# Computational Approaches to Modeling Speaker State in the Medical Domain

Julia Hirschberg - Computer Science Department, Columbia University
Anna Hjalmarsson - Speech, Music and Hearing, KTH
Noémie Elhadad - Department of Biomedical Informatics, Columbia University

**Abstract**
Recently, researchers in computer science and engineering have begun to explore the possibility of finding speech-based correlates of various medical conditions using automatic, computational methods. If such language cues can be identified and quantified automatically, this information can be used to support diagnosis and treatment of medical conditions in clinical settings and to further fundamental research in understanding cognition. This chapter reviews computational approaches that explore communicative patterns of patients who suffer from medical conditions such as depression, autism spectrum disorders, schizophrenia, and cancer. There are two main approaches discussed: research that explores features extracted from the acoustic signal and research that focuses on lexical and semantic features. We also present some applied research that uses computational methods to develop assistive technologies. In the final sections we discuss issues related to and the future of this emerging field of research.

# 1 Introduction

Many medical conditions (e.g., depression, autism spectrum disorders, schizophrenia, as well as cancer) affect the communication patterns of the individuals who suffer from them. Researchers in psychology and psycho-linguistics have a long tradition of studying the speech and language of patients who suffer from these conditions to identify cues, with the hope of leveraging these cues for both diagnosis and treatment. Like with other observational data based on patient behavior, clinicians follow rigorous training to elicit and analyze patients' speech. More recently, researchers in computer science and engineering have begun to explore the possibility of finding speech-based correlates of various medical conditions using automatic, computational methods. If cues to medical disorders can be quantified and detected automatically with some degree of success, then this information can be used in clinical situations. Thus, automatic methods can assist clinicians in not only in screening patients for conditions, but also in assessing the progress of ongoing treatment. Furthermore, automatic methods can provide cost- and time-effective general screening methods for disorders, such as autism spectrum disorders, which often go undiagnosed. Finally, they can also provide useful input for assistive technologies that can be used in clinical situations or made available to patients in the home.

In this chapter, we discuss some of these approaches and suggest possibilities for future computational research on language cues to medical conditions and on assistive technologies being developed to make use of them.

## 2 Computational Approaches to Speaker State

Computational approaches to the study of language correlates of medical conditions have largely arisen from related work on computational modeling of emotional state. Numerous experiments have been conducted on the automatic classification of the classic emotions, such as anger, happiness, sadness; secondary emotions such as confidence or annoyance; or simply positive from negative emotions, from acoustic, prosodic, and lexical information (c.f. Mozziconacci & Hermes 1999, Yuan et al 2002, Ang et al 2002, Batliner et al 2006, Lee & Narayanan 2004, Liscombe et al 2005, Ai et al 2006,Liscombe et al 2006). Motivation for these studies have come primarily from call center and Interactive Voice Response (IVR) applications, for which there is interest in distinguishing angry and frustrated callers from the rest, either to hand them off to a human attendant or to flag such conversations as problematic for off-line study (Lee & Narayanan 2004, Liscombe et al 2005a, Devillers & Vidrascu 2006, Gupta & Rajput 2007). Other research has focused on assessment of students' emotional state in automatic tutoring systems (Liscombe et al 2005b, Ai et al 2006).

More recently, emotional speech researchers have expanded the range of phenomena of interest beyond studies of the classic emotions to include emotion-related states, such as deception (Hirschberg et al 2005), sarcasm (Tepperman et al. 2006), charisma (Biadsy et al 2008), personality (Mairesse &Walker 2006), romantic interest (Ranganath et al. 2009), "hotspots" in meetings (Wrede & Shriberg 2003), and confusion (Kumar et al 2006). To encompass this expansion of a research space which typically uses similar methods and a common set of features for classification, some have termed research of this larger class the study of *speaker state***.** A recent focus of this area has been the use of techniques and features developed in studies of emotional speech in the analysis of medical and psychiatric conditions.

Most computational studies of emotion and other speaker state make use of statistical machine learning techniques such as Hidden Markov Models (HMMs), logistic regression, rule-induction algorithms such as C4.5 or Ripper, or Support Vector Machines to distinguish among possible states. Corpus-based approaches typically examine acoustic and prosodic features, including pitch, intensity, and timing information (e.g., pause and turn durations and speaking rate), and voice quality, and less often lexical and syntactic information, extracted from large amounts of hand-labeled training data. Many corpus-based studies suffer from poor agreement among labelers, making the training data noisy. Since human annotation is expensive and labelers often disagree, unsupervised clustering methods are sometimes used to sort data into states automatically, but it is not always clear what the resulting clusters represent. Laboratory studies attempt to induce the desired states from professional actors or non-professional subjects in order to compare the same linguistic features in production studies or to elicit subject judgments of acted or natural emotions in perception studies. However, characteristics of emotions elicited from actors have been found to differ significantly from those evinced by ordinary subjects, making it unclear how best to design representative laboratory studies. More recently, Magnetic Resonance Imaging (MRI) studies have sought to localize various emotions and states within the brain (e.g., Johnstone et al 2006, Lee et al 2006). While these experiments sometimes produce intriguing results, it is still not clear what we can conclude from them, beyond the activation evidence in different locations of the brain associated with different speaker states. Thus, it is not always clear how best to study speaker state. Medical conditions, however, provide the

possibility of correlating medical diagnoses with the same sorts of language-based features used to examine states like anger, confidence, and charisma.

# 3 Computational Approaches to Language Analysis in Medical Conditions

Computational approaches to the study of language in medical conditions can be classified in several ways – by the condition studied, the methods used, or the end goal of the study. Research has been done on prosodic cues for the assessment of coping strategies in breast cancer survivors (Zei Pollerman 2002), for evaluations of head and neck cancer patients (Maier et al 2010), for diagnosis of depression and schizophrenia (Alpert et al 2001, Moore et al 2003, Mundt et al 2007, Bitouki et al 2009), and for the classification of Autism Spectrum Disorders (ASD)(Le Normand et al 2008, Paul et al 2008, Hoque 2008, Diehl et al 2009, Van Santen et al 2009).Textual analysis methods, which rely on patient speech transcripts or texts authored by patients, have also been leveraged for cancer coping mechanisms (Graves et al 2005, Bantum & Owen 2009), psychiatric diagnosis (Oxman et al. 1988, Elvevag et al. 2009), and analysis of suicide notes (Pestian 2008). While much research has focused primarily on assessment, other researchers have explored possibilities for treatment, providing assistive technologies for those suffered from aphasia (Fink et al 2009) or ASD (Kaliouby et al 2006). Other work has examined the success of assistive technologies, such as the evaluation of cochlear implants (Luo et al 2006). In this section we will discuss some of this research to illustrate the approaches that have been taken and the current state of the field.

## *3.1 Assessment and Diagnosis*

### 3.1.1 Cancer

Computational methods have been leveraged for diagnosing cancer based on a patient's speech transcript and quantifying the effect of cancer on speech intelligibility. One open research question in the field of psycho-social support for cancer patients and survivors is how to identify coping mechanisms. There has also been work on studying the presence and extent of emotion expressions in cancer patients' speech.

Oxman et al. (1988) describe an early attempt at using automatic textual analysis to diagnose four possible conditions. The authors collected speech samples from 71 patients (25 with paranoid disorder, 17 with lung or breast cancer, 17 with somatization disorder, and 12 with major depression). Patients were asked to speak for five minutes about a subject of their choice. The speech transcripts were analyzed through two different textual analysis methods: (1) automatic word match against a dictionary of psychological dimensions (Stone et al. 1969) and (2) manual rating according to hostility and anxiety scales derived from the Gottschalk-Gleser scales (Gottschalk et al. 1969). The Gottschalk-Gleser scales are an established textual analysis method, traditionally used to support psychiatric diagnosis. The scales operate at the clause level, thereby taking into account a larger context than dictionaries. Two psychiatrists were also asked to read the transcripts and diagnose the patients with one of the four conditions. Neither the raters nor the two psychiatrists had knowledge of the patient**'s** condition. The authors found that

the pure lexical lookup method identified the best predictors for diagnosis classification, above the manual analysis and the expert diagnoses.

Automatic Speech Recognition (ASR) has been shown to be an effective means of evaluating intelligibility for patients suffering from cancer of the head and neck. In 2010, Maier et al. experimented with this method, recording German patients suffering from cancer of the larynx and others suffering from oral cancer. ASR was performed on the recordings; the word recognition rate (i.e. ratio of correctly recognized words to all words spoken by the speaker) was then compared to perceptual ratings by a panel of experts and to an age-matched control group. Both patient groups showed significantly lower word recognition rates than the control group. Automatic speech recognition yielded word recognition rates which correlated with experts' evaluation of intelligibility on a significant level. They thus concluded that word recognition rate from ASR can serve as a good means with low effort to objectify and quantify this important aspect of pathologic speech.

Zei Pollerman (2002) found a positive correlation between patients considered to be coping well with their treatment and mean pitch range. It was hypothesized that *active coping,* defined as "tonic readiness to act upon an event," could be reflected in the prosody of spontaneous speech. Ten breast cancer patients were diagnosed by clinicians as to their coping behavior, active or passive. Patients' voice recordings were recorded in high and low arousal conditions, and analyzed for mean f0, f0 range (defined as f0 maximum – f0 minimum), standard deviation of f0, mean intensity, intensity ratio expressed as decibel (dB) maximum vs. dB minimum, and speaking rate. For each parameter the difference between values for high and low arousal conditions were measured. The study found that those with adaptive adjustment to their cancer (active coping) showed a higher difference in f0 range than those with passive coping behavior.

More recently, Graves et al (2005) discovered some differences between cancer survivors and controls with respect to emotional expression in textual samples in a study of emotional expression in breast cancer patients. Comparing 25 breast cancer patients with 25 healthy patients, this study asked subjects to complete a verbal 'emotion expression' behavioral task. James Pennebaker and his research collaborators 'Linguistic Inquiry and Word Count (LIWC) paradigm was used to identify positive and negative emotion words in text(Pennebaker, Francis, & Booth, 2001).The authors found that, while there was no difference between cancer sufferers and healthy subjects, cancer patients used significantly fewer "inhibition words" and were in fact rated by trained raters as expressing *more* intense emotion.

The use of lexical resources to recognize expressions of emotion in text was also investigated in the work of Bantum and Owen (2009). They compare two automatic resources, LIWC and the Psychiatric Content Analysis and Diagnosis system (PCAD), based on the Gottschalk-Gleser scales mentioned above, for the recognition of positive and negative emotions, as well as more particular emotions of anxiety, anger, sadness and optimism. The authors compiled a corpus of texts written by 63 women with breast cancer in an Internet discussion board. On average, each text contained 2600 words. Trained raters annotated the texts according to positive and negative emotions, as well as presence of anxiety, anger, sadness, and optimism. Along with the texts, self-reports of emotional well-being were collected from the 63 participants. The authors found that LIWC was more accurate in identifying emotions than PCAD when compared to the manual

raters, despite the context-sensitive nature of PCAD. Interestingly, when comparing the self reports to manual and automatic ratings, there was no significant correlation between the self-reported positive and negative emotions and the rater, LIWC or PCAD codes of positive and negative emotions.

### 3.1.2 Diabetes

Zei Pollerman (2002) presents an early study of the potential use of acoustic-prosodic features in diagnosis of various conditions. In a study of diabetic patients at the University Hospitals of Geneva, the relationship between autonomic lesions and diminished emotional reaction was examined. Forty diabetic patients' autonomic functions were assessed by quantification of their heart rate variability (HRV). Emotional states (anger, joy and sadness) were then induced via verbal recall of personal experience. Subjects were then asked to pronounce a short sentence in a manner appropriate to the emotion induced and to report the degree to which they had felt the emotion on a scale from 1 to 4. Their utterances were analyzed for f0, energy, and speaking rate and these features were then correlated with their HRV indices. The f0 ratio, that is, the difference between F0 maximum and F0 minimum, energy range, and speaking rate were significantly correlated with HRV. A combined measure based on these features was then used to compare between subjects' productions of angry utterances and sad utterances. The study found that indeed subjects with a higher degree of autonomic responsivity displayed a higher degree of differentiation between anger and sadness in their vocal productions. This suggested that poor prosodic differentiation between anger and sadness could be interpreted as a symptom of poor autonomic responsivity. The study also found that groups with higher HRV reported a higher degree of subjective feeling for the induced emotions than those with lower HRV.

### 3.1.3 Depression

Researchers studying the acoustic correlates of depression have generally distinguished between studies of *automatic speech*, such as counting or reading, from studies of *free speech*, since the latter requires cognitive activity such as word finding and discourse planning in addition to simple motor activity. Research on automatic speech includes research at Georgia Tech and the Medical College of Georgia (Moore et al 2003) on the use of features extracted from the glottal waveform to separate patients suffering from clinical depression from a control group. These researchers analyzed speech from a database of 15 male (6 patients, 9 controls) and 18 female (9 patients, 9 controls) recording reading a short story, with at least 3 minutes of speech for each subject; glottal features were used to classify within each gender group. While the data set was quite small, the researchers reported promising results for some of the features.

Other researchers have compared subjects diagnosed as *agitated* vs. those diagnosed as *retarded* depressives. Alpert et al (2001) examined acoustic indicators in the speech of patients diagnosed with depression to assess results of different treatments. In a 12-week double-blind treatment trial, that compared response to nortriptyline (25–100 mg/day) with sertraline (50–150 mg/day).Twelve male and ten female elderly depressed patients and an age-matched normal control group ($n$=19) were studied. Patients were divided into retarded or agitated groups on the basis of prior ratings. Measures of fluency (speech productivity and pausing) and prosody (emphasis and inflection) were examined. Depressed patients showed "less prosody" (emphasis and inflection) than the normal subjects. Improvement in the retarded group was reflected in

briefer pauses but not longer utterances. There was a trend in the agitated group for improvement to be reflected in the utterance length but not the length of pauses. The authors concluded that clinical impressions were substantially related to acoustic parameters. These findings suggest that acoustic measures of the patient's speech may provide objective procedures to aid in the evaluation of depression.

Mundt et al. (2007) presented a study in which they elicited, recorded and analyzed speech samples from depressed patients in order to identify voice acoustic patterns associated with depression severity and treatment response. An IVR telephone system was developed by Healthcare Technology Systems (Madison, WI) and used to collect speech samples from 35 patients. All subjects got a personal pass and an access code to the IVR system and were then asked to repeatedly call a toll-free telephone number over a period of 6 weeks. The IVR system requested the subjects to respond to different types of questions such as "describe how you've been feeling physically during the past week" and additional tasks including to count from 1 to 20 or to recite the alphabet. The subjects' depression severity was evaluated using three different clinical measures including clinician rated HAMDs, IVR HAMDs and IVR QIDS. During the 6 weeks of data collection, 13 subjects showed a treatment response whereas 19 did not. Comparisons in vocal acoustic measurements between the group that showed treatment response and the group that did not were performed. The results show that there were no significant differences between subjects at baseline, that is, at the beginning of the six weeks. In a comparison of the acoustic measures between the baseline and the end of the 6 weeks, the group with no treatment response only showed a reduction in total pause time. In contrast, responders to treatment, showed a number of significant differences, including increased pitch variability (f0), increased total recording sample duration, a reduction in total pause time, fewer number of pauses and an increase in speaking rate.

### 3.1.4 Schizophrenia

Schizophrenia is a neuro-developmental disorder with a genetic component. Patients typically show disorganized thinking and their language is correspondingly affected, especially at the discourse level. Research has found that healthy, non-schizophrenic relatives also exhibit some subtle peculiar communication patterns at the lexical and discourse level. Elvevag et al. (2007) show that Latent Semantic Analysis (LSA) is a promising method to evaluate patients based on their free-form verbalizations. In a follow-up study, Elvevag and colleagues (2009) collected 83 speech transcripts from three groups: schizophrenic patients, their first-degree relatives, and healthy unrelated individuals. They analyzed the transcripts according to three types of measures: statistical language models, measures based on the semantic, LSA-based similarity of a text sample to patient or control text samples, and surface features such as sentence length. They found that the three populations could be discriminated based on these three types of measures. When discriminating between patients and non-patients, surface features and language model features were predictive enough on their own (patients tended to have shorter sentences, with unusual choice of words). However, when discriminating between patients and their healthy relatives, a successful model required features related to syntactic and semantic level in addition to surface features.

Bitouki et al (2009) have begun to examine the use of automatic emotion recognition approaches in speech to the diagnosis of schizophrenia. In their initial work, they have focused on

identifying new features for emotion recognition. They have experimented with the use of segmental spectral features to capture information about expressivity and emotion by providing a more detailed description of the speech signal. They describe results of using Mel-Frequency Spectral Coefficients computed over three phoneme type classes: stressed vowels, unstressed vowels and consonants in the utterance to identify emotions in several available speech corpora, the LDC (Linguistic Data Consortium) Emotional Speech Corpus and the Berlin Emo-DB (Emotional Speech Corpus). Their experimental results indicate that both the richer set of spectral features and the differentiation between phoneme type classes improved performance on these corpora over more traditional acoustic and prosodic features. Classification accuracies were consistently higher for the new features compared to prosodic features or utterance-level spectral features. Combination of the phoneme class features with prosodic features leads to even further improvement. These features have yet, however, to be applied to the diagnosis of schizophrenia.

### 3.1.5 Autism Spectrum Disorders

Autism Spectrum Disorder (ASD) is a range of neurodevelopment disorders that affect communication, social interaction and behavior. Symptoms range from mild to severe and include a lack of interest in social interaction, trouble communicating and repetitive and restrictive behavior. ASD has several diagnostic categories including autism, Aspberger syndrome and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). Kanner (c.f. 1946, 1948) was one of the first to describe the behavioral disorders in the autistic spectrum and many of them are related to speech and language. Frequently mentioned language impairments include: unusual word choices, pronoun reversal, echolalia, incoherent discourse, unresponsiveness to questions, aberrant prosody, and lack of drive to communicate (Rapin & Dunn, 2003).

ASD has no simple confirmatory test, but is diagnosed by a set of physical and psychological assessments. Influenced by early observations made by Kanner (1946) and Asperger (1944), much work within psycholinguistics has been devoted to identifying and studying the language disorders related to ASD. The majority of this research has been qualitative rather than quantitative, but recently researches have started to use computational methods to study some of these disorders.

ASD is associated with having an odd or peculiar sounding prosody (Mesibov, 1992). Frequently mentioned deficits include both observations of a “flat” or monotonic voice as well as an abnormally large variation in f0. Deviations in prosody are difficult to isolate since prosody interacts with several levels of language such as phonetics, phonology, syntax and pragmatics. Moreover, f0 varies a great deal between speakers, within speakers and over different contexts. An early study of f0 and autism suggests that compared controls, a group of individuals with autism appeared to have either a wider or a narrower range of f0 (Baltaxe, 1984). Shriberg et al., (2001) used the Prosody-Voice Screening Profile (PVSP), a standardized screening method, to study prosodic deficits in individuals with High-Functioning Autism (HFA) and Asperger Syndrome (AS). The results show that utterances spoken by the group of individuals with HFA and AS were often marked as inappropriate in terms of phrasing, stress and resonance.

Diehl et al. (2009) use Praat (Boersma & Weenink, 2005) to extract f0 in order to explore if there are any differences in f0 range between individuals with HFA and typically developing children.

The results show that the HFA individuals had a higher average standard deviation in f0 than controls, however, the groups did not differ in average f0. The clinical manual judgments (ADOS) of the individuals in the HFA group also turned out to be significantly correlated with the average standard deviation across f0 samples. That is, subjects with a higher variation in f0 were judged as having greater language impairment by trained clinicians. However, there is considerable overlap between the two groups, which suggests that f0 alone cannot be used to identify deviations in expressive prosody for ASD individuals.

There are also studies that have explored deviations in specific functions of prosody. Le Normand et al. (2008) study prominence and prosodic contours in different types of speech acts. The speech samples were spontaneous speech taken from eight French speaking autistic children in a free play situation. The hypothesis was that children with a communicative disorder, such as autism will fail to produce appropriate prominence and prosodic contours related to different communicative intent such as declarative, exclamation or question. The speech acts and prominence were labeled manually. The prosodic contour was extracted and judged manually by visualizing the sound in the Praat editor (Boersma & Weenink, 2005). The results suggest that the there is a large proportion of the utterances with low prominence and flat prosodic contour.

Van Santen et al. (2009) present another study that investigates how ASD individuals produce specific functions of prosody. The prosodic functions explored include lexical stress, focus, phrasing, pragmatic style and affect. The tasks that were used to elicit data were specially designed to make the subjects produce speech with the targeted prosodic functions. The subjects were recorded and scored in real time by clinicians and later also judged by a set of naïve subjects. Automatically extracted measures of f0 and amplitude were collected as well. The results show that a combined set of the automatic measures correlated approximately as high with the naïve subjects' mean scores as the clinicians' individual judgments. However, the real time judgments of clinicians correlated substantially less with the mean scores than the automatic measures.

Paul et al. (2008) investigate stress production by ASD individuals in a nonsense syllable imitation task. The aim was to establish whether ASD speakers produce stressed syllables differently from typically developing (TD) peers. The hypothesis was that the ASD patients would not perform differently from the control (the TD speakers). The study included speech samples from 20 TD speakers and 46 speakers with ASD. Subjective judgments and automatically extracted acoustic measures were correlated with diagnostic characteristics (e.g., PIQ, VIQ, Vineland and ADOS scores). The results show significant but small differences in the production of stressed and unstressed syllables between the ASD and TD speakers. First, the speakers with ASD were less likely to get the right subjective judgment of their produced stress than the TD speakers. Second, the analysis of the acoustic measures revealed that both TD and ASD speakers produced longer stressed than unstressed syllables but that the duration differences between stressed and unstressed syllables were smaller for the ASD group.

Hoque (2008) analyzes a number of different voice parameters in individuals with ASD, Down syndrome (DS) and Neuro-Typicals (NT). The parameters analyzed included f0, duration, pauses rhythm, formants and voice quality intensity. The parameters were explored using data mining methods in order to find a set of optimal features that can be used to identify distinguishable

speech features for the ASD, DS and NT groups. The results show that the average duration per turn was longer for NT than for ASD and DS. Moreover, the magnitudes of maximum rising and falling edges in a turn/utterance is much higher for NT than in DS and ASD. Yet, the number of rising and falling edges is comparable between NT and ASD. The future aim of this initial analysis of speech parameters is build assistive technologies that can give individuals with ASD and DS real time feedback helping them produce more intelligible speech.

### 3.1.6 Suicide

There has been preliminary work on the analysis of suicide notes. The goal of such an analysis is to gain deeper understanding of the psychological state of individuals committing suicide, as well as to help prevent suicide of such individuals. Pestian et al. (2008) envision a screening tool in place at a psychiatry emergency room that predicts the likelihood of an individual being in a suicidal state rather than merely depressed. The authors present preliminary results on the use of linguistic analysis when applied to suicide notes. Notes from 33 completers (individuals who completed suicide) and 33 simulators (individuals not contemplating suicide who were asked to write a suicide note) were collected. The authors trained a classifier for completer/simulator. Features included word count, presence of pronouns, unigrams, Kincaid readability index, and presence of emotional words based on an emotion dictionary match. The best classifier reached 79% accuracy in discriminating between completers and simulators. The most significant linguistic differences between completers and simulators were in fact at the surface level, such as word count. For comparison, five mental health professionals were asked to read the notes and classify them as originating from a completer or a simulator. Experts classified the notes with 71% accuracy.

## *3.2 Assistive Technologies*

Much research has been done on the use of speech technologies to assist persons with medical disabilities, such as using Text-to-Speech systems as aids for the blind or for those who have lost their ability to speak. In this section however we focus on the use of recognition technologies to aid those who are being treated for disabilities.

### 3.2.1 ASD

Research at the MIT Media Lab led by Rosalind Picard has proposed a number of methods to assist those diagnosed with ASD. In El Kaliouby et al (2006) a wearable device is described which is designed to monitor social-emotional information in real time human interaction. Using a wearable camera and other sensors, and making use of various perception algorithms, the system records and analyzes the facial expressions and head movements of the person with whom the wearer is interacting. The system creators propose an application of individuals diagnosed with ASD, to help them in perceiving communication in social settings and enhancing their social communication skills.

Hoque et al (2008) analyzed the acoustic parameters of individuals diagnosed with ASD and Down syndrome. The idea is to use these parameters to visualize subject's speech productions in real time in order to provide them with live feedback that can help them modify their productions. In further work (Hoque et al 2009), explores the effect of using an interactive game to help individuals with ASD produce intelligible speech. Nine subjects diagnosed with ASD and 1 subject with Down's syndrome participated in the study. Most of the participants had

difficulties with amplitude modulation and speech rate and the interactive game was designed to target these problems. The subjects alternately received sessions with a computerized game and traditional speech theory. A number of different acoustic measures were extracted automatically, including Relative Average Perturbation (RAP), Nosie Harmonic Ratio (NHR), Voice Turbulence Index (VTI) and prosodic features including pitch, intensity and speaking rate. A preliminary analysis suggests that one participant significantly slower his speech rate when interacting with the computerized game. Furthermore, two other participants' had significant reduction in pitch brakes when interacting with the computerized program, suggesting that they were able to better control their pitch.

### 3.2.2 Aphasia

Aphasia is a condition in which people lose some of their ability to use language due to an injury (often stroke) or disease that affects the language-production and perception areas of the brain. While aphasics are generally very motivated to improve their speech and language abilities and are receptive to using computer programs, they vary in their ability to use a mouse or keyboard, read, speak, and understand spoken language. Typically treatment currently includes speech therapy by trained therapists, which can be quite costly and is rarely covered by insurance. To address this problem, Fink et al (2002, 2009) have developed software to provide aphasia sufferers with structured practice targeted at improving their speech on a long-term basis. MossTalkWords 2.0 was developed by Moss Rehab Hospital, to lead users through several different types of exercises in a self-paced manner. One of these exercises, Cued Naming, involves presenting a picture of an item or action and asking the user to name it, with cues available as memory aids. In the initial version of MossTalk, users self-monitored the correctness of their responses, or worked with a clinician. The need for the clinician could be reduced by using an ASR engine (Microsoft 6.1) with the grammar dynamically modified to include only the description of the picture being presented. An enhanced version of the system integrated speech recognition with MossTalkWords so that users would get immediate and automatic feedback from the ASR on whether the picture was correctly named. Advantages of using ASR instead of a human clinician are not only lower cost but also 24/7 availability of the system. An evaluation of the on mild to moderate aphasia sufferers with good articulation found acceptable levels of accuracy for the ASR and considerable reported user satisfaction.

### 3.2.3 General Evaluation

Researchers at Universität Erlangen-Nürnberg have fielded a web-based system called PEAKS (Program for Evaluation and Analysis of all Kinds of Speech disorders) to evaluate speech and voice disorders automatically. They particularly target speech evaluation after treatment, which is typically performed subjectively by speech pathologists, who are asked to assess intelligibility. The essence of their system is an ASR system developed at Erlangen-Nürnberg for use in spoken dialogue systems. The subject reads a known text, which is recognized by the system, which weights acoustic features higher than other components for this application. System output is just word error rate and word accuracy. This information is combined with information from a prosody module, which extracts pitch, energy, and duration along with jitter, shimmer, and information on voiced/unvoiced segments. These features are used to create a classifier trained on expert judgments. The resulting classifier is then used to assign scores to test patient recordings. Using their system, they report that their evaluation of patients whose larynx has been removed due to cancer and who have received tracheoesophageal (TE) substitute voices

correlates .90 (P < .001) with human expert judgments. The correlation of PEAKS judgments with experts is .87 (P<.001) for children with cleft lip and palate who have undergone reconstructive surgery. The system can be accessed over the phone or on the web and is intended to provide a "second opinion" for pathologists working alone or in other cases where speech therapists might use additional information in further treatment.

# 4 Discussion

The use of computational methods to identify speaker state in the medical domain is an emergent field of research. This research builds on previous findings in the fields of psychiatry, psychology and cognition. Speech and textual analysis can help us gain a deeper understanding of medical conditions, and they can also contribute to the design of systems that can be used clinically, whether as an aid to diagnosis/screening or to assess the effectiveness of treatment. While the methods described in this chapter are far from being readily usable in a clinical setting, they are nonetheless very promising, since they promise to help with many conditions which are currently very difficult to diagnose and treat.

## *4.1 Exploring Different Methods for Identifying Speaker State Identification*

There are two main approaches to the analyses described above: one relies on the *speech signal* and one operates on the *speech content.* Historically, speech-processing researchers have focused on features derived from speech signal, while psychology researchers have relied on textual analysis methodology. More specifically, computational analyses of speech and language in the context of medical conditions operate at two primary levels:

**1)** Aspects of the speech signal, including durational features, intensity, and F0;and
**2)** Surface features of the textual, such as word count and lexical patterns, primarily examined through matching against lexical resources (see Pennebaker et al. (2003) for a review of textual analysis methods and applications).

One open research question is whether combining these two levels, which have generally been investigated separately, could yield more accurate models of speaker state. Furthermore, as NLP technology progresses in part-of-speech tagging, syntactic parsing, semantic inference, and discourse modeling, more and more tools are now readily available and can be used in a variety of settings. It is worth investigating whether incorporating additional linguistic features yields better computational models of speaker state. So far, there is conflicting evidence that higher-level linguistic features are more helpful than shallow, lexical ones or speech signal information. In two studies, for instance, purely lexical features (as derived from lexical resources like LIWC) performed better than context-aware features, such as PCAD and the Gottschalk-Gleser scales (Oxman et al 1988; Bantum & Owen 2009). On the other hand, Le Normand (2008) shows that semantic and syntactic information derived from manually labeled speech acts can help target specific functions of prosody, which have been described previously as difficult for individuals with ASD to process. Similarly, in the study of schizophrenia, evidence shows that distributional semantics, without the necessity of factoring in speech signals, can be leveraged to discriminate patients from their first-degree healthy relatives (Elvevag et al. 2009).

## *4.2 Comparing Different Methods of Data Collection*

One important characteristic shared by all the studies described in this chapter is their data-driven approach in characterizing speaker states. However, many open questions about data collection remain, making it a primary concern for future research in this area.

One data collection issue, which needs to be carefully considered is the methodology used to elicit data. Many of the conditions and their associated speaker states have already been described in detail by clinicians in the literature (e.g., prosody in ASD patients). For computational methods to succeed, however, they must analyze actual speech samples which are representative of the behaviors under study and which can be easily segmented to target the representative examples. In the case of ASD patients, for instance, the goal is to collect speech samples which may exhibit particular prosodic patterns. In the study of coping mechanisms for cancer patients, speech samples with emotion expressions are desired.

### 4.2.1 Spontaneous Speech versus Scripted Speech

The research discussed in this chapter presents several strategies for collecting data:
1) free spontaneous speech (e.g., Oxman et al. 1988, Le Norman et al 2008);
2) free speech, albeit about a particular topic (e.g., Bantum & Owen 2009);
3) proxy texts (e.g., Pestian et al. 2008);and
4) specific tasks such as imitation, where subjects repeat words or sentence read to them (e.g., Van Santen et al. 2009).

Let's discuss imitation for a moment. Collecting data through specific tasks or repetitions facilitates segmentation, because it is possible to control what the subjects say and when they say it. Another advantage is that such tasks allow researchers to elicit larger amounts of the target behavior. Spontaneous speech, on the other hand, is difficult to analyze in a controlled fashion. The target behavior may occur sparsely, if at all. Moreover, in order to identify critical segments, spontaneous speech data requires hand labeling or other types of pre-processing, which can be challenging for research. Against all these advantages of scripted speech, spontaneous dialogue data has the benefit of *ecological validity* – it is more representative of subjects' behavior in a natural environment. For example, in the specially designed tasks presented by Van Santen et al. (2009), the children appeared to have little problem conveying the target prosody. It is possible that ASD individuals can imitate appropriate prosody in a laboratory setting, but they may still have problems using these accurately in a real-world setting.

### 4.2.2 Annotation Difficulties and Shortcomings

In many cases, the speech samples, or their transcripts, must be annotated with gold standard information. For computational methods to be successful, one must pay attention to the annotation process. Because most annotations related to speaker state are largely observational, agreement among annotators can be low. Careful annotation is often tedious and follows extensive annotation guidelines. For instance, annotating emotions in speech transcripts is a difficult task for humans. It requires trained annotators as well as established annotation schemas. Yet, while there are several emotion taxonomies developed in the field of psychology, it is unclear whether they are readily usable for computational purposes. While it makes sense from a psychological standpoint to differentiate between "anger" and "hostile anger" for example, it might be necessary to merge the two emotions when training an emotion-detection

tool from annotated texts. Besides the annotation costs, the validity of the annotation is important. In several studies, expert opinions are not always reliable (e.g., Oxman et al. 1988, Pestian et al. 2008), but neither are patients' self-reports (Bantum & Owen 2009).

### 4.2.3 Protecting Human Subjects' Privacy

Finally, privacy concerns cannot be ignored when collecting language samples from patients. In the United States, for instance, the Health Insurance Portability and Accountability Act (HIPAA) laws and the institutional review boards ensure that the privacy of patients is kept. As such, health evaluations and also audio recordings are considered protected health information. Researchers must obtain institutional approval prior to collecting and processing speech samples. Furthermore, for datasets to be available to the scientific community, they must first be anonymized.

## 5 Conclusion

This is an exciting time for researchers in speech and language processing to investigate methods to recognize speaker state from a medical standpoint. With recent advances in speech processing, core natural language processing technologies, and data mining, the time is ripe to apply these methods to clinical applications. The resulting tools can impact medicine in several ways. Clinicians are more and more accustomed to having technology as part of their every-day activities and are more open to recognizing the value of technology in their decision-making processes. Thus, screening tools for conditions that are difficult to diagnose, partly because diagnosis of such conditions rely [for which diagnosis may be difficult because it relies] on [careful] close observation of patients over time, can be developed in tandem with the needs and skills of clinicians. Such tools can have economic and public health benefits, in that a wider population – particularly individuals who live far from major medical centers – can be efficiently screened for a broader spectrum of neurological disorders. Fundamental research on mental disorders, like post-partum depression and post traumatic stress disorder, and on coping mechanisms for patients with chronic conditions, like cancer and degenerative arthritis, can [also] likewise benefit from computational models of speaker state. A successful research endeavor, which brings together computational and clinical expertise, will ultimately provide better understanding of computational models as well as cognition.

## References

H. Ai et al (2006), "Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs," Interspeech 2006, Pittsburgh.

M. Alpert et al (2001), "Reflections of depression in acoustic measures of the patient's speech," *Journal of Affective Disorders*, 66:59–69.

J. Ang et al (2002), "Prosody-based automatic detection of annoyance and frustration in human-computer dialog", ICSLP 2002, Denver.

H. Asperger (1944) (tr. U. Frith (1991)), "Autistic psychopathy in childhood," in U. Frith. *Autism and Asperger syndrome*. Cambridge University Press. pp. 37–92.

E. Bantum and J. Owen (2009), "Evaluating the Validity of Computerized Content Analysis Programs for Identification of Emotional Expression in Cancer Narratives," Psychological Assessment, 2009, 21(1): 79–88.

Emo-DB. Berlin Emotional Speech Corpus. (http://pascal.kgw.tu-berlin.de/emodb/).

D. Bitouk et al. (2009), "Improving Emotion Recognition using Class-Level Spectral Features," Interspeech 2009, Brighton.

C. Baltaxe (1984). "Use of contrastive stress in normal, aphasic, and autistic children," *Journal of Speech and Hearing Research*, 27:97–105.

A. Batliner et al, (2003) "How to find trouble in communication," *Speech Communication*, 40, pp. 117–143.

P. Boersma.&D. Weenink (2005). PRAAT: Doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved from http://www.praat.org.

F. Burkhardt et al. (2005), "A Database of German Emotional Speech," Interspeech 2005, Lisbon.

J. Diehl et al (2009), "An acoustic analysis of prosody in high-functioning autism", *Applied Psycholinguistics*, 30(3).

R. el Kaliouby et al. (2009). "An Exploratory Social-Emotional Prosthetic for Autism Spectrum Disorders," in *Body Sensor Networks*. 2006. MIT Media Lab.

R.B Fink.et al (2009). "Evaluating Speech Recognition in a Computerized Naming Program for Aphasia," American Speech-Language Hearing Association Conference. New Orleans, November.

R. B. Fink et al. (2002). "A computer implemented protocol for treatment of naming disorders: Evaluation ofclinician-guided and partially self-guided instruction," A*phasiology*,16(10/11):1061-1086.

B. Elvevaag, P. Foltz, D. Weinberger, and T. Goldberg (2007), "Quantifying Incoherence in Speech: an Automated Methodology and Novel Application to Schizophrenia," Schizophrenia Research, 93:304-316.

B. Elvevaag, P. Foltz, M Rosenstein, and L. DeLisi (2009), "An automated method to analyze language use in patients with schizophrenia and their first degree-relatives," Journal of Neurolinguistics.

W. Goldfarb et al. (1972),. "Speech and language faults in schizophrenic children. Journal of Autism and Childhood Schizophrenia, 2(3):219-233, 1972.

P. Gupta &N. Rajput, (2006), "Two-Stream Emotion Recognition For Call Center Monitoring", Interspeech 2006, Pittsburgh.

Gottschalk, L., Winget, C., & Gleser, G. (1969). Manual of instructions for using the Gottschalk-Gleser content analysis scales: Anxiety, hostility, and social alienation-personal disorganization. Berkeley: University of California Press.

K. Graves et al. (2005), "Emotional expression and emotional recognition in breast cancer survivors: A controlled comparison," *Psychology and Health*, 20:579-595.

M. E. Hoque et al. (2009),"Exploring Speech Therapy Games with Children on the Autism Spectrum," Interspeech 2009*,* Brighton.

T. Johnstone et al (2006), "The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions,"*Social, Cognitive and Affective Neuroscience, 1(3),* 242-249.

L. Kanner (1946), "Irrelevant and metaphorical language in early infantile autism,"*American Journal of Psychiatry*, 103:242-246.

L. Kanner (1948), "Autistic Disturbances of Affective Contact," *Nervous Child*, 2:217-2520.

C. M. Lee and S. Narayanan (2004), "Towards detecting emotionsin spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, 2004.

S. Lee et al (2006), "A Study of Emotional Speech Articulation using a Fast Magnetic Resonance Imaging Technique," Interspeech 2006, Pittsburgh.

M. Le Normand et al (2008), "Prosodic disturbances in autistic children speaking French, Speech Prosody," Campinas, Brazil.

M. Lehtinen (2008), "The prosodic and nonverbal deficiencies of French- and Finnish-speaking persons with Asperger Syndrome," Proceedings of the ISCA Workshop on Experimental Linguistics, Athens.

M. Levit et al (2001), "Use of prosodic speech characteristics for automated detection of alcohol intoxication," ISCA Workshop on Prosody in Speech Recognition and Understanding, Red Bank NJ.

Linguistic Data Consortium, "Emotional prosody speech and transcripts," LDC Catalog No.: LDC2002S28, University of Pennsylvania.

J. Liscombe et al (2005), "Using Context to Improve Emotion Detection in Spoken Dialog Systems," Interspeech 2005, Lisbon.

J. Liscombe et al (2006), "Detecting Certainness in Spoken Tutorial Dialogues," Interspeech 2006, Pittsburgh.

X. Luo et al (2006), "Vocal Emotion Recognition with Cochlear Implants," Interspeech 2006, Pittsburgh.

A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth (2009), "PEAKS – A systems for the automatic evaluation of voice and speech disorders," Speech Communication 51 (2009):425-437.

F. Mairesse and M. Walker (2006), "Automatic Recognition of Personality in Conversation," HLT-NAACL 2006, New York City.

G. Mesibov (1992). "Treatment issues with high-functioning adolescents and adults with autism," In E. Schopler & G. Mesibov (Eds.), *High-functioning individuals with autism* (pp. 143–156). New York: Plenum Press.

Elliot Moore II, Mark Clements, John Peifer and Lydia Weisser (2003), "Investigating the Role of Glottal Features in Classifying Clinical Depression," IEEE EMBS, Cancun.

S. Mozziconacci and D. J. Hermes (1999), "Role of intonation patterns in conveying emotion in speech," ICPhS 1999, San Francisco.

Mundt, J. et al (2007), "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of Neurolinguistics*, 20(1):50-64.

P. Oudeyer (2002), "Novel useful features and algorithms for the recognition of emotions in human speech," Speech Prosody 2002, Aix-en-Provence.

T. Oxman, S Rosenberg, P. Schurr, and G. Tucker (1988), "Diagnostic Classification Through Content Analysis of Patient Speech," American Journal of Psychiatry. 1988. 145:464-468.

Pennebaker, J. et al (2001), *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum.

J. Pennebaker, M. Mehl, and K. Niederhoffer (2003), "Psychological Aspects of Natural Language Use: our Words, our Selves," Annu. Rev. Psychol. 2003. 54:547–77.

J. Pestian, P. Matykiewicz, J. Grupp-Phelan, S. Arszman Lavanier, J. Combs, and R. Kowatch (2008), "Using Natural Language Processing to Classify Suicide Notes," ACL BioNLP Workshop, pp. 96-97.

Paul, R et al (2008) “Production of syllable stress in speakers with autism spectrum disorders,” *Research in Autism Spectrum Disorders,* 2:110-124.

R. Ranganath, D. Jurafsky, and D. McFarland (2009), “It’s Not You, it’s Me: Detecting Flirting and its Misperception inSpeed-Dates,” EMNLP 2009, Singapore.

Rapin, I., and Dunna, M. (2003), “Update on the language disorders of individuals on the autistic spectrum,” *Brain Development*. 25:166–172.

Shriberg, L. et al, (2001), “Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome,” *Journal of Speech, Language, and Hearing Research*; 44(5)

P. Stone, D. Dunphy, M. Smith, et al (1969), “The General Inquirer: A Computer Approach to Content Analysis,” Cambridge, Mass. MIT Press.

van Santen, J. et al (2009), “Automated assessment of prosody production,” *Speech Communication* 51:1082–1097.

J. Yuan et al (2002), “The acoustic realization of anger, fear, joy, and sadness in Chinese,” ICSLP, Denver.

Zei Pollerman, B. (2002), “A Study of Emotional Speech Articulation using a Fast Magnetic Resonance Imaging Technique,” Speech Prosody 2002, Aix-en-Provence.

E. Zetterholm (1999), “Emotional speech focusing on voice quality,”FONETIK: The Swedish Phonetics Conference, Gothemburg.