses also occurred more frequently in truthful than in deceptive speech; the lie contains significant less pauses than the truthful statement, and the pauses before lie is longer than before the truth. Secondly, they also did research for the differences between um and uh, and found that um was more likely to be followed by a silent pause than uh; a tendency for uhs to occur in utterances that were locally truthful but the subjects were expressing a global lie. In conclusion, people tend to be more careful about their word during lying than during telling the truth, and local deception does correlate with the use of um more than with the use of uh, um is longer, and has longer latencies with more silent pause surrounded.

Figure 2.1 contains the result of speech based deception detection system from following paper [25, 26, 27, 28, 30]. From this, we can observe that the performance state-of-art deception detection system is about 65-70%, and it is almost close or even better than the best human performance for deception detection task. But we think there is still chance for improving the system to reach higher performance. Because the state-of the art approach have not yet considered personal difference such as personality, and cultural differences. Personal difference is the one of the most difficult challenges for detecting deception. If we can have some successes in this part, we believe there will be a lot of improvement in deception detection performance.

Figure 2.1: *Experimental Results on Deception Detection.*



Chapter 3

Emotion Recognition

Automatic emotion recognition might be a efficient way to help deception detection because of intra-personal difference during deception. A same person may experience different emotions at different times when telling a lie versus telling the truth. [11] states that, compared with truth tellers, liars may experience stronger emotions, may experience higher levels of cognitive load, and incline to use more and different strategies to make a convincing impression on others. Therefore we can apply automatic emotion recognition techniques to obtain the change of emotion state for deception detection.

Automatic emotion recognition from speech has been a very active research area for the last decade, and many researchers have been working on this topic. There are plentiful successes in emotion recognition, and we will discuss some approaches from recent years in this article. Human language carries various kinds of information, which is not only from lexical level, but also from acoustic and prosodic level. The original motivation of automatic emotion recognition is that how to use speech to let the computer understand human instead of using a traditional input device to the computer, not only make computer understand the lexical content, but also let computer understand the emotional content of speech. It's a difficult problem, but there are more and more successes in this field recently.

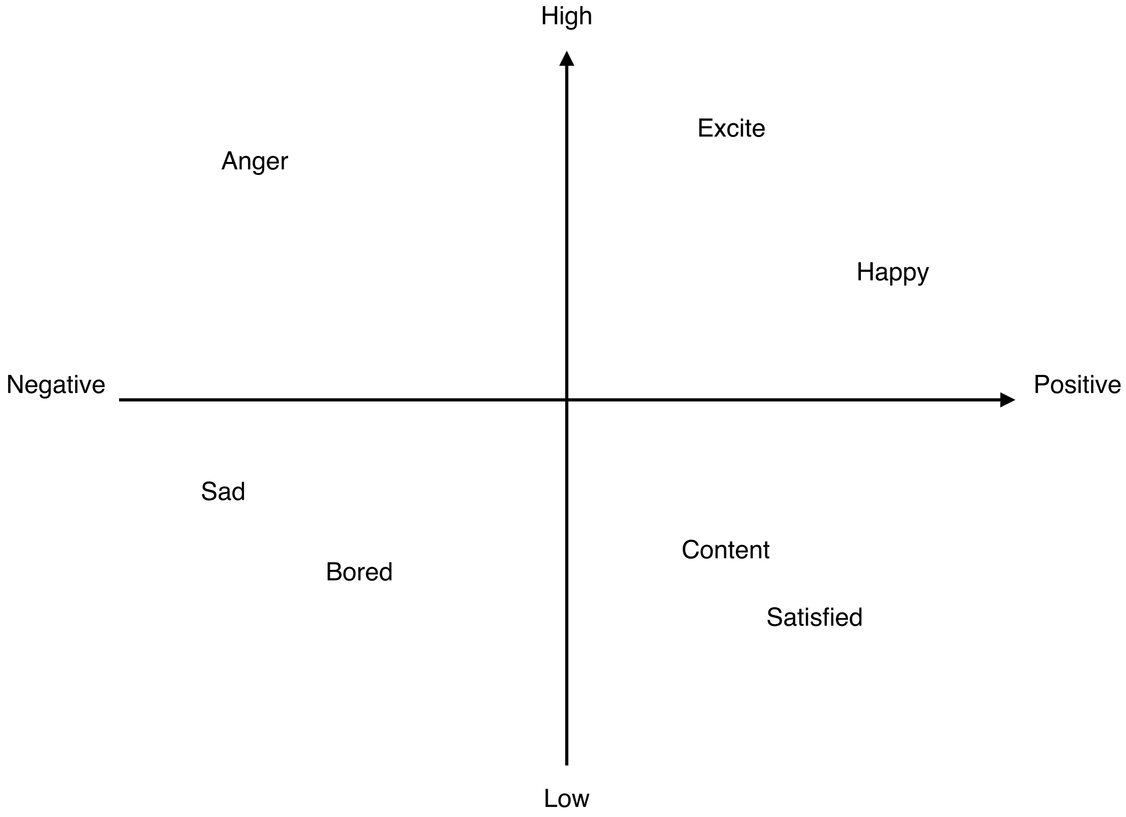
The process of emotion recognition is very similar to deception detection process, there are usually three main parts of research challenge for automatic emotion recognition, which are finding appropriate audio units, feature extraction, and classification. These three parts are also the major difficulties for automatic emotion recognition, although there are also some other details to care about, such as data collection part. We will review some approaches in this section to introduce the most useful techniques of emotion recognition for deception detection.

3.1 Emotion Data

In human language, information about emotion is conveyed via speech, in how and what the speech is being spoken. How it is spoken is more important than what is spoken, the pitch, intensity, tempo and energy of the speech is more important than the lexical content in the emotion aspect. [50] The acoustic of emotional speech different between people depending, in part, on personality. There are three types of emotion databases, simulated or acted emotions, elicited emotion and natural emotion. The data collection for automatic emotion recognition is also very challenging depending on the type of data. In general, automatic emotion recognition research begins with acted speech [51], and becomes to use more and more realistic data now [52, 53]. Simulated or acted speech is speech expressed intent in a professional manner. In most studies that use acted speech, the expression of a particular emotion is both intentional and overt. Indeed, many such studies simply instruct the actor to utter a given text with emotion X. The Berlin database of emotional speech [54] and the Danish Emotional Speech Corpus [55] are good examples of acted emotional speech databases. Natural speech is the spontaneously produced speech from real life. Elicited speech is neither neutral nor simulated, rather emotions are intentionally explicated by context. For example people are recorded in the lab with some task intended to elicit anger or irritation in the subjects. So far the acted speech from professionals is still the most reliable for emotion recognition research, because professionals can deliver emotion with a great strength and quality. Comparing data used for emotion recognition to that used in deception detection, the currently available corpora are closest to elicited emotion, because the emotion appearance of subjects is neither natural since it happens in laboratory environment, nor simulated because subjects really feel these emotions.

The labeled emotions usually are labeled along two dimensions, arousal and valence. Arousal is ”a state of heightened activity in both our mind and body that makes us more alert [46]. Valence is ”the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation” [56]. Valence ranges from positive to negative, and arousal ranges from high to low. Using these ranges, emotions can be mapped onto four quadrants, specifically: positive high (e.g. happy), positive low (e.g. content), negative high (e.g. angry) and negative low (e.g. bored). We find very few datasets labeled using such a continuous scale for arousal and valence, perhaps because producing regressions for these scales is highly complex. More commonly, a lot of emotion datasets use labels for basic emotions like anger, happiness, sadness, and fear, because these basic emotions are easier to recognize, measure and classify.

Figure 3.1: *Two dimensional emotion space Valence and Arousal.*



3.2 Audio of Emotion Unit

In order to recognize emotion from speech, digitalization and acoustic preprocessing is required, and also speech must be segmented into meaningful units. Previously, there has not been too much research attention to this topic. In part, this is because most of research has used acted emotion speech datasets, which have obvious and unchanged emotions within each single utterance. However, for spontaneous speech, these emotion unit does not exist, a person can demonstrate an emotion for a relatively brief amount of time within a longer utterance or dialog. We need to find the proper unit of emotion efficiently and accurately. A good unit should fit two requirements, first, it should be long enough to reliably calculate features by statistical functions, second, it should be short enough to have stable acoustic properties with emotion segmentation. For features calculation of global statistics over the whole unit, the unit has to have a minimum length to do the feature extraction. If the length of the segment is too short, then it is unstable and inaccurate to extract statistical measure based on the large size of values. On the other hand, in order to capture the emotional state, the unit should be short enough that no emotion variation exists in the unit. So segmenting the proper unit of audio is very important not only in acted speech but also more for spontaneous speech, because it is more difficult to find the proper unit from spontaneous speech than in acted speech.

There are several major types of approaches for segmenting audio for emotion recognition. These include segmenting the audio based on dialogue turn, utterance, word, and phrase [52, 53, 57, 58, 59, 60, 61, 62]. It is difficult to say which one is best, because different approaches are more effective for different datasets. For the deception detection, we think the dialogue turns may be the best for emotion recognition, because truth and lie domains are labeled on the dialogue turn, and it will give the consistency between the deceptive label and the emotion label which obtained from emotion recognition.

3.3 Feature of Emotion Recognition

After selecting the unit of analysis, extracting the relevant features is the second step for an emotion classification system. The goal for the feature extraction is to represent the acoustic audio signal as a multidimensional feature vector. There is not yet agreement on what the feature set is most effective for automatic emotion recognition, and it is very much data dependent. A high number of features often does not perform well because most of classifiers are negatively influenced by redundant features. Thus people usually include many features in the beginning of the process, and apply feature selection methods afterward. Many approaches use energy and pitch related features, while formant and Mel Frequency Cepstral Coefficients (MFCC) are also used in many places. Some papers also mentioned durational, pause related features and different type of voice quality features as the starting feature set.

Pitch and energy features are the most frequently used features in the automatic emotion recognition process. These features calculated from acoustic variables by applying statistical functions within an emotion unit. The adoption of pitch and energy features becomes more detailed and specific as emotion recognition techniques develop and evolve. In [63], they found that short term acoustic features (pitch, formants, MFCCs) are the most useful features for emotion recognition, and the statistics of acoustic and prosodic features (mean, range, variance, pitch contour trends; mean and range of the intensity contour, rate of speech and transmission duration between utterances) are also contributed to the emotion recognition. In [64], the pitch and energy related features was also adopted to emotion recognition, they extracted pitch, intensity, distribution energy in the spectrum, and MFCC features for voice unit, and calculated global features (speech rate, intensity and pitch contour) from whole segment unit. Another interesting point in this paper is that they also used lexical features by using term frequency-inverse document frequency(TFIDF) method.

It is difficult to compare the features across published work, since different condition of each work made the performance of each system incomparable. Often different research use different datasets, and the emotion classes also varies for each research. For example, 50% accuracy of one four emotion class dataset is excellent, while another dataset meets 70% to 80% accuracy of two emotion classification. This does not mean the first approach did not do well, but rather that it is more difficult task than second one, and that can be due to many different reasons. Comparisons of features can be made through the relevance ranking by the information gain of single features or by rank in a sequential selection method. Usually, single feature can be selected by the relevance ranking method, but single feature relevance does not determine the usefulness of entire feature set. Another way of comparing features is grouping features by property to look at the performance of different combination of groups. People generally agree that pitch and energy related features are the most useful features for emotion recognition task until now [65, 66, 63, 64]. Almost every emotion recognition approach used pitch and energy features in their base feature set, but different approaches usually have their own best features from these feature sets depend on the specific task. Therefore, we think pitch and energy features are the most important features for emotion recognition task, and these features will be the starting feature set for emotion recognition of deception detection process, then apply the feature reduction method and feature analysis method to select the best performed features for the deception detection task.

Chapter 4

Personality

Personality psychology is the modern answer to such an ancient question: What is the personality? As a construct, personality aims at capturing stable individual characteristics, typically measurable in quantitative terms, that explain and predict observable behavioral differences [67]. [68] Finding consistent individual differences between humans is the main reason for the interest in this task within computing and human machine interaction.

Personality plays an important role in deception detection. [11] states that one of major difficulties of deception detection is interpersonal difference, and we think personality recognition is a key to solving this difficulty. Psychologists [33] state that there are large individual differences in peoples behavior such as lying, some people raised their pitch when lying, while some lowered it significantly; some would to laugh when deceiving, while others laughed more while telling the truth. They also found that judges with different personalities perform differently when they detect deceit [29]. Therefore, we can hypothesize that personality tests may also provide useful information in predicting individual differences in deceptive behavior of the speakers they judged. Furthermore, personality analysis will be a useful cue to analyze interpersonal differences in different lying styles. For doing the personality analysis, we need to first detect the personality accurately in order to do further process. In recent years, there are many interesting approaches for personality recognition.

Personality in computing is similar to any other affective computing topic; understanding, predicting, and synthesizing human behavior are the main goals of the research. The main problems of the computational personality area are automatic personality recognition, automatic personality perception and automatic personality synthesis regardless of data and methodology. The difference among them is that automatic personality recognition is the recognition of the true personality of an individual, automatic personality perception is the prediction of the personality others attribute to a given individual, and automatic personality synthesis is the generation of artificial personalities through embodying agents. Therefore, personality recognition is best suited for deception detection purpose, because we want to find the relation between deception behavior and the personality for different individuals. Then in order to do the personality analysis for deception, personality recognition is the first step.

By definition, personality refers to individual differences in characteristic patterns of thinking, feeling and behaving [69]. Many approaches used the five factor model to measure personality. The big five or five factor model [31] is the dominant paradigm in personality research, and one of the most influential models in all of psychology. [32]

The big five traits as defined by [31] are as follows. For clarity and precision we are including the definitions verbatim.

Openness to Experience. This factor is designed to capture imagination, aesthetic sensitivity, and intellectual curiosity. Those who score low on this dimension prefer the familiar and tend to behave more conventionally. Openness is of particular interest in deception detection since it addresses the degree to which a listener might be willing to set aside preconceptions and take in all aspects of an immediate situation, which in the case of our experiment comprises the behavior of the speaker in the given context. It also seems reasonable to expect that a listener who scores high in Openness would be more able to defer judgment (specifically in terms of the deceptiveness of the speaker) until s/he has observed all available information rather than making facile conclusions, a trait surely of use in this context.

Conscientiousness. Conscientiousness addresses individual differences in the realm of self-control. Contrasts measured by this dimension include those between determination, organization, and self-discipline in high-scorers and laxness, disorganization, and carelessness in low-scorers.

Extraversion. This factor is intended to capture an individuals proclivity for interpersonal interactions, and describes variation in sociability. It reflects contrasts between those who are reserved and outgoing, quiet and talkative, and active and retiring.

Agreeableness. Agreeableness is a measure of a class of interpersonal tendencies, and its meaning is slightly unintuitive when compared to the usage of agreeableness in common parlance. At its base, Agreeableness is a measure of an individuals fundamental altruism, and individuals high in Agreeableness are sympathetic to others and expect that others feel similarly.

Neuroticism. Neuroticism contrasts emotional stability with maladjustment. It is intended to capture differences between those individuals prone to worry versus calm, emotional versus unemotional behavior, and hardiness versus vulnerability.

4.1 Personality Data

Not surprisingly, data plays a significant role in personality computing, and there are many different datasets specifically for personality recognition. However, one of the biggest limitations for data is that there is no widely accepted standard benchmark to test the different approaches. One widely used dataset is the SSPNet Speaker personality corpus, which was used in Interspeech 2012 Speaker Trait Challenge [70]. This corpus includes 640 clips from 322 subjects who spoke French, each clip is about 11 seconds long. There is only one speaker in each clip to avoid any confusion, and the speech is emotionally neutral and does not contain any words could be easily understood by an individual who does not speak French. There are 11 judges who listened to all the clips, and filled the BFI-10 [71] questionnaire for each clip. The BFI-10 questionnaire is a personality assessment questionnaire commonly applied, and it aimed to calculate the score of the big five traits dimensions. The judges worked separately through the whole process, and they didn't know each other. They werent allowed to work more than 60 minutes per day for these tasks, and the order of the clips for each judge was changed to present the task from being exactly the same and repetitions to avoid error or tiredness of the judges. The final label of the clip is an average of the score from all judges for the given trait.

For the SSPNet Speaker personality corpus, the biggest issue is how accurate is the labelling of the big five traits. The label of the personality is really depends on the judges subjective perceptions. though they used the average score of the 11 judges. So the potential solution is let the subjects fill the NEO-FFI inventory [72] to identify the big five trait for each subject, but then it will sometimes be impossible or more expensive than reusing audio from broadcast news and having observers labelling the perceived personality.

4.2 Personality Features

There are many different approaches for personality computing, we will discuss several typical approaches for this task. For the text based personality recognition, we will review Mairesses paper from 2006 [73, 74] which used a lexical analysis of personality recognition which performed well. Then, we will go over various approaches from the Interspeech 2012 Speaker Trait Challenge. [75, 76, 77, 78, 79, 80, 81, 82, 83]

Psycholinguistics show that the choice of words is not only driven by the meaning of the words, but also by psychological conditions, such as emotion, personality and relational attitude. Therefore, it is possible to detect personalities through text analyses associated with psycholinguistic techniques. For lexical analysis for personality recognition, a lot of approaches used several very fundamental and widely known techniques which we mentioned in the deception section, such as LIWC, N-gram and parts of speech. Moreover, the MRC psycholinguistic database [84] is used for lexical analysis of personality. The MRC psycholinguistic database contains approximately 150000 psychological words, and this dictionary can be used in computational experiment of psychology or linguistic.

In [73], they used a combination of LIWC,MRC, utterance type features and some prosodic features as their base feature set. They extracted a set of linguistic features of essay and conversation transcripts with 88 word categories from LIWC, and calculated the ratio of the words in each category motivated by the correlation between those features with big five dimensions of personality [85]. Next, they also extracted 14 additional features by averaging word feature counts from the MRC Psycholinguistic database [84], which contains 150,000 words. Then, utterance type features were introduced; they used parse tree to tag each utterance as a command, prompt, question and assertion. In total, 4 utterance type features were added to final feature sets. At last, they added 11 prosodic features computed by Praat characterizing the voices pitch, intensity, and speech rate. Finally, they trained the personality model by using RankBoost for each big five trait, and they found that LIWC is the most useful feature for most dimensions except neuroticism, on other hand, MRC performed outstandingly well on neuroticism but not for other dimensions; prosodic features were valuable in predicting extraversion by themselves; the utterance type features did not perform well on any dimension. In another paper from [74], they used an exact same feature set, but the Weka toolkitwas used to train and evaluate the different statistical models such as linear regression, regression tree, decision tree, Naive Bayes, and SVM. Finally, they found a linear regression perform poorly on every dimension, regression tree worked best for extraversion, Naive Bayes improved the prediction of extraversion, neuroticism, conscientiousness, however, nothing performed better regarding openness and agreeableness dimensions.

The Interspeech 2012 Speaker Trait Challenge [70] has led to the first rigorous comparison of different approaches over the same data and using the same experimental protocol. A large number of features and machine learning approaches have been proposed, and the results of the speaker trait challenge show that no particular approach is clearly outperforms than the others. Each approach has a dimension which it performs well and poorly in. Therefore, there is no one best strategy which works for every dimension. In general, it is a better choice to develop model for each personality trait separately instead of doing them together.

The corpus used in this challenge includes 640 speech clips in French which were 10 seconds for each, and there were total 322 subjects. The number of judges for each clip was 11, each judge listened to all the clips, and completed a personality questionnaire and calculated a score for the Big-five dimensions.

The challenge organizer gave all participants a baseline feature set with the performance result for this feature set as the start point, and they also released the train, developed and tested data without the personality rating for test data, and they only provide the performance of the test data at final submission. The baseline feature set is the 6125 features of openSMILE [86], and the performance measure was the unweighted average recall (UAR).

$$UAR = \frac{1}{K}\Sigma_{K}^{i=1}\frac{A_{ii}}{\Sigma_{K}^{j=1}A_{ij}} \tag{4.1}$$

[87]

Some of the papers submitted to this challenge worked on feature selection methods. The first one [75] started work from the challenge baseline feature set, and minimized the challenge feature set until to achieve a satisfactory performance with Gaussian Mixture models. They used a set covering framework to select the minimum number of features, and also selected the features with mutual information with respect to the personality trait. The best performance of this approach was 79.7%(UAR) on conscientiousness dimension, and this approach only improved the openness by 2% compared to the challenge baseline, but decreased the performance on all other dimensions of test set.

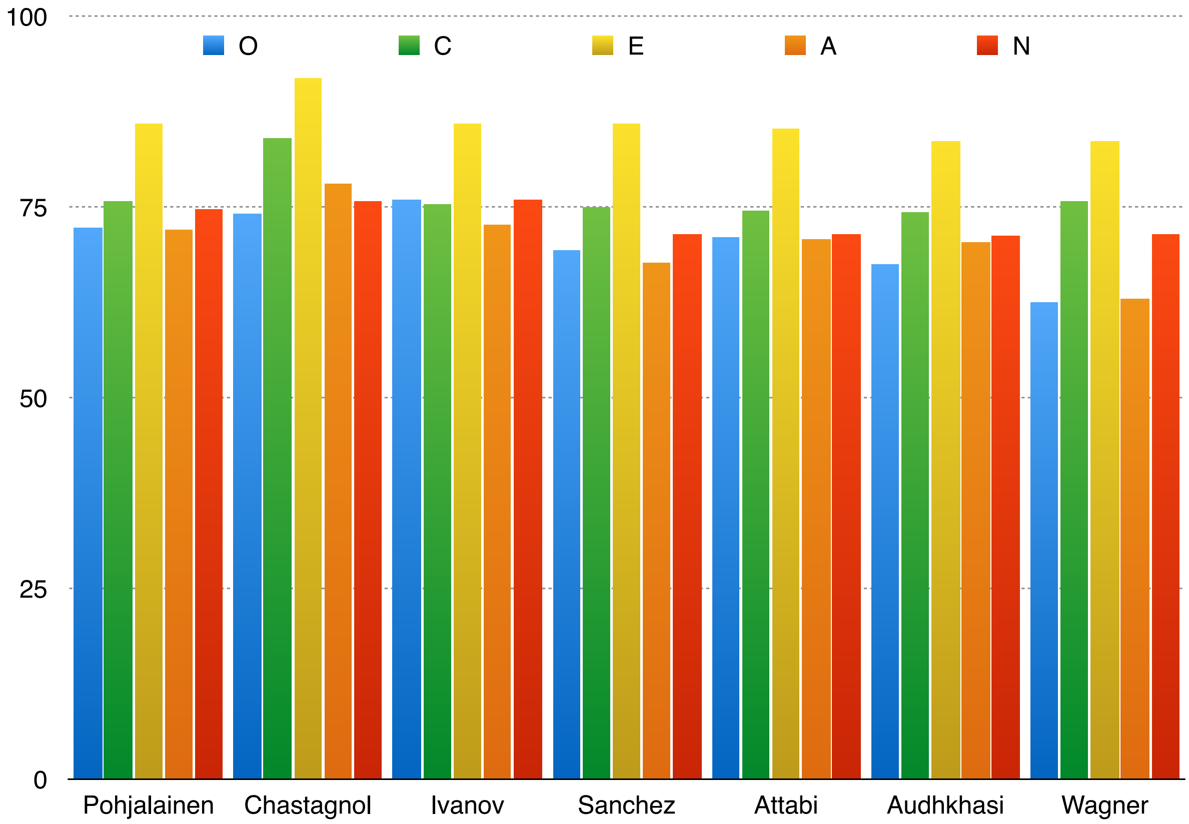
The next feature selection approach is the Sequential Floating Forward Search [76], they applied this method to the challenge baseline feature set. The Sequential Floating Forward Search(SFFS) is a combination of Sequential Forward Search (SFS) and Sequential Backward Search (SBS), SFS tries to add a feature to an existing well performing feature subset to increase the performance, and SBS tries to remove a feature from a subset to increase the performance without decreasing the performance. After feature reduction, the selected feature set is trained by Support Vector Machine, and the system performed well on dev set, it increased the performance about 8-14% for different dimensions. However, for the test set, the performance only improved 5% for the agreeableness dimension, but decreased for other dimensions. The obvious reason for this is the model was over trained by the development set. Therefore, it performed worse for the test set.

There are two papers mainly concentrated on prosody features. The first approach [78] added pitch and intonation features to challenge baseline feature set. First, they extracted more pitch features by using OpenSmile [86] combined with SHS algorithm and Viterbi smoothing besides baseline features, total of 365 more pitch features were extracted (97 LLD from F0 and delta F0 features, and 268 LLD from pitch except F0 and delta F0). Then they used MOMEL (MOdelling Melody) and INTSINT (INternational Transcription System for INTonation) software[88] to analyze and generate intonation patterns from pitch features, there were 34 intonation features extracted in this step. Then they used support vector machines to train the features, and they found that pitch is useful for this task, but not the intonation (contour stylization of pitch curve). In the second approach [79], they extracted prosodic features including several statistics of pitch and energy, such as mean, maximum, minimum, etc. Then they also added the prosodic polynomial coefficient features and Mel Frequency Cepstral Coefficients. After applying feature reduction to the challenge baseline feature set, they used SVM to train the model by using the combined feature sets of reduced features, prosodic and MFCC. Finally, the system improved 9% for openness, 5% for extraversion, 3% for neuroticism, and there were no improvements on conscientiousness and agreeableness.

There are several other approaches in this challenge focusing on another aspect of speech such as classification algorithm and text analysis through ASR system. The one [81] from USC SAIL is the most interesting approach for this task among them. They used an automatic speech recognition system to get the lexical content of the corpus, and applied LIWC to obtain the lexical features at first. Then they also added some prosodic features such as rate, duration of words, characters, phonemes, fillers, silence, filled pause, breathiness and laughter. Finally they used Bayesian Networks to train the model, and the system improved the baseline about 3% on average by using the dev set, and they did not provide the result of the test set.

Since we reviewed all the approaches for the Interspeech 2012 Speaker Trait Challenge, there are a variety of features and classification method were introduced from this challenge. The most commonly used method is feature reduction, since the challenge baseline used 6125 features. The feature reduction method worked well for some of them, but not always; the main issue here is over fitting, some of the approaches got a very impressive results on dev set, but when it came to the test set, the performance was disappointing (Figure 4.1). For the feature part, we found that prosodic features are performing well on personality recognition as other effective computing, such as emotion recognition, and surprisingly the lexical feature worked well on personality recognition also.

Figure 4.1: *Experimental Results on Personality recognition.*



Chapter 5

Conclusion and Future Work

This paper is aimed mainly at deception detection, and we also introduced automatic emotion recognition and personality recognition to aid in the process. We presented a series of analyses and experiments for deception detection, emotion recognition, and personality recognition.

Current approaches for detecting deception build upon the experiment of computer scientist based on the psychological theory of deception, and there are some findings and successful detections of deception. For the feature extraction, a variety of techniques were introduced for extracting lexical features, acoustic and prosodic features. For classification, various kinds of classifiers were used for detecting deception, although, Support Vector Machine is most widely used and successful classification algorithm for deception detection task, so we can consider SVM as a standard and basic classifier for this task. There have been some successful approaches in recent years for deception detection, such as emotional words type, filled pause features, and acoustic and prosodic features. However, by comparing the performance of different approaches, we can find that even the best performing approach is still not quite accurate on detecting deception of speech.(Figure 2.1) Its not only because detecting detection is difficult, but also we think there are quite number of cues that nobody has tried or are less adopted, such as emotion, personality and cultural difference. People know that emotion plays an important role in deception detection, but there is still not a clear answer for what kind emotion is the most important cue to detect deception and how does emotion correlate with deception. For the personality, [29] found that judges with different personalities performed differently when they detect deceit, but there is no further research for how personality appears in different lying style. Therefore, we want to research these topics for our future direction of work on detecting deception.

Emotion recognition is important for deception detection because people may experience stronger and variable emotions when they lie. Therefore, we think emotion recognition can find the emotion variation for different time periods to detect deception. The state-of-art emotion recognition does this job pretty well, and there are a lot of different approaches we can use for this task. Until now, there is general agreement on this task for features, that prosodic and acoustic features are the most important feature types for emotion recognition task, and the lexical features also helps this task. So we can try to start using these features as a start feature set for emotion recognition for deception detection.

Personality recognition is another important point for deception detection because of interpersonal differences and interpersonal variations during detecting deception. Psychologists [33] state that different people lie in different ways, some people raised their pitch when lying, while some lowered it significantly; some would laugh when deceiving, while others laughed more while telling the truth. [29] also found that judges with different personalities perform differently when they detect deceit, and we can hypothesize that personality test may also provide useful information in predicting individual differences in deceptive behavior of the speakers they judged. There are a lot of state of the art personality recognition systems in recent years, and many of these are performing quite well based on speech. People generally agree that it is helpful to use feature reduction, acoustic and prosodic features, lexical features for personality recognition among these approaches, and the Interspeech 2012 Speaker Trait Challenge base feature set were used in many approaches as well as SVM. We aimed to develop our own system for personality recognition for deception detection.

Although in this paper, we focused on verbal cues for detecting deception with emotion and personality, others [5, 89] have noticed major non-verbal differences in deceptive behavior between people of different cultures. We want to examine the differences in verbal cues of deception across cultures, and find the best extractable features and appropriate machine learning method for identifying the deception across cultures.

Dealing with humans is one of the most difficult and important challenges for computing, and automatic deception detection using computers is one of the more challenging tasks among human computer interactions. Deception is a common human behavior, and deception detection has been a huge interest during human history. If we can solve the deception detection task or even improve tasks performance, then it will help in many fields, such as business, jurisprudence, law enforcement, and national security.

Bibliography

[1] J. S. Seiter, J. Bruschke, and C. Bai, "The acceptability of deception as a function of perceivers' culture, deceiver's intention, and deceiver-deceived relationship," *Western Journal of Communication (includes Communication Reports)*, vol. 66, no. 2, pp. 158–180, 2002.

[2] P. A. Granhag and L. A. Strömwall, *The detection of deception in forensic contexts.* Cambridge University Press, 2004.

[3] R. L. Bassett, D. Basinger, and P. Livermore, "Lying in the laboratory: Deception in human research from psychological, philosophical, and theological perspectives," *Psychology & Christianity Integration: Seminal Works That Shaped the Movement*, vol. 34, p. 333, 2007.

[4] L. K. Guerrero, P. A. Andersen, and W. A. Afifi, *Close encounters: Communication in relationships.* Sage Publications, 2013.

[5] R. S. Feldman, L. Jenkins, and O. Popoola, "Detection of deception in adults and children via facial expressions," *Child development*, pp. 350–355, 1979.

[6] P. Ekman and W. V. Friesen, "Detecting deception from the body or face.," *Journal of Personality and Social Psychology*, vol. 29, no. 3, p. 288, 1974.

[7] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[8] N. Sebanz and M. Shiffrar, "Detecting deception in a bluffing body: The role of expertise," *Psychonomic bulletin & review*, vol. 16, no. 1, pp. 170–175, 2009.

[9] W. Kang, D. Cao, and N. Liu, "Deception with side information in biometric authentication systems," 2014.

[10] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, "Lying in everyday life.," *Journal of personality and social psychology*, vol. 70, no. 5, p. 979, 1996.

[11] A. Vrij, P. A. Granhag, and S. Porter, "Pitfalls and opportunities in nonverbal and verbal lie detection," *Psychological Science in the Public Interest*, vol. 11, no. 3, pp. 89–121, 2010.

[12] P. V. Trovillo, "A history of lie detection," *Journal of Criminal Law and Criminology (1931-1951)*, pp. 848–881, 1939.

[13] A. Vrij, "Nonverbal dominance versus verbal accuracy in lie detection a plea to change police practice," *Criminal Justice and Behavior*, vol. 35, no. 10, pp. 1323–1336, 2008.

[14] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception.," *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[15] A. Vrij, *Detecting lies and deceit: Pitfalls and opportunities.* John Wiley & Sons, 2008.

[16] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[17] J. K. Burgoon, J. P. Blair, and R. E. Strom, "Cognitive biases and nonverbal cue availability in detecting deception," *Human Communication Research*, vol. 34, no. 4, pp. 572–599, 2008.

[18] T. Lindholm, "Who can judge the accuracy of eyewitness statements? a comparison of professionals and lay-persons," *Applied Cognitive Psychology*, vol. 22, no. 9, pp. 1301–1314, 2008.

[19] T. Qin, J. Burgoon, and J. F. Nunamaker Jr, "An exploratory study on promising cues in deception detection and application of decision tree," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pp. 23–32, IEEE, 2004.

[20] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker Jr, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *Intelligent Systems, IEEE*, vol. 20, no. 5, pp. 36–43, 2005.

[21] A. Freire, M. Eskritt, and K. Lee, "Are eyes windows to a deceiver's soul? children's use of another's eye gaze cues in a deceptive situation.," *Developmental psychology*, vol. 40, no. 6, p. 1093, 2004.

[22] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 309–319, Association for Computational Linguistics, 2011.

[23] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 171–175, Association for Computational Linguistics, 2012.

[24] S. Feng, L. Xing, A. Gogar, and Y. Choi, “Distributional footprints of deceptive product reviews.,” in *ICWSM*, 2012.

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

[26] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, *et al.*, “Distinguishing deceptive from non-deceptive speech,” 2005.

[27] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, “Combining prosodic lexical and cepstral systems for deceptive speech detection,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[28] S. Benus, F. Enos, J. Hirschberg, and E. Shriberg, “Pauses in deceptive speech,” in *Speech Prosody*, vol. 18, pp. 2–5, 2006.

[29] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, “Personality factors in human deception detection: comparing human to machine performance.,” in *INTERSPEECH*, 2006.

[30] F. Enos, E. Shriberg, M. Graciarena, J. B. Hirschberg, and A. Stolcke, “Detecting deception using critical segments,” 2007.

[31] W. T. Norman, “Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings.,” *The Journal of Abnormal and Social Psychology*, vol. 66, no. 6, p. 574, 1963.

[32] R. R. McCrae and P. T. Costa Jr, “A five-factor theory of personality,” *Handbook of personality: Theory and research*, vol. 2, pp. 139–153, 1999.

[33] B. M. DePaulo and H. S. Friedman, “Nonverbal communication.,” 1998.

[34] A. Vrij, R. Fisher, S. Mann, and S. Leal, “A cognitive load approach to lie detection,” *Journal of Investigative Psychology and Offender Profiling*, vol. 5, no. 1-2, pp. 39–43, 2008.

[35] A. Vrij, S. A. Mann, R. P. Fisher, S. Leal, R. Milne, and R. Bull, “Increasing cognitive load to facilitate lie detection: the benefit of recalling an event in reverse order.,” *Law and human behavior*, vol. 32, no. 3, p. 253, 2008.

[36] F. A. Kozel, K. A. Johnson, Q. Mu, E. L. Grenesko, S. J. Laken, and M. S. George, “Detecting deception using functional magnetic resonance imaging,” *Biological psychiatry*, vol. 58, no. 8, pp. 605–613, 2005.

[37] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, “Pitch changes during attempted deception.,” *Journal of Personality and social Psychology*, vol. 35, no. 5, p. 345, 1977.

[38] K. L. Landry and J. C. Brigham, “The effect of training in criteria-based content analysis on the ability to detect deception in adults.,” *Law and Human Behavior*, vol. 16, no. 6, p. 663, 1992.

[39] J. Rosenfeld, “Alternative views of bashore and rapp’s (1993) alternatives to traditional polygraphy: A critique.,” 1995.

[40] N. R. C. C. on National Statistics *et al.*, *The polygraph and lie detection.* National Academies Press, 2003.

[41] M. G. Aamodt and H. Custer, “Who can best catch a liar? a meta-analysis of individual differences in detecting deception,” *Forensic Examiner*, vol. 15, no. 1, pp. 6–11, 2006.

[42] D. C. Raskin, “Polygraph techniques for the detection of deception.,” 1989.

[43] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker Jr, “A comparison of classification methods for predicting deception in computer-mediated communication,” *Journal of Management Information Systems*, vol. 20, no. 4, pp. 139–166, 2004.

[44] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count (liwc): A computerized text analysis program,” *Mahwah (NJ)*, vol. 7, 2001.

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

[46] J. Berger, “Arousal increases social transmission of information,” *Psychological science*, vol. 22, no. 7, pp. 891–893, 2011.

[47] M. K. Johnson and C. L. Raye, “Reality monitoring.,” *Psychological review*, vol. 88, no. 1, p. 67, 1981.

[48] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, “Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender,” *Language and speech*, vol. 44, no. 2, pp. 123–147, 2001.

[49] H. H. Clark, “Managing problems in speaking,” *Speech communication*, vol. 15, no. 3, pp. 243–250, 1994.

[50] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[51] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1970–1973, IEEE, 1996.

[52] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[53] D. J. Litman and K. Forbes-Riley, “Predicting student emotions in computer-human tutoring dialogues,” in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 351, Association for Computational Linguistics, 2004.

[54] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.,” in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[55] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–941, IEEE, 2007.

[56] N. H. Frijda, *The emotions.* Cambridge University Press, 1986.

[57] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, “How to find trouble in communication,” *Speech communication*, vol. 40, no. 1, pp. 117–143, 2003.

[58] R. Fernandez and R. W. Picard, “Classical and novel discriminant features for affect recognition from speech.,” in *Interspeech*, pp. 473–476, 2005.

[59] O. Pierre-Yves, “The production and recognition of emotions in speech: features and algorithms,” *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 157–183, 2003.

[60] B. Schuller, R. Müller, M. K. Lang, and G. Rigoll, “Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles.,” in *INTERSPEECH*, pp. 805–808, 2005.

[61] G. Nicholas, M. Rotaru, and D. J. Litman, “Exploiting word-level features for emotion prediction,” in *Spoken Language Technology Workshop, 2006. IEEE*, pp. 110–113, IEEE, 2006.

[62] A. Batliner, V. Zeißler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, “We are not amused-but how do you know? user states in a multi-modal dialogue system.,” in *INTERSPEECH*, Citeseer, 2003.

[63] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[64] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, “Speech-based recognition of self-reported and observed emotion in a dimensional space,” *Speech communication*, vol. 54, no. 9, pp. 1049–1063, 2012.

[65] T. S. Polzin and A. Waibel, “Detecting emotions in speech,” in *Proceedings of the CMC*, vol. 16, Citeseer, 1998.

[66] T. S. Polzin and A. Waibel, “Pronunciation variations in emotional speech,” in *Modeling Pronunciation Variation for Automatic Speech Recognition*, 1998.

[67] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality traits*. Cambridge University Press, 2003.

[68] A. Vinciarelli and G. Mohammadi, “A survey of personality computing,” 2014.

[69] A. E. Kazdin, “Encyclopedia 0p psychology,” 2000.

[70] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, *et al.*, “The interspeech 2012 speaker trait challenge.,” in *INTERSPEECH*, 2012.

[71] B. Rammstedt and O. P. John, “Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german,” *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[72] P. T. Costa and R. R. MacCrae, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. Psychological Assessment Resources, 1992.

[73] F. Mairesse and M. Walker, “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, 2006.

[74] F. Mairesse and M. Walker, “Words mark the nerds: Computational models of personality recognition through language,” in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 543–548, 2006.

[75] J. Pohjalainen, S. Kadioglu, and O. Räsänen, "Feature selection for speaker traits.," in *INTERSPEECH*, 2012.

[76] C. Chastagnol and L. Devillers, "Personality traits detection using a parallelized modified sffs algorithm," *computing*, vol. 15, p. 16, 2012.

[77] A. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recognition," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[78] C. Montacié and M.-J. Caraty, "Pitch and intonation contribution to speakers' traits classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[79] M. H. Sanchez, A. Lawson, D. Vergyri, and H. Bratt, "Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification.," in *INTERSPEECH*, 2012.

[80] Y. Attabi and P. Dumouchel, "Anchor models and wccn normalization for speaker trait classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[81] K. Audhkhasi, A. Metallinou, M. Li, and S. Narayanan, "Speaker personality classification using systems based on acoustic-lexical cues and an optimal tree-structured bayesian network.," in *INTERSPEECH*, 2012.

[82] H. Buisman and E. Postma, "The log-gabor method: speech classification using spectrogram image analysis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[83] J. Wagner, F. Lingenfelser, and E. André, "A frame pruning approach for paralinguistic recognition tasks.," in *INTERSPEECH*, 2012.

[84] M. Coltheart, "The mrc psycholinguistic database," *The Quarterly Journal of Experimental Psychology*, vol. 33, no. 4, pp. 497–505, 1981.

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

[86] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[87] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall.," in *INTERSPEECH*, 2012.

[88] D. Hirst, “A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation,” in *Proceedings of the XVIth International Conference of Phonetic Sciences*, vol. 12331236, 2007.

[89] M. J. Cody, W.-S. Lee, and E. Y. Chao, “Telling lies: Correlates of deception among chinese,” *Recent advances in social psychology: An international perspective*, pp. 359–368, 1989.