Algorithms for Extracting Robust and Accurate Speech Features and Their Application in Clinical Domain

Meysam Asgari

Presented to the Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Computer Science & Engineering

August 2014

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of
Meysam Asgari
has been approved.

Izhak Shafran, Thesis Advisor
Research Scientist, Google Inc.

Brian Roark,
Research Scientist, Google Inc.

Alexander Kain,
Associate Professor, OHSU

John H.L. Hansen,
Professor, University of Texas at Dallas

Kemal Sonmez,
Associate Professor, OHSU

Acknowledgment

First of all, my deepest thanks go to my adviser , Izhak Shafran, for his continuous support and encouragement he has generously offered in the past five years. I would also like to thank him for the insightful guidance he has given to me through this journey.

I would like to sincerely thank all the members of thesis committee, Brian Roark, Alexander Kain, John Hansen,Kemal Sonmez for their invaluable suggestions helped me improving the quality of the thesis. I am grateful to all of the faculty at CSLU, who not only taught me everything I know about speech and language processing but also willingly offered their input and encouragement. I extend my thanks to my fellow students, both past and present, specially Maider Lehr, Alireza Bayestehtashk, Géza Kiss, Eric Morley, Masoud Rouhizadeh, Mahsa Yarmohammadi, Mahsa Elyasi, and Hamidreza Mohammadi.

I would like to thank Mathias Mehl for making his corpus of naturalistic samples of everyday conversations available for this thesis. I would also like to thank Lisa Sheeber for providing me with the adolescents' speech database for conducting experiment on detection of clinical depression.

Finally, words cannot adequately express my gratitude to my family for their sage advice and their unshakable confidence in me; and to my dear wife, Narges, for her patience support.

Contents

List of Tables

List of Figures

Abstract

Algorithms for Extracting Robust and Accurate Features and Their Application in Clinical Domain

Meysam Asgari

Doctor of Philosophy
Center for Spoken Language Understanding within
Oregon Health & Science University
School of Medicine

August 2014
Thesis advisor: Izhak Shafran

Abstract

Samples of everyday conversations are being collected and analyzed in a growing number of applications, ranging from studying behavior in social psychology to clinical assessment of voice pathology and even cognitive function. Aside from the spoken words, the acoustic properties of speech samples can provide important cues in these applications. The goal of this study is developing novel algorithms for robust and accurate estimation of speech features and employing them to build probabilistic speech models for characterizing and analyzing clinical speech. We aim to achieve accurate and reliable estimation of voiced segments, fundamental frequency, harmonic-to-noise ratio (HNR), jitter, and shimmer for clinical speech analysis. Towards this goal, we adopt a harmonic model (HM) of speech. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer. We evaluate the performance of our improved HM in the context of voicing detection and pitch estimation with other state-of-the-art techniques on the Keele data set. Through extensive experiments on several noisy conditions, we demonstrate that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments. We also employ our improved harmonic model for developing a novel algorithm to estimate jitter, shimmer, and HNR that is less sensitive to noise and can also capture variations within the frame. We further verify the robustness of these measures on detecting disordered voices due to Parkinson's disease (PD).

Next, we turn our attention to investigate the utility of developed measures in clinical applications. We perform empirical studies on the speech-based assessment of cognitive impairments including PD, autism spectrum disorder (ASD), and clinical depression. We demonstrate that the severity of PD can be inferred from speech with a mean absolute error of about 5.5, explaining 61% of the variance and consistently well-above chance. Leveraging the same mechanisms developed for inferring PD, we detect children with ASD and classify them into four categories. We find that our features improve the performance,

measured in terms of unweighted average recall (UAR), of detecting ASD by 2.3% and classifying the disorder into four categories by 2.8% compare to a state-of-the-art baseline performance. We also examine the use of our features in detection of clinical depression in adolescents. We conduct experiments to compare the performance our developed features with that obtained from *openSMILE*, a standard feature extraction tool. Our experiment show that our extracted features from HM improve the performance of detecting depression from spoken utterances for speaker-level decisions. Finally, we explore the feasibility of detecting social contexts from audio recordings of everyday life such as in life-log. Again, we find that the features developed in this thesis perform better than MFCC and *OpenSMILE* features in these tasks. This is true even when we apply recently developed deep neural networks (DNNs) for classification, achieving classification accuracy as high as 87.7% and 86.8% for speakers' *location* and *activity*.

Chapter 1

Introduction

Analysis of acoustic signals of the human voice has many purposes. Our voice reveals considerable insight into the structure and function of certain organs involved in speech and language production. For instance, sometimes the first symptom of a neurological disorder such as Parkinson's disease (PD) is a speech impairment [1]. PD can affect all components of speech production including breathing, laryngeal function, articulation as well as their coordination for the production of smooth speech. Resulting dysarthric speech often exhibits monotonous pitch, variable speech rate, harsh and breathy voices [2]. Likewise, researchers have shown the effects of psychological disorders, such as depression, in patients' voice [3]. Speech pathologists have characterized depressed speech as monotonous, mono-loud, and low in range of pitch frequency [4]. Also, a number of studies have shown that emotional arousal considerably influences phonatory and articulatory aspects of the speech production system [5]. In addition to PD and depression, autism spectrum disorder (ASD) is another example that affects patient's voice. Children with autism often exhibit unusual pitch and intonation, for example, monotonous pitch, reduced stress, odd rhythm, large pitch range [6], and even differences in harmonic structure of their speech [7].

The severity of aforementioned diseases is typically assessed subjectively by an expert practitioner and often requires patient's presence at the clinic. This assessment is often costly and time-consuming, and can be burdensome in some situations, for example, when a patient must undergo frequent reassessments. For several decades now, observed symptoms in speech of patients with such diseases have motivated researches to explore alternative approaches based on speech processing techniques. Researchers have measured these symptoms more objectively with the hope of augmenting or simplifying the assessment. It is often cheaper and easier to automatically elicit, record, and analyze speech than conducting in-person clinical assessment. Furthermore, speech-based assessment can be remotely administered and can objectively monitor changes over time. Easier methods of assessments, such as automated screening and telemonitoring, can play a crucial role in the early detection of the aforementioned diseases.

The main focus of the thesis is developing novel algorithms for robust and accurate estimation of pitch-related features; and employing them to build probabilistic speech models for characterizing and analyzing clinical speech. There are a number of approaches

in time and frequency domain to estimate pitch-related features. Time domain methods often ignore frequency and amplitude variations of speech over the analysis frame, and on the other hand, the resolution of short time Fourier transform does not provide the necessary time-frequency resolution to capture small amount of perturbation observed in, for example, Parkinson's disease (PD). Alternatively, we adopt the harmonic model of speech, a model that has recently gained a considerable attention. This model takes into account the underlying harmonic nature of voiced speech and decomposes it into a harmonic and a non-harmonic component. We overcome certain drawbacks of this model and introduce an improved version of HM that leads us to accurate and reliable estimation of voiced segments, fundamental frequency, HNR, jitter, and shimmer. We then turn our attention to employing these features and applying several supervised machine learning algorithms for clinical assessment of speech impairments in PD, ASD, and clinical depression. We also explore the feasibility of detecting social context from audio recordings of everyday life such as life-log using our developed features. The amount of labeled data for this task is relatively small compared to the available recordings. This lends itself naturally to unsupervised feature extraction using deep neural networks (DNNs), followed by supervised learning of a classifier for social contents.

In the end, this thesis does not create models inspired by the biological process associated with clinical conditions. Instead, we adopt the approach of developing algorithms to extract robust and accurate features from speech that is broadly applicable to a variety of clinical problems.

1.1 Objectives and contributions of the thesis

In the following of this chapter, we briefly describe the thesis motivations and objectives, followed by introducing our approach and contributions.

Robust detection of voiced segments

With the widespread speech interfaces in smartphones and robots, there is a growing demand for robust algorithms to detect speech in presence of a variety of noise conditions.

As in most speech processing systems, voice activity detection (VAD) is the first component of our acoustic feature extractor. Most previous algorithms reported in literature were developed and evaluated on task with at most a handful of noise types [8]. The performance of these algorithms cannot be easily extrapolated to diverse noise backgrounds encountered, for example, in everyday life.

Our goal in this work is to develop a robust algorithm to detect voiced segments in the presence of diverse background noise. Toward this goal, we adopt the harmonic model of speech to exploit the rich harmonic structure of voiced speech that is preserved even in adverse noisy conditions leading to a powerful feature for segregating voiced from unvoiced speech and noise signals. We also overcome the weakness of harmonic model in differentiating speech from stationary harmonic noise using the non-stationary property of speech. Unlike the previous works where model parameters were estimated independently using maximum likelihood framework, we estimate model parameters more robustly using a maximum *a-posteriori* (MAP) criterion. Given the estimates of model parameters, we detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level is computed by formulating a two-state hidden Markov model (HMM). The state could either be voiced or unvoiced, with likelihood given by the per-frame harmonic model. The transition model consists of a simple zero-mean Gaussian. We compute voicing decision over the utterance using Viterbi alignment. Finally, we empirically demonstrate the advantage of our algorithm on two data sets. On the Keele dataset, for all levels of additive noise our algorithm is substantially better than the popular tool, *get-f0*. More importantly, on a large collection of 30-second samples of ambient recordings of everyday life, we show that our algorithm performs significantly better than other popular alternatives.

The work has been published in [9].

Accurate and robust estimation of pitch frequency

Accurate and robust estimation of pitch plays a central role in speech processing. Pitch-related statistics convey considerable information about the emotional state of speakers. Pitch estimators are widely employed in diverse applications which can benefit from better

accuracy and better robustness. For example, pitch tremors in early stages of Parkinson's disease can be subtle and accurate estimator will be useful in automated screening tasks for the disease [10].

Various methods in time, frequency and cepstral domain have been proposed for generating pitch candidates. Most algorithms excel when the background noise is minimal or for specific types of background noise. In this work, we aim to improve the robustness and accuracy of pitch estimation across a wide variety of background noise conditions. For this, we adopt the harmonic model of speech and address two major weakness of this model in the context of pitch estimation: 1) the problem of pitch halving and doubling, and 2) the need to specify the number of harmonics. We propose a local smoothing function that exploits the fact that there is more energy in the harmonics near the true pitch than at the corresponding neighborhoods of half or double the pitch. We use a local smoothing function to accumulate this energy and improve the robustness of the pitch candidates in each frame. Also, using a model complexity term with a Bayesian information criterion (BIC), we chose the optimal number of harmonics. We evaluated our proposed pitch estimation method with other state of the art techniques on the Keele dataset in terms of gross pitch error and fine pitch error. Through extensive experiments on several noisy conditions, we demonstrate that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments.

The work has been published in [11].

Estimating jitter, shimmer, and harmonic-to-noise ratio for quantifying voice quality

Jitter and shimmer refer to a short-term (cycle-to-cycle) perturbation in pitch period and amplitude of voiced speech, respectively. Speech pathologists often measure jitter and shimmer to characterize abnormalities in voice. Automatic measurement of time period and amplitude of each cycle is sensitive to noise. The harmonic model (HM) allows an alternate method to measure jitter and shimmer that is less sensitive to noise. However, standard version of this model is not able to follow cycle-to-cycle variations within the frame of analysis. Alternatively, we employ a harmonic model with varying amplitudes

(HM-VA) that allows capturing cycle-to-cycle variations associated with jitter and shimmer. The key idea of our approach is reconstructing two versions of input waveform in each frame with HM and HM-VA, and then using parameters of both models for estimating jitter and shimmer.

Researchers have been attempting to quantify voice quality for the hope of finding quantitative measures for assessment of pathological speech [12, 13]. Harmonic-to-noise ratio (HNR) is one of the key features that measures the degree of hoarseness in speech. Given the estimated parameters of HM-VA, we decompose a frame of speech into two harmonic and noise components by subtracting the reconstructed signal from the original speech signal. We then compute the HNR and the ratio of energy in first and second harmonics.

These quantities are difficult to evaluate independently due to the lack of grand truth. Instead, we evaluate them with alternative measures extracted from *Praat*, a standard features extraction tool, in the context of detecting disordered voices belong to subjects with Parkinsons disease.

The work has been published in [14].

Investigating the utility of proposed acoustic features for clinical applications

Now, we turn our attention to investigate the utility of developed measures in clinical applications. We apply several machine learning algorithms to perform empirical studies on the speech-based assessment of cognitive impairments including PD, autism spectrum disorder (ASD), and clinical depression. In the following, we will briefly describe our method and our evaluation on aforementioned diseases.

1- Assessing the severity of Parkinson's disease

For several decades now, there has been sporadic interest in automatically characterizing the speech impairment due to Parkinson's disease. In this work, we characterize the disease using a relatively large cohort of 168 subjects, collected from multiple (three) clinics. We elicited speech using three tasks – the sustained phonation task, the diadochokinetic task and a reading task. Apart from standard features from time domain (e.g.,

energy, speaking rate), spectral domain (e.g., pitch, spectral entropy) and cepstral domain (e.g, mel-frequency warped cepstral coefficients), we also estimate harmonic-to-noise ratio, shimmer and jitter using our developed algorithms. Our results show that the severity of the disease can be inferred from speech with a mean absolute error of about 5.5 with respect to the clinical assessment using the Unified Parkinson's Disease Rating Scale (UPDRS); the range of target motor-scale is 0 to 108. Results are consistently well-above chance across all clinics. We also find that the reading task is significantly better at capturing cues than diadochokinetic or sustained phonation task. Moreover, our experiments show that our proposed features are better than alternative features extracted from *openSMILE*.

The work has been published in [10] and [15].

2- Detecting and diagnosing of autism spectrum disorders

Autism spectrum disorder (ASD) covers a range of developmental disabilities that can cause significant social, communication, and behavioral challenges. There has been continual interest in objectively characterizing language impairments in ASD. In this work, we employ speech measures extracted from harmonic model, along with standard features such as energy, cepstral, and spectral features to assess speech disorders in children with autism spectrum disorders (ASD) on the *Autism Sub-Challenge* of Interspeech 2013. The challenge consists of two tasks: 1) a binary 'Typicality' classification task with classes – TYPically developing (TYP) and ATYpically developing (ATY), and a four-way 'Diagnosis' task for classifying children into 4 categories – TYP, pervasive developmental disorders (PDD), PDD not otherwise specified (PDD-NOS), and specific language impairment such as DYSphasia (DYS). We find that our proposed features improved the performance, measured in terms of unweighted average recall (UAR), of detecting ASD by 2.3% and diagnosing the disorder into four categories by 2.8% over the baseline in this challenge, and won the challenge.

The work has been published in [16].

3- Inferring clinical depression from speech and spoken utterances

There has been considerable interest on analyzing acoustic properties of speech for the hope of quantitative assessment of clinical depression. In this work, we investigate the problem of detecting clinical depression from recordings of adolescents collected during their family interactions. The cues for depression may be present in what is said (content) and how it is said (prosody). Given the scarcity of the text data for training models with n-grams, we explore an alternative method to extract content information related to affect by encoding words in terms of valence and arousal, using a look up table that has been compiled by averaging responses from large number of raters [17, 18]. For extracting prosody, we employ standard feature extractors such as those implemented in *openSMILE* and compare them with features extracted from harmonic model. Our experiments show that our features from harmonic model outperform standard features, such as those computed from *openSMILE*, on detecting depression from spoken utterances. The textual features provide additional gain, achieving a classification accuracy of about 74%.

The work has been submitted to [19].

4- Investigating the utility of proposed acoustic features for inferring social contexts

In this study, we investigate the problem of detecting social contexts from the audio recordings of everyday life such as in life-logs. Inferring social contexts from audio life logs are useful to screen large groups and monitor patients in real-world scenarios [20]. This is particularly relevant for older adults who live independently as well as for kids with autism. Our corpus for this study consists of samples of snippets of audio recordings from everyday life of university students. Recording are 30-seconds long recorded every 12 minutes. Unlike the standard corpora of telephone speech or broadcast news, these recordings have a wide variety of background noise.

We investigate the use of our developed features along with deep neural networks (DNNs) to detect social contexts, such as speakers location (e.g., in transit) or activity (e.g., watching TV or eating) in this recordings. DNNs can potentially extract useful

features in an unsupervised manner. The layers of the network that extract these features are then modified in a supervised manner to fine-tune the network for a classification task with limited amounts of labeled data. We find that DNNs can be employed effectively to infer social contexts from audio snippets of everyday life, achieving classification accuracy as high as 87.7% and 86.8% for speakers‘ location and activity. We also find that the features extracted using harmonic model are better than MFCC and *openSMILE* features in these tasks.

The work has been submitted to [21].

1.2 Outline

The remainder of this thesis is organized into four chapters. Chapter 2 describes the harmonic model of speech and the details of parameter estimation framework. Describing the structure of voicing detection algorithms, we briefly review some proposed methods and formulate the harmonic model of speech for detecting voiced segments. We empirically evaluate the efficacy of the model in detecting voiced segments under different noisy conditions and then measure the performance on a large collection of utterances recorded from several speakers over their everyday life. Next, we address certain drawbacks of harmonic model and employ it for the pitch estimation problem. We briefly review popular algorithms that are presented in the empirical evaluations. We empirically evaluate and characterize the proposed improvements to harmonic model using a series of experiments with several background noise types and at different SNRs. Next, we employ a harmonic model with time-varying amplitudes along with a standard harmonic model to estimate voice quality measures including jitter, shimmer, and harmonic-to-noise ratio. Finally, we empirically evaluate the performance of the proposed measures on detecting subjects with voice impairments due to Parkinson's disease (PD) from normal subjects.

Chapter 3 describes the use of speech features along with machine learning algorithms in clinical application. Through supervised approaches, we attempt to automatically characterize speech impairments in PD, autism spectrum disorder (ASD), and clinical depression. We learn probabilistic speech models employing our developed and standard speech

features to objectively assess each of these diseases. The utility of the proposed approach is evaluated on predicting clinical ratings of 168 subjects diagnosed with PD. In the same vein, we detect 35 children with ASD from 54 normal children and diagnose the disease into four categories. Finally, we examine the use of our developed speech models on detecting 71 clinically depressed adolescents from 77 healthy controls.

Chapter 4 describes the feasibility of detecting social contexts from audio recordings of everyday life. We employ deep neural networks (DNNs) and investigate different strategies for training these models using different speech features. We report results on a real-world application, such as detecting speakers location (e.g., in transit) or activity (e.g., watching TV or eating) in audio life logs.

Chapter 5 summarizes the main contributions of the thesis and suggests directions for future work.

Chapter 2

Speech Analysis using Harmonic Model

This chapter starts with the review of the computational model of speech production. Next, we describe the harmonic model of speech and formulate an approach to robustly detect voiced segments. We empirically evaluate the efficacy of the model in detecting voiced segments under efferent noise conditions and measure the performance on a large collection of utterances recorded from several speakers over the course of their everyday lives. Then, we employ this model to estimate pitch frequency. The straightforward application of this model, however, leads to certain types of systematic errors – pitch halving and doubling errors. We propose a method to mitigate these errors. We empirically evaluate and characterize the proposed improvements to harmonic model using a series of experiments with several background noise types and at efferent SNRs. Finally, we formulate an approach using harmonic model to to quantify voice quality. The model allows robust estimation of HNR, shimmer, and jitter. We evaluate these quantities in the context of detecting disordered voices due to Parkinson's disease.

2.1 Speech production model

Our approach is motivated by the computational model of speech production. During voiced sounds, rhythmic opening and closing of vocal folds converts the airflow from the lungs into a sequence of short glottal pulses. These excitation pulses are rich in harmonics and considered as the source of voiced speech. They are subsequently modulated by resonances of the vocal tract and the transfer function of the lip radiation. Unvoiced sounds are generated in a similar manner except they are driven by a noisy source while the vocal folds remains open. The noisy source comprises frication noise, aspiration noise, and the fluctuations produced by the turbulences of the glottal airflow. Individuals with voice disorders usually cannot seamlessly switch between the two sources and therefore, excitation pulses are contaminated by the noise signal.

From a signal processing point of view, speech production process can be modeled by a linear system as shown in Figure 2.1. The voiced and unvoiced sounds are modeled by two separate sources as we mentioned earlier. The effect of the shape of the vocal tract is modeled by $\mathbf{V(z)}$, and the radiation characteristics of the lips are taken into

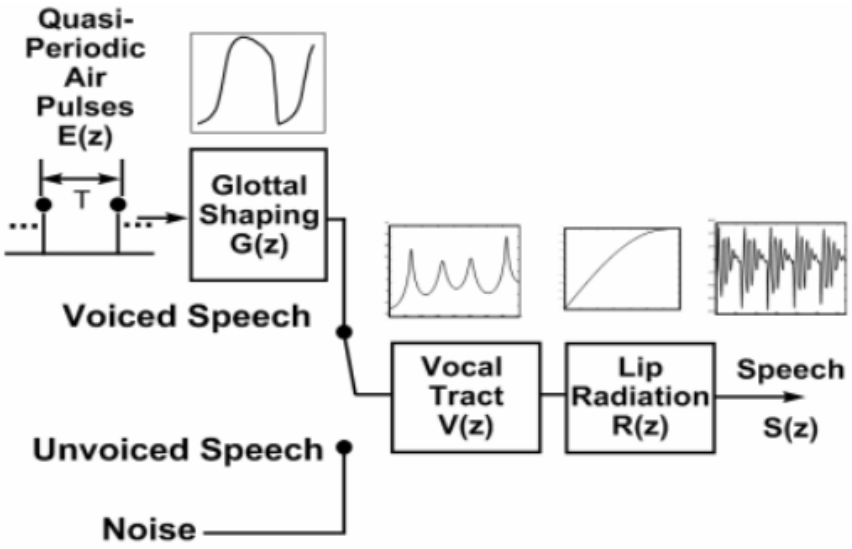

Figure 2.1: A computational model of speech production.

account by **R(z)**. Since the glottal pulses carry the harmonic information of voiced speech, the resulting voiced sounds can be modeled with a harmonic model that separates the harmonic parts from the noise. Such a model has been successfully employed for periodic signal [22].

2.2 Harmonic model

The popular source-channel model of voiced speech considers glottal pulses as a source of period waveforms which is being modified by the shape of the mouth assumed to be a linear channel. Thus, the resulting speech is rich in harmonics of the glottal pulse period. The harmonic model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related, that is, the frequencies of the sinusoids are multiples of the fundamental frequency [23]. This assumption arises from the harmonic nature of speech signal and reduces the number of parameters in the general sinusoidal model. Stylianou [23] introduced a harmonic plus noise model (HNM) for speech analysis and synthesis. The observed voiced signal in HNM is represented in terms of a harmonic component and a non-periodic component related to noise. Speech decomposition using a

HNM is useful for applications in speech synthesis, voice conversion, speech enhancement, and speech coding.

2.2.1 Model expression

Let $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_N)]^T$ denote the N speech samples in a voiced frame, measured at times $t_1, t_2, \ldots, t_N$. The samples can be represented with a harmonic model with an additive noise $\mathbf{n} = [n(t_1), n(t_2), \ldots, n(t_N)]^T$ as follow:

$$\begin{aligned} s(t) &= a_0 + \sum_{h=1}^{H} a_h \cos(2\pi f_0 h t) + b_h \sin(2\pi f_0 h t) \\ y(t) &= s(t) + n(t) \end{aligned} \tag{2.1}$$

where H denotes the number of harmonics and $2\pi f_0$ stands for the fundamental angular frequency. The harmonic signal can be factored into coefficients of sinusoidal functions, α, β, and the harmonic components which are determined by the given angular frequency, $2\pi f_0$, and model order, H.

$$\begin{aligned} s(t) &= \begin{bmatrix} 1 & A_c(t) & A_s(t) \end{bmatrix} \begin{bmatrix} a_0 \\ \alpha \\ \beta \end{bmatrix} \\ A_c(t) &= \begin{bmatrix} \cos(2\pi f_0 t) & \cdots & \cos(2\pi f_0 H t) \end{bmatrix} \\ A_s(t) &= \begin{bmatrix} \sin(2\pi f_0 t) & \cdots & \sin(2\pi f_0 H t) \end{bmatrix} \\ \alpha &= \begin{bmatrix} a_1 & \cdots & a_H \end{bmatrix}^T \\ \beta &= \begin{bmatrix} b_1 & \cdots & b_H \end{bmatrix}^T \end{aligned} \tag{2.2}$$

Stacking rows of $[1 \quad A_c(t) \quad A_s(t)]$ at $t = 1, \cdots, T$ into a matrix $\mathbf{A}$, Equation 2.2 can be compactly represented in a matrix notation as:

$$\mathbf{y} = \mathbf{A}\,\mathbf{m} + \mathbf{n} \tag{2.3}$$

where $\mathbf{y} = \mathbf{A}\,\mathbf{m}$ corresponds to a expansion of the harmonic part of voiced frame in terms of windowed sinusoidal components, and $\Theta = [f_0, \mathbf{m}, \sigma_n^2, H]$ is the set of unknown parameters.

2.2.2 ML estimation of model parameters

Assuming the noise samples, $\mathbf{n}$, are independent and identically distributed random variables with a zero-mean Gaussian distribution, the likelihood function of the observed vector, $\mathbf{y}$, given the model parameters can be formulated as following equation.

$$\begin{aligned} \mathbf{L}(\Theta) &= \log p(\mathbf{y}|\mathbf{\Theta}) \\ &= -\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}||\mathbf{y} - \mathbf{Am}||^2 \end{aligned} \tag{2.4}$$

Assuming that unknown parameters are independent with respect to each other, the parameters of vector $\mathbf{m}$ (while the other parameters are kept constant) can then be estimated by a maximum likelihood (ML) approach.

$$\hat{\mathbf{m}}_{ML} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{2.5}$$

We then substitute $\hat{\mathbf{m}}$ into Equation 2.5 and the likelihood function can be written as follow:

$$\begin{aligned} \mathbf{L}(f_0, \hat{\mathbf{m}}, \sigma_n^2) &= -\frac{N}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}||^2 \\ &= -\frac{N}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}(\mathbf{y} - \mathbf{\Gamma})^T(\mathbf{y} - \mathbf{\Gamma}) \end{aligned} \tag{2.6}$$

where $\mathbf{\Gamma} \equiv \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the projection matrix into the subspace spanned by the columns of $\mathbf{A}$. The noise variance is maximized by taking the derivative of Equation 2.6 with respect to σ_n^2 and making it zero.

$$\begin{aligned} \frac{\partial}{\partial\sigma_n^2}\mathbf{L}(f_0, \hat{\mathbf{m}}, \sigma_n^2) &= -\frac{N}{2\sigma_n^2} + \frac{1}{2\pi\sigma_n^4}\mathbf{y}^T(I - \Gamma)\mathbf{y} \\ \hat{\sigma_n^2} &= -\frac{1}{N}\mathbf{y}^T(\mathbf{I} - \mathbf{\Gamma})\mathbf{y} \end{aligned} \tag{2.7}$$

We then substitute the estimate of σ_n^2 into the likelihood function as follows:

$$\begin{aligned} \mathbf{L}(f_0, \hat{\mathbf{m}}, \hat{\sigma_n^2}) &= -\frac{N}{2}\log(\mathbf{y}^T(\mathbf{I} - \mathbf{\Gamma})\mathbf{y}) - \frac{N}{2}(1 + \log(\frac{2\pi}{N})) \\ &= -\frac{N}{2}\log(\mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{y}^T\mathbf{\Gamma}\mathbf{y}) + C \end{aligned} \tag{2.8}$$

where C is a constant factor does not affect the likelihood. It can be shown that $\mathbf{\Gamma}$ is a symmetric matrix where $\mathbf{\Gamma}^T\ \mathbf{\Gamma} = \mathbf{\Gamma}$. Also, under the harmonic model, the reconstructed

signal $\hat{\mathbf{s}}$ is given by $\hat{\mathbf{s}} = \mathbf{A}\ \mathbf{m}$ and the noise component is obtained by subtracting the reconstructed signal from the original speech signal, $\hat{\mathbf{n}} = \mathbf{y} - \hat{\mathbf{s}}$. Now, using the properties of $\mathbf{\Gamma}$, we can rewrite the reconstructed signal and the likelihood function in the Equation 2.9 as follows.

$$\begin{aligned} \hat{\mathbf{s}} &= \mathbf{\Gamma}\ \mathbf{y} \\ \mathbf{L}(f_0) &= -\frac{N}{2}\log(\mathbf{y}^T\ \mathbf{y} - (\mathbf{\Gamma}\ \mathbf{y})^T(\mathbf{\Gamma}\ \mathbf{y})) + C \\ &= -\frac{N}{2}\log(||\mathbf{y}||^2 - ||\hat{\mathbf{s}}||^2) + C \\ &= -\frac{N}{2}\log(||\hat{\mathbf{n}}||^2) + C \end{aligned} \tag{2.9}$$

Maximizing the likelihood function in the above equation is equivalent to minimizing the energy of residual noise component ignoring the additive constant factor. The energy of residual part is minimum when the reconstructed signal is fitted well to the original signal. Thus, pitch frequency can be estimated by seeking the frequency that maximizes the energy of the reconstructed signal over the pre-determined grid of discrete f_0 values ranging from $f_{0\ min}$ to $f_{0\ max}$.

$$\hat{f}_{0\ ML} = \arg\max_{f_0}\ \hat{\mathbf{s}}^T\hat{\mathbf{s}} \tag{2.10}$$

2.2.3 MAP estimation of model parameters

The ML estimate of model parameters ignores the continuity of pitch across frames and can potentially over-fit the data. We propose to improve the robustness of the model by exploiting the fact that model parameters cannot vary arbitrarily across frames from the same speaker. Additionally, the physical shape of vocal tract limits the variation of harmonic coefficients into a bounded subspace, and they are not allowed to vary in an arbitrary subspace. From a Bayesian point of view, it is equivalent to imposing a constraint over the space of harmonic components in the model. The maximum *a posteriori* estimate of the model parameters can be factored as follow.

$$\begin{aligned} \hat{\Theta}_{MAP} &= \arg\max_{\Theta} p(\Theta|\mathbf{y}) \\ &= \arg\max_{\Theta} p(\mathbf{y}|\Theta)p(\Theta) \end{aligned} \tag{2.11}$$

The likelihood of a voiced frame, $p(\mathbf{y}|\Theta)$, is estimated from Equation 2.4. For the simplicity, the prior term, $p(\Theta)$, is factored as $p(\Theta) = p(\mathbf{m})p(\omega)$ where $p(\omega)$ is a uniform distribution from 50 to 500 Hz. In our data, we observe that the coefficients of harmonic estimated independently per frame using the Equation 2.5, are approximately Gaussian, as illustrated in Figure 2.2. Hence, we model the prior $p(\mathbf{m})$ as multivariate Gaussian

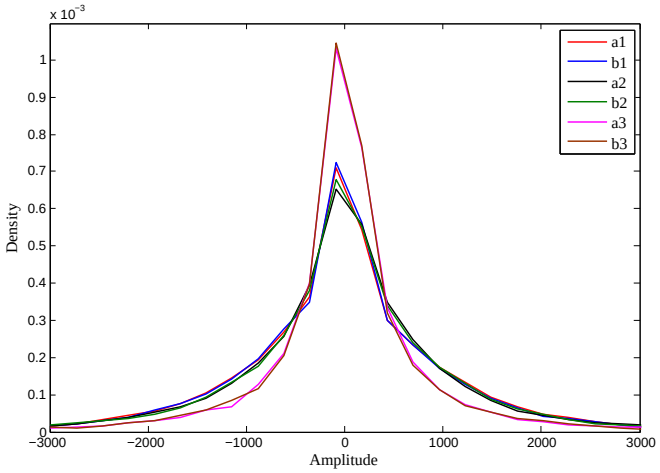

Figure 2.2: The empirical distribution of the first three coefficients of harmonic sinusoids and cosines in the voiced frames of the Keele dataset.

distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Since the likelihood and the prior are in the same distribution family, Gaussian, the prior is *conjugate prior* and we can obtain a closed form MAP estimate by differentiating the Equation 2.12.

$$\frac{\partial}{\partial \mathbf{m}} \log p(\Theta|\mathbf{y}) = \frac{2\mathbf{A}^T}{\sigma_n^2}(\mathbf{y} - \mathbf{A}\,\mathbf{m}) + 2\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \mathbf{m})$$

The derivative is set to zero and the closed form analytical expression for the MAP estimate can be computed.

$$\hat{\mathbf{m}}_{MAP} = (\mathbf{A}^T\mathbf{A} + \sigma_n^2\boldsymbol{\Sigma}_m^{-1})^{-1}(\mathbf{A}^T\mathbf{y} + \sigma_n^2\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m) \tag{2.12}$$

In practice, for each utterance we compute the per-frame estimate of the model parameters and from those estimates, we estimate the prior distribution $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Note, Bayesian estimate of all the model parameters is significantly more complex and requires expensive

numerical approximations compared to our simpler MAP estimate with its closed form analytical solution [24]. The MAP estimate derived in a related previous work smooths the likelihood using a first order HMM transition model and hence differs from our approach [25].

2.2.4 Model order selection

Another problem with the harmonic model is the need to specify the number of harmonics considered. This is typically not known *a priori* and the optimal value can be different in different noise conditions. Davy et al. proposed a sampling-based method for estimating the number of harmonics [24]. Their approach is based on Monte Carlo sampling and requires computationally expensive numerical approximations. Mahadevan employs the Akaike information criterion (AIC) for tackling the problem of model order selection [26]. The AIC attempts to make a balance between the *goodness of fit* of the model and the model complexity by adding a penalty term to the likelihood. Here, we follow a Bayesian approach trying to maximize the likelihood function given by:

$$\hat{H} = \arg\max_{H} p(\mathbf{y}, \Theta_H) \tag{2.13}$$

where Θ_H denotes the model constructed by H harmonics. The likelihood increases as a function of increasing model order and often leads to overfitting. We adopt the Bayesian information criterion (BIC) as a model order selection criterion, where the increase in the likelihood is penalized by a term that depends on the model complexity or the number of model parameters. The BIC has been been widely used for model order selection in machine learning problems, such as linear regression and time series [27, 28]. For the harmonic model, we include a term that depends on the number of data points in analysis window, N, as follow:

$$\begin{aligned} BIC(H) &\approx -2\log p(\mathbf{y}|\Theta_H) + H\log N \\ &\approx -2\mathbf{L}(\Theta_H) + H\log N \end{aligned} \tag{2.14}$$

where the first term is likelihood and the second therm is a penalty. In a related work, Mads et al. [29] propose a similar Bayesian approach with an additional penalty term

compare to the Equation 2.15 as follow.

$$BIC(H) \approx -2\mathbf{L}(\Theta_H) + H \log N + \frac{3}{2} \log N \tag{2.15}$$

Then, for each frame of speech, they select the optimum model order by seeking the number of harmonics, ranging from $H = 2, 3, \ldots, H_{max}$, that minimizes the BIC metric in Equation 2.15. Alternatively, we compute the average frame-level BIC using Equation 2.15 for different model orders. This is a suboptimal solution leading to a unique number of harmonics for all frames of analysis. One drawback of this suboptimal model order selection is possibility of pitch halving or doubling estimates for some frames due to either overfitting or underfitting of the model. But, we show in Sub-section 2.4.5 under *Model sensitivity* that smoothing of likelihood using a local smoothing function tackles this drawback and makes the model robust against model order variation. We will describe the procedure of likelihood smoothing in detail in Sub-section 2.4.3.

2.3 Voicing detection

2.3.1 Related research

Advances in various speech applications, such as the widespread interface, require reliable methods for detecting voiced segments, particularly, in presence of a wide variety of noise types. Detection of voiced segments, often known as voice activity detection (VAD), is the problem of segregating voiced speech from non-speech signals including silence, background noise, and unvoiced speech. VAD is a key module in most speech applications, such as speech coding, speech recognition, speech enhancement, and speaker verification. The accuracy of VAD directly affects the overall performance of these systems and inaccurate voicing determination is an important cause of systematic errors. Most previous algorithms reported in the literature effectively operate when the background noise is minimal or for stationary noises. However, they generally suffer at low signal-to-noise ratios (SNRs) or in presence of non-stationary noises.

Traditionally, VAD consists of two stages. First, an inherent acoustic feature that can discriminate speech from non-speech is extracted for each frame. Next, speech/non-speech decision is made by a decision rule. A variety of acoustic features have been

proposed in VAD algorithms, such as short-time energy, zero-crossing rate [30], spectral entropy [31], cepstral coefficients [32], and periodicity measures [33]. After extracting a set of features for a frame, a binary classifier determines the class which the frame belongs to. The simplest approach is comparison of the feature value with a pre-defined threshold usually extracted from non-speech segments. However, this simple decision rule in not able to track the background noise variations leading to misclassification errors. Several methods based on statistical models have been proposed to enhance the performance of threshold-based VAD algorithms [34, 35, 36, 37, 38]. Sohn et al. [35] proposed a statistical model based VAD where the decision rule is derived from the likelihood ratio test (LRT). This method assumes that discrete Fourier transform (DFT) coefficients of noisy speech and background noise are independent Gaussian random processes. Thus, they can be independently modeled by Gaussian distributions. Then, a hypothesis test where true and null hypotheses represents speech and non-speech frames is conducted. Finally, to take into account the correlation of consecutive occurrences of speech or non-speech frames, a first-order Markovian process is employed for smoothing the decisions. There are other alternative statistical distributions that have been proposed to obtain better speech and noise models. Gazor et al. [37] employ a zero-mean Laplacian distribution to model discrete cosine transform (DCT) coefficients of speech, and Martin [39] uses a Gamma distribution to model model DFT coefficients of speech. Recently, Fisher et al. [40] proposed a harmonic model (HM) in a generalized LRT framework for voiced/unvoiced decision. The key idea of their method is computing the ratio of energy between harmonic and non-harmonic components of speech decomposed by harmonic model.

In summary, researchers have proposed several supervised approaches based on machine learning techniques, such as Gaussian mixture model (GMM) [41, 42], hidden Markov models (HMM) [43, 44], support vector machine (SVM) [45, 46], and deep neural network (DNN) [47, 48]. These techniques can benefit from advantages of learning from multiple features that can be potentially extracted from speech signal. When a representative training corpus is available, parametric models of speech and noise such as HMMs, can be

trained over a set of features and obtain high performance reliably. However, the diversity of background noise in some applications precludes the possibility of collecting and labeling representative samples for every type of noise.

2.3.2 Formulating harmonic model for detecting voiced segments

The harmonic structure of voiced speech is preserved even in adverse noisy condition and this is a powerful feature for segregating it from the unvoiced and noise signals.

Detecting voiced frames can be cast into a hypothesis test problem, in which true and null hypotheses, H_1 and H_0, are defined as follow

$$\begin{aligned} H_1 &: \mathbf{y} = \mathbf{s} + \mathbf{n} \\ H_0 &: \mathbf{y} = \mathbf{n} \end{aligned} \tag{2.16}$$

where, H_1 and H_0 represent the voiced and unvoiced frames. Note that in this problem, our definition of unvoiced includes silent and noise frames as well. Once the model parameters are estimated, we can readily compute the likelihood of observing voiced (v) and unvoiced (u) frames under H_1 and H_0 hypotheses.

$$\begin{aligned} L(H_1) &= -\frac{N}{2}\log(||\mathbf{y}||^2 - ||\hat{\mathbf{s}}||^2) + C \\ L(H_0) &= -\frac{N}{2}\log(||\mathbf{y}||^2) + C \end{aligned} \tag{2.17}$$

where the constant factor C effects both likelihoods equally and it is dropped. By comparing the likelihoods, a frame is classified as either voiced or unvoiced. Now that we can compute probability of observing a voiced or unvoiced frame, the frame-level scores is smoothed to obtain a segment-level decision using a hidden Markov model (HMM), as in [25]. Specifically, this is achieved using an HMM with two states, the voiced (v) and the unvoiced (u) states, whose observation probabilities are modeled using Equation 2.17. Figure 2.3 shows a two-stated HMM for voicing detection. There is an inter-frame temporal dependency in the sequence of speech and non-speech frames that can be modeled by a first-order Markov process. The transition probabilities, probabilities of staying in the same state and transition across states, represented by two parameters can be tuned for a task. With this HMM, the voiced segments of any utterance is computed using a

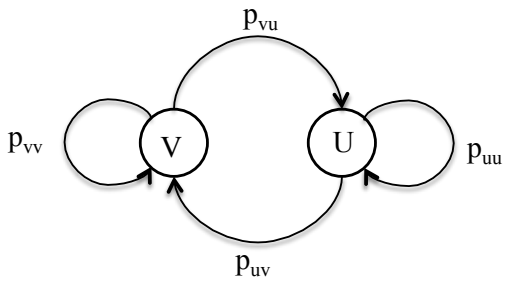

Figure 2.3: A two-state HMM for voicing detection

Viterbi search. Unlike HMMs trained on cepstral features for speech recognition tasks, the parameters of the observation probability are estimated for each frame from the observations themselves. One additional concern that needs to be addressed is the possibility that the background noise (e.g., fan noise) may also be rich in harmonics. We utilize the non-stationary property of speech to distinguish it from stationary harmonic noise. We condition the observation probability $p(\mathbf{y}|v)$ with additional indicator variable t, which is then factored.

$$p(\mathbf{y}|v,t) \quad = \quad p(\mathbf{y}|v)p(\delta\mathbf{m}|t,v) \tag{2.18}$$

Here, $\delta\mathbf{m}$ is the difference between the harmonic coefficients of the current frame and its neighbors. This difference will be low for stationary harmonic noise and is modeled as a univariate Gaussian with just two variables. Note that the number of adjacent frames in the neighborhood of current frame is tuned to a new task with very few examples.

2.3.3 Evaluation

Keel dataset

We evaluate the performance of our proposed method on a task of detecting voiced frames using controlled experiments on the well-studied Keele dataset [49]. The dataset contains 10 phonetically balanced audio files from 10 speakers, 5 males and 5 females. This dataset provides a reference pitch and voicing labels obtained from the simultaneously recorded laryngograph signal. For evaluation, we exclude frames for which the voicing label in the

corpus is uncertain. The speech was recorded in noise free conditions and for testing the robustness of our algorithm, we contaminated them at different noise level ranging from 0 db to 20 db in several noisy environments, including *restaurant*, *subway*, *white*, *car*, *street*, *exhibition*, *babble*, and *airport* taken from Aurora noise dataset [8]. For adding the noise signal, we used Filtering and Noise adding Tool (FaNT) [8] with the telephone speech characteristics configuration using G.712 filter, a narrow-band telephone speech bandpass filter with a flat frequency response between approximately 300 to 3400 Hz. This filtering makes the task of pitch estimation more challenging compare to the full-band scenario due to the spectral attenuation at harmonics bellow 300 Hz.

We compare the performance of our algorithm under both ML and MAP estimation with that of *get-f0*, an algorithm employed in many popular tools (wavesurfer, praat,etc). The *get-f0* employs hand-tuned pre-processing followed by normalized cross correlation (NCC) as a periodicity measure for both discriminating voiced from unvoiced frames and estimating pitch frequency. For our algorithm, the pitch estimation was performed by searching over the frequency range of 50-500Hz with a resolution of 1Hz.

Results

We chose detection trade-off (DET) curve, illustrated further in Figure 2.3.3, to evaluate the accuracy of proposed method. In this curve, false accept is the percentage of voiced frames incorrectly classified as unvoiced frames and false alarm is the percentage of unvoiced frames incorrectly identified as voiced frames. DET curves plotted in the Figure 2.3.3 summarizes the performance of ML and MAP version of HM with that of *get-f0*. For better understanding of the robustness of proposed method under diverse noise conditions, we averaged the results over all aforementioned noise types. As expected, the error in detecting voiced segments increases with noise irrespective of the method employed. The results show that all methods have comparable performance under clean condition. But, the MAP version of HM consistently outperforms the *get-f0* and ML version of HM in noisy conditions.

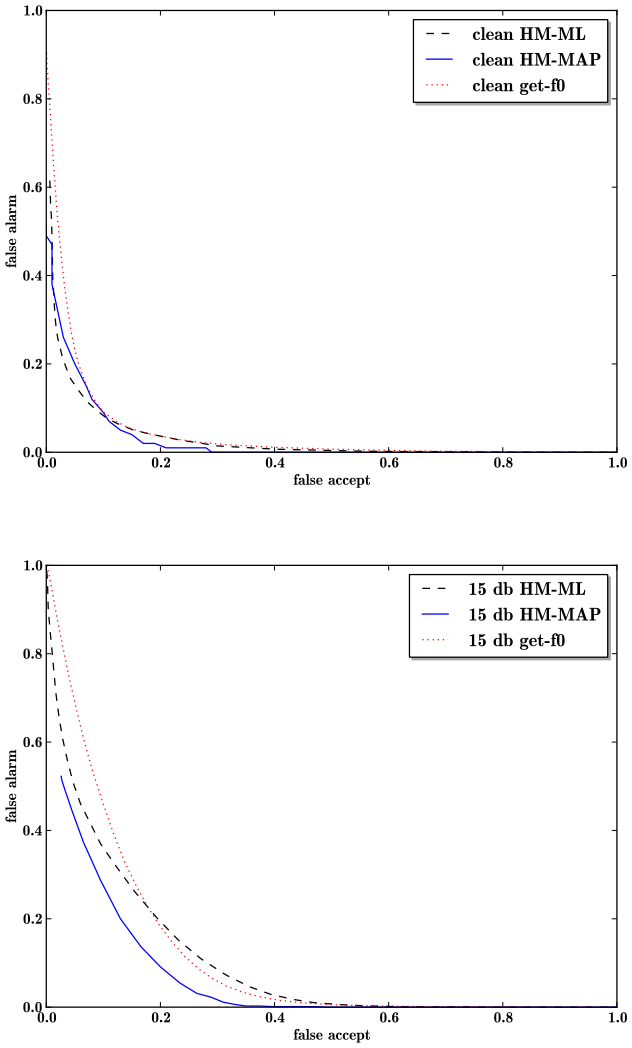

Figure 2.4: DET curve for detecting voiced frames with *get-f0*, ML and MAP versions of harmonic model in the Keele dataset under clean and noisy conditions at 15 db SNRs

Corpus of Everyday Conversations

Next, we evaluate the performance of our algorithm on a large corpus of samples of everyday conversations [50]. These recordings were collected 97 students using a lavaliere microphone and a digital recorder. The recorder was timed to record 30-seconds clips every 12 minutes during the participant's day. The recordings have been transcribed by research

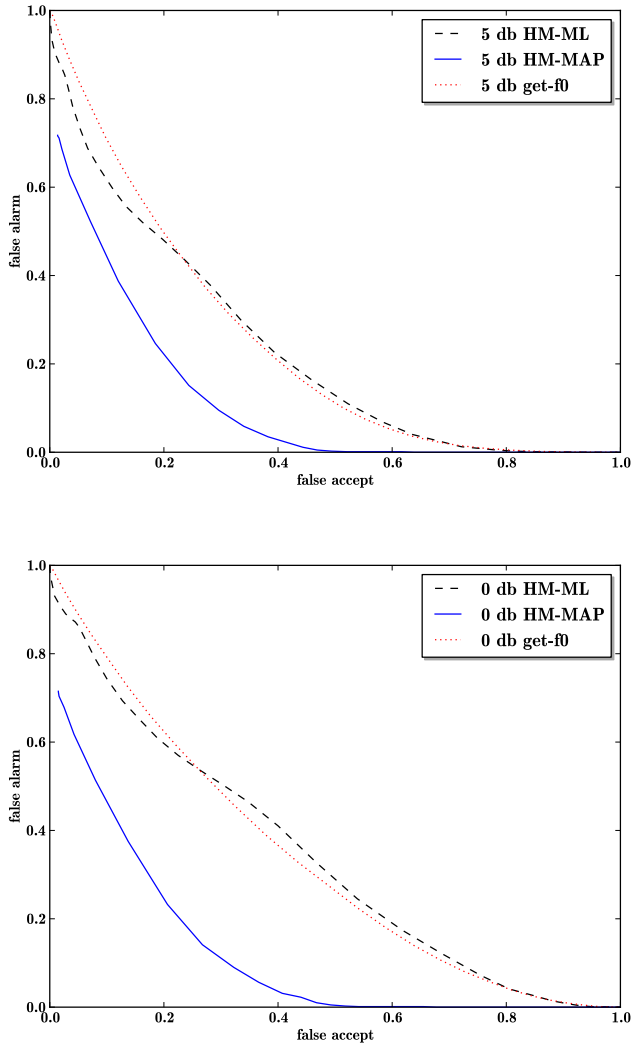

Figure 2.5: DET curve for detecting voiced frames with *get-f0*, ML and MAP versions of harmonic model in the Keele dataset under noisy conditions at 0, and 5 db SNRs

assistants. We created a corpus of non-speech and mostly speech utterances. Utterances with no reference transcripts were treated as non-speech. We identified utterances with very few silences by estimating the distribution of number of words in the utterances, shown in Figure 2.6, and picking the top (more than 18 words) quintile. This gave us an evaluation set of 4620 recordings containing no words and 1106 utterances containing mostly speech. The performance of MAP-version of harmonic models, the *get-f0* and

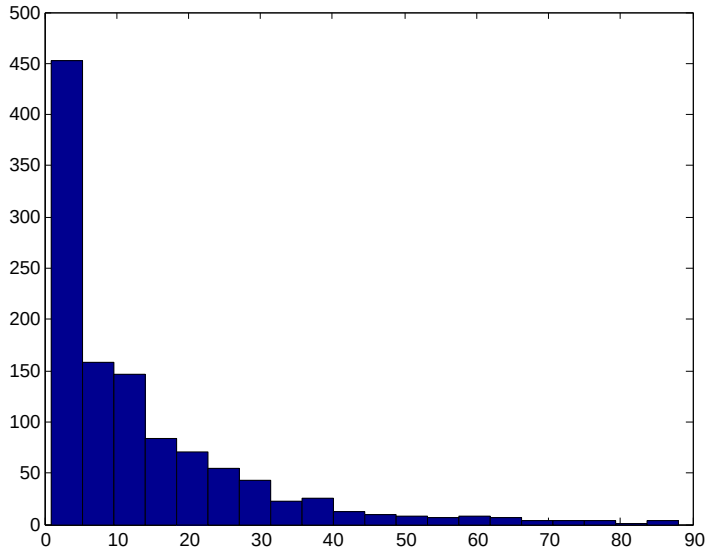

Figure 2.6: Distribution of number of words per utterance (30-seconds clips) in the recordings of everyday conversations.

the *openSMILE* are shown in the Figure 2.7. The total number of voiced and unvoiced frames were computed in each utterance and they were classified as speech or non-speech utterances using a threshold. By varying the threshold, we obtained the DET curve. The results show that the MAP-version of our algorithm outperforms both the baseline methods by a substantial margin.

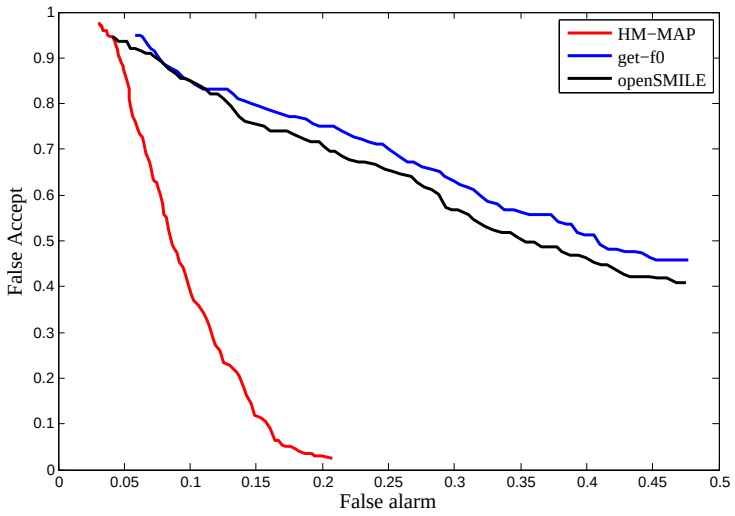

Figure 2.7: Comparing performance on detecting utterance with only noise (or speech) using MAP-version of harmonic model and two popular tools, *get-f0* and *openSMILE*.

2.4 Pitch estimation

2.4.1 Related research

Fundamental frequency, also referred as pitch period, is a key feature in speech analysis. Due to importance of robust pitch estimation on speech-related applications, it has been an interesting topic for many years. There are a variety of pitch detection algorithms in the literature, which generally consist of two stages. First, several pitch candidates are generated at each frame using a function that measures the self-similarity, such as autocorrelation function. Next, a dynamic programming algorithm, such as Viterbi algorithm, estimates the most probable pitch trajectory across all candidates after applying smoothing constraints.

Various methods in time, frequency and cepstral domain have been proposed for generating pitch candidates. Since the literature on this topic is extensive, we limit our brief overview to popular or recent algorithms that are directly relevant to this work and the final empirical evaluations. *Praat* obtains candidates from local peaks in either

autocorrelation or normalized cross-correlation function [51]. YIN uses a autocorrelation-based squared difference function followed by post-processing techniques to calculate candidates [52]. Such methods have been used in standard pitch detector tools such as *WaveSurfer* [53] and *Snack* [54].

Analogous to convolution in time domain, methods in frequency domain locate peaks in power spectrum. Hermes [55] proposed an algorithm that estimates the f_0 by seeking the frequency that maximizes the summation of harmonics on the logarithmic power spectrum as follow:

$$\hat{f}_0 = \arg\max_f \sum_{k=1}^{n} \log|X(k \cdot f)| \tag{2.19}$$

where $X(\cdot)$ denotes the power spectrum and n is the maximum number of harmonic peaks considered in the power spectrum. This method, however, ignores the information present in frequencies that are not harmonically related. To addresses this drawback, Sun [56] proposed the subharmonic to harmonic ratio (SHR) algorithm, where the height of the peaks with respect to the valleys are considered as follow.

$$\hat{f}_0 = \arg\max_f \frac{\sum_{k=1}^{n} \log|X(k \cdot f)|}{\sum_{k=1}^{n} \log|X((k-\frac{1}{2}) \cdot f)|} \tag{2.20}$$

Unlike proposed methods by Hermes and Sun where power spectrum of speech signal is inspected, Drugman et al. [57] focus on spectrum of the residual signal, $e(t)$. This method first estimates the residual signal by inverse filtering of the spectral envelop obtained by an auto-regressive framework. Next, from the spectrum of residual signal, summation of residual harmonics (SHR) is evaluated as follow:

$$\begin{aligned} SHR(f) &= E(f) + \sum_{k=2}^{n} [E(k \cdot f) - E((k-\frac{1}{2}) \cdot f) \\ \hat{f}_0 &= \arg\max_f \; SHR(f) \end{aligned} \tag{2.21}$$

where $E(\cdot)$ denotes the amplitude spectrum of residual signal, $e(t)$. Finally, f_0 is estimated by seeking the frequency that maximizes the SHR.

Recently, Kawahara proposed a time-frequency method called TANDEM-STRAIGHT for voice analysis and pitch extraction [58]. It first employs a power spectrum estimation

method called TANDEM that adaptively represents the signal and eliminates periodic temporal fluctuations. Then, pitch frequency is calculated using a fixed-point algorithm called STRAIGHT. Their time-frequency algorithm is computationally expensive.

Generally speaking, the above mentioned algorithms excel when the background noise is minimal or for specific types of background noise and their performance drop at low signal-to-noise ratios (SNRs). However, pitch estimators are widely employed in diverse applications which can benefit from better accuracy and better robustness.

Recently, Tabrikian and his colleagues [25] integrated a Harmonic model with a MAP framework to robustly estimate pitch period at low SNRs. However, the straightforward application of harmonic model leads to common errors in pitch detection, pitch halving and doubling errors. We propose a method in continue to mitigate these error while choosing the candidates per frame.

2.4.2 Segmental pitch tracking using harmonic model

As we showed earlier at Equation 2.10, pitch frequency can be estimated by maximizing the energy of the reconstructed signal. In the other hand, the rate of pitch variation is inherently limited by the motion of the articulators in the mouth during speech production [59] and hence, it cannot vary arbitrarily between adjacent frames. This smoothness constraint can be enforced using a first order Markov dependency between pitch estimates of successive frames. Adopting the popular hidden Markov model framework, the estimation of pitch over utterances can be formulated as follows. Let $\mathbf{Y} = \{\mathbf{y}_0, \ldots, \mathbf{y}_M\}$, and $\mathbf{F}_0 = \{f_0^{(0)}, \ldots, f_0^{(M)}\}$ be M length sequences of observed frames and candidate pitch estimates, respectively. The observation probabilities are assumed to be independent given the hidden states or candidate pitch frequencies here and expressed as follow.

$$p(\mathbf{Y}|\mathbf{F}_0) \quad = \quad \prod_{i=1}^{M} p(\mathbf{y}_i|f_0^i) \tag{2.22}$$

Then, following the Bayes rule for driving the posterior probability, the *maximum a posteriori* (MAP) estimation of F_0 is obtained as follow.

$$\hat{F}_0 \quad = \quad \arg\max_{\mathbf{F}_0}\{p(\mathbf{Y}|\mathbf{F}_0)p(\mathbf{F}_0)\} \tag{2.23}$$

A zero-mean Gaussian distribution with the defined over the pitch difference between frames is a reasonable approximation for the first order Markov transition probabilities [25].

$$\begin{aligned} p(\mathbf{F}_0) &= p({f_0}^{(1)}, {f_0}^{(2)}, \cdots, {f_0}^{(M)}) \\ &= p({f_0}^{(1)}) \prod_{i=2}^{M} p({f_0}^{(i)} | {f_0}^{(i-1)}) \\ p({f_0}^{(i)} | {f_0}^{(i-1)}) &\sim \mathcal{N}(f_0^{(i)}; {f_0}^{(i-1)}, \sigma_t^2) \end{aligned} \tag{2.24}$$

where $p({f_0}^{(1)})$ is the prior probability of f_0 at the first frame, and σ_t^2 is the variation of pitch transition function.

Putting all this together and substituting the likelihood from the Equation 2.10, the pitch over an utterance can be estimated as follows.

$$\hat{\mathbf{F}}_0 = \arg\max_{\mathbf{F}_0} \left[\sum_{i=0}^{M} \hat{\mathbf{s}_\mathbf{i}}^T \hat{\mathbf{s}_\mathbf{i}} | f_0^{(i)} + \log \mathcal{N}(f_0^{(i)}; f_0^{(i-1)}, \sigma_t^2) \right] \tag{2.25}$$

Thus, the estimation of pitch over an utterance can be cast as an HMM decoding problem where its states represent the possible discrete values of f_0, and can be efficiently solved using the Viterbi algorithm.

2.4.3 Pitch halving and doubling

The most common errors in pitch detection algorithms are pitch halving and doubling mostly known as gross pitch error (GPE). They often occur due to the strong subharmonics located in the range of pitch frequency; in time domain, their counterpart are the alternating cycles appear in both amplitude and period of the speech signal. These alternate can be found in either disordered or noisy voices. Like in other pitch detection algorithms, harmonic models suffer from pitch halving and doubling too. The harmonics of $f_0/2$ (halving) include all the harmonics of f_0. Similarly, the harmonics of $2f_0$ (doubling) are also the harmonics of f_0. The true pitch, f_0, may be confused with $f_0/2$ and $2f_0$ depending on the number of harmonics considered and the noise.

In many conventional algorithms, the errors due to halving and doubling are minimized by heuristics such as limiting the range of allowable f_0 over a segment or an utterance. This

requires prior knowledge about the gender and age of the speakers. Alternatives include median filtering and constraints in Viterbi search [54], which remain unsatisfactory.

We propose a method to capture the probability mass in the neighborhood of the candidate pitch frequency. The likelihood of the observed frames falls more rapidly near candidates at halving $f_0/2$ and doubling $2f_0$ than at the true pitch frequency, f_0. This probability mass in the neighborhood can be captured by convolving the likelihood function with an appropriate window.

Figure 2.8 illustrates the problem of pitch halving and demonstrates our solution for it. The top plot represents a voiced frame of clean speech contaminated with the babble noise at 10 db SNR, and in the bottom plot, the dotted line shows its estimated likelihood function. A maximum of this function will erroneously pick the candidate $f_0/2$ as the most likely pitch candidate for this frame. However, notice that the function has a broader peak at f_0 than at $f_0/2$. The solid line shows the result of convolving the likelihood function with a hamming window. In our experiments, we employed a hamming window with the length of $f_{0-min}/2$ where f_{0-min} is the minimum pitch frequency. The locally smoothed likelihood shows a relatively high peak at the true pitch frequency f_0 compared to $f_0/2$, thus overcoming the problem of pitch halving.

2.4.4 Evaluation

We evaluate the performance of our proposed method on a task of estimating pitch frequency on the Keele dataset with the similar experimental paradigm described earlier at voicing evaluation.

We assessed performance of the pitch trackers using following measures [57]: gross pitch error (GPE), which is defined as the percentage of f_0 estimates that deviate more than 20% of the ground truth; and the fine pitch error (FPE) that is the mean absolute error computed for estimates that are bellow than 20% of reference f_0.

We compared the performance of our proposed method with the following pitch estimation methods – (a) **STRAIGHT-TANDEM** (S-T), based on the fixed-point analysis

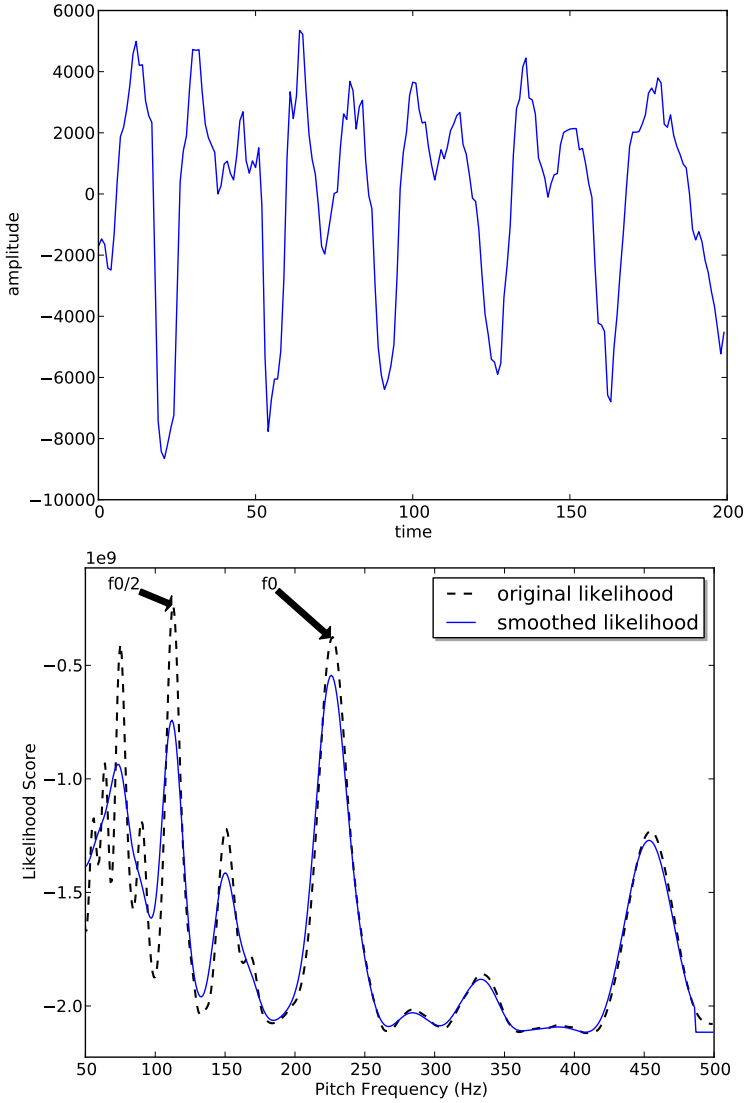

Figure 2.8: The top plot is a voiced frame of speech contaminated with the *babble* noise at 10 db SNR. The solid and dash lines in the bottom plot illustrate original and smoothed likelihood functions, respectively. The likelihood function has maxima near f_0 and $f_0/2$. Smoothing the likelihood locally solves pitch halving problem.

on modified power-spectrum [58]; (b) **YIN**, a template matching method with the autocorrelation function in time-frequency domain and *ad hoc* post-processing [52]; and (c) **SHR**, a method based on Subharmonics to Harmonic Ratio [56]. In all cases, the search for optimal pitch frequency was performed over a range from 50 Hz to 500 Hz was and

	Pitch Detection Algorithms				
	SHR	S-T	YIN	HM	HM-SL
Clean	4.18	2.05	3.19	4.81	**2.04**
Restaurant	15.61	12.03	14.76	19.08	**8.20**
Subway	13.79	9.72	12.73	15.15	**5.49**
White	12.84	9.17	10.17	**3.45**	4.76
Car	15.61	21.58	14.05	15.83	**6.82**
Street	16.66	20.52	16.25	20.36	**9.77**
Exhibition	14.03	7.40	12.20	10.17	**4.52**
Babble	16.11	13.37	14.97	19.31	**8.23**
Airport	16.65	19.73	15.31	18.32	**8.16**

Table 2.1: Comparison of the proposed method with other popular methods in terms of gross pitch error (GPE) under clean and different noisy conditions. In noisy conditions, the table reports the average over all SNRs ranging from 0 db to 20 db.

the frame rate was fixed to 100 frames per second. Note that the performance of pitch trackers are compared using only the frames corresponding to the reference voiced frames.

Results

In noisy conditions, Table 2.1 reports the average errors over all SNR bins ranging from 0 db to 20 db. On both clean and noisy speech, harmonic model with smoothed likelihood (HM-SL) clearly outperforms all other approaches in terms of GPE except in *white* noisy condition where the smoothing appears unnecessary and the HM outperforms all.

In Table 2.2 HM approach outperforms others except for SHR in *restaurant* noisy condition. As it is clear in Table 2.2, HM-SL has a comparable performance to the HM. This may be explained by the fact that smoothing of the likelihood score may reduce the precision of the harmonics.

	Pitch Detection Algorithms				
	SHR	S-T	YIN	HM	HM-SL
Clean	2.23	2.72	2.62	**2.22**	2.62
Restaurant	**2.52**	5.61	4.07	3.63	3.87
Subway	4.02	5.18	3.62	**3.00**	3.27
White	3.60	4.49	3.37	**2.80**	3.04
Car	3.98	5.61	3.97	**3.31**	3.46
Street	3.56	5.81	3.78	**3.44**	3.98
Exhibition	3.87	4.87	3.68	**3.09**	3.30
Babble	3.96	5.86	4.10	**3.59**	3.86
Airport	4.06	5.75	4.11	**3.58**	3.78

Table 2.2: Comparison of the proposed method with other popular methods in terms of fine pitch error (FPE) under clean and different noisy conditions. In noisy conditions, the table reports the average over all SNRs ranging from 0 db to 20 db.

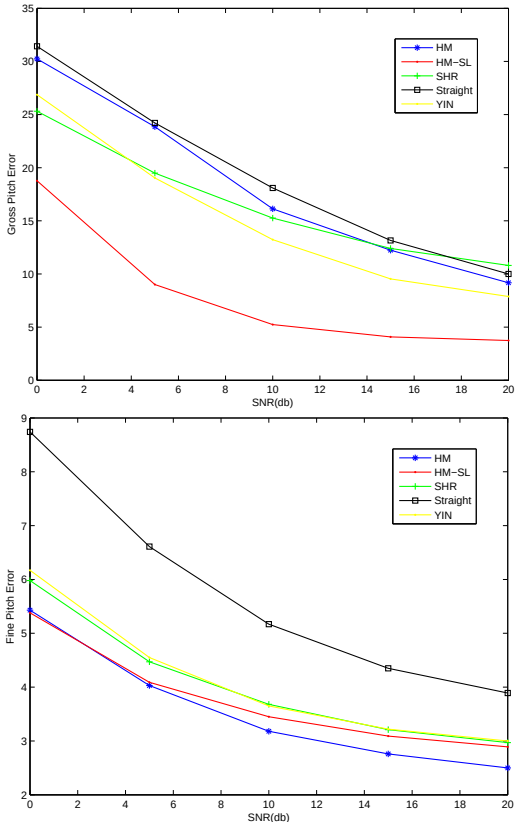

Figure 2.9: Gross pitch error (top) and fine pitch error (bottom) for all methods and averaged over all 8 noisy conditions.

The Figure 2.9 summarizes the overall gross pitch error and fine pitch error across all the noise conditions. The proposed model (HM-SL) substantially outperforms the other popular pitch trackers in gross pitch error under all levels of noise conditions. The performance HM-SL is also better than HM, which shows that the smoothing contributes to the performance gains. The model also performs better in fine pitch error when the noise level is high. At low noise levels, the proposed model degrades fine pitch estimate, which is not entirely surprising and is due to the smoothing. In fact, at high SNRs standard HM is sufficient and smoothing is not necessary.

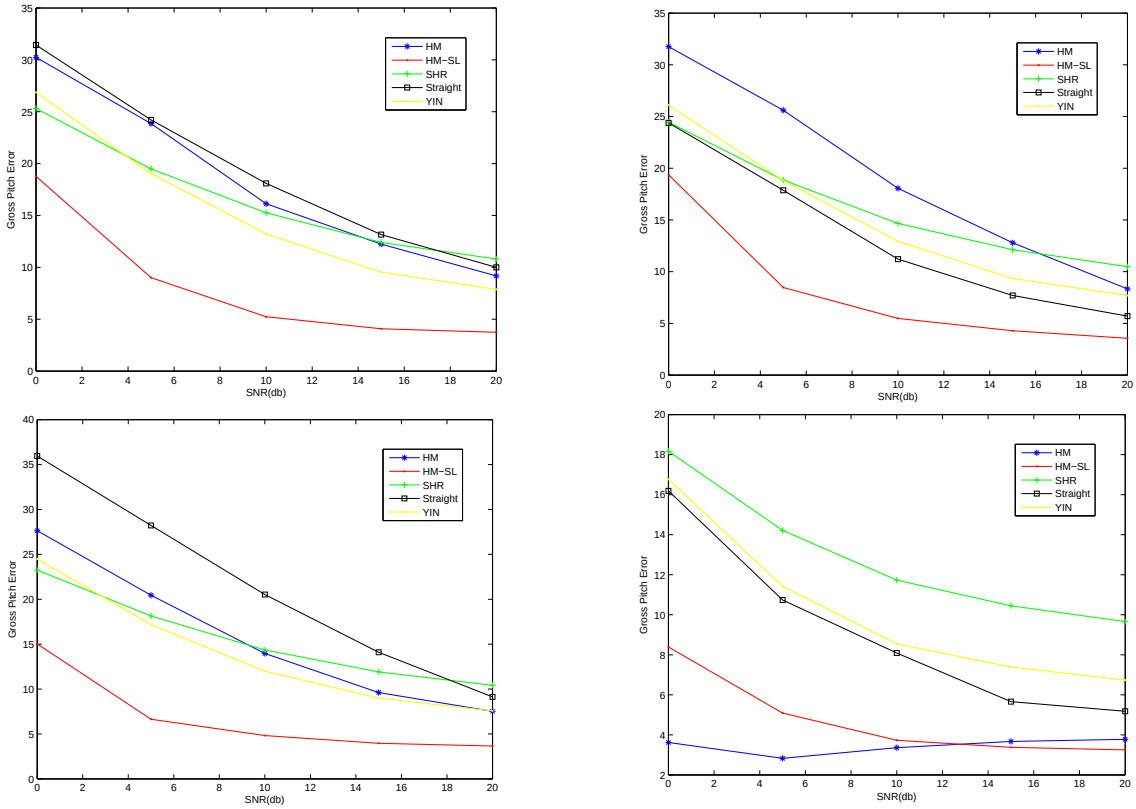

Figure 2.10: Comparison of the proposed model with other popular methods in terms of gross pitch error under four representative noise conditions – air (top,left), babble (top,right), car (bottom, left) and white (bottom,right).

2.4.5 Model sensitivity

In order to measure the sensitivity of proposed pitch tracker to variation of the model order, we evaluated the overall GPE under both clean and noisy conditions as a function of number of harmonics, H. For noisy condition, we averaged the GPE over over all noise types. Figure 2.11 indicates that under clean condition, the model is robust to model order variations. However, as SNR increases, GPE varies as a function of model order. We computed average frame-level BIC for a several number of harmonics, ranging from $H = 2, \cdots, H = 9$, on a sub-sample of data. In this experiment, ten percent of recordings (noisy and clean) was randomly selected and seven harmonics appeared to be an optimal value across clean and noisy conditions at different SNRs.

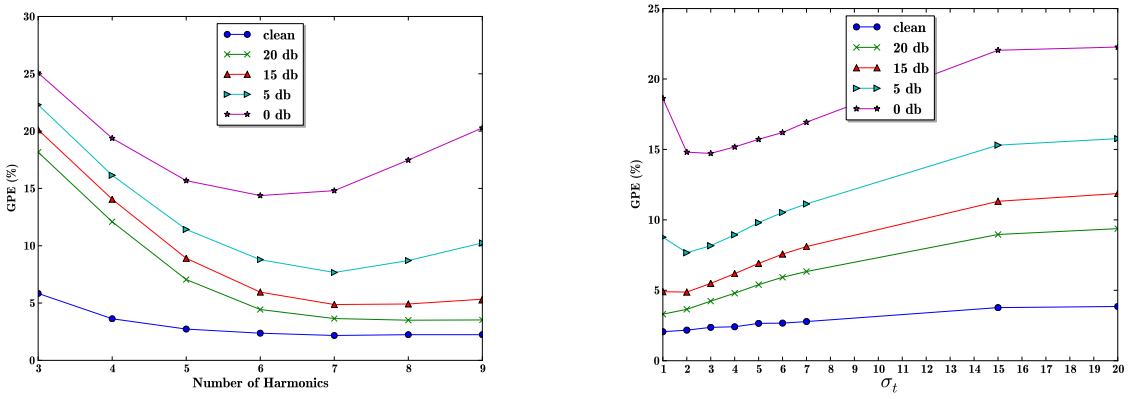

Figure 2.11: Comparison of overall gross pitch error (averaged over all noise types) as a function of number of harmonics (left) and as a function of variance of transition probability (right).

The same experiment were performed to evaluate the model sensitivity to variation of transition function in adjacent frames. The results, illustrated in Figure 2.11, shows that the optimal variance in transition function is irrelevant of SNR to some extent and appears to be 2 Hz in this case.

2.5 Quantifying voice quality

Speech pathologists often describe voice quality in voice disorders as breathy or harsh, which has been largely attributed to incomplete closure of vocal folds. Traditional assessment of voice quality relied on perceptual measurements of perceived abnormalities in patients voice. For several years now, researchers have been attempting to quantify voice quality more subjectively employing acoustic measures that reflect inherent abnormalities of disordered voices. Jitter, shimmer, and HNR are important cues with regard to voice quality assessment. As the degree of perceived hoarseness increases, more noise appears to replace the harmonic structure and as a result HNR decreases. Irregular vocal fold vibration causes random modulation of the source signal, in both amplitude (shimmer) and time period (jitter).

2.5.1 Jitter and shimmer

Related research

Jitter and shimmer are the prominent acoustic measures that are often used in the context of voice quality assessment. Small cycle-to-cycle fluctuations in glottal pitch period and amplitude are defined jitter and shimmer respectively. They may occur during voice production and cause voice roughness, especially in pathological voices [60]. Perturbation analysis is based on the fact that small fluctuations in frequency, and amplitude of waveform reflect the inherent noise of voice. Speech pathologist often measure shimmer and jitter to characterize abnormalities in voice production [61]. Speech pathologist often measure shimmer and jitter to characterize abnormalities in voice production [61]. A number of methods in time and frequency domain have been proposed for the computation of jitter and shimmer [62, 63]. In the frequency domain, Vasilakis and Stylianou [64] proposed a mathematical model for estimation of jitter. Assuming that the magnitude of spectrum can be separated into a harmonic part and a sub-harmonic part, they showed that jitter can be estimated by counting the number of intersections between harmonic and sub-harmonic spectra. Time domain [51, 65, 66] methods often employ relative pitch (T_i) and amplitude (A_i) differences between consecutive pitch periods for estimating jitter

and shimmer as expressed in following equations, respectively:

$$J = \frac{1/(N-1)\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{1/N\sum_{i=1}^{N} T_i} \tag{2.26}$$

$$S = \frac{1/(N-1)\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{1/N\sum_{i=1}^{N} A_i} \tag{2.27}$$

where N denotes the number of cycles in the analysis frame. Automatic measurement of time period and amplitude of each cycle is prone to errors due to noise. Most automated methods sidestep this problem by measuring the variation across frames using the average time period and amplitude per frame [67, 68, 69]. We alleviate this problem by estimating shimmer and jitter from the reconstructed signal using the estimated parameters of the harmonic model [14]. This is less sensitive to noise and can also capture the variation within a frame as explained in the following subsection.

2.5.2 Estimating shimmer and jitter using harmonic models

Shimmer

Shimmer is defined as the variation in amplitude between the adjacent cycles of the glottal waveform. It can be viewed as a slow amplitude modulation (AM) of glottal waveform due to the inability of humans to keep constant the tension of their vocal folds [70]. Our approach for estimating shimmer is based on the fact that speech signal can be formed as a combination of two AM and frequency modulated (FM) components. The key idea here is to reconstruct two versions of the input waveform in each frame; a version where the amplitudes of the harmonics are assumed to be constant, as in Equation 2.1, and another without that assumption. Both reconstructions are estimated to minimize the effect of noise. The model for the voiced speech that allows harmonic amplitude to vary with time can be represented as follow [24].

$$\begin{aligned} s_{tv}(t) &= a_0 + \sum_{h=1}^{H}[a_h(t)\cos(2\pi f_0 h t)] \\ &+ \sum_{h=1}^{H}[b_h(t)\sin(2\pi f_0 h t)] \end{aligned} \tag{2.28}$$

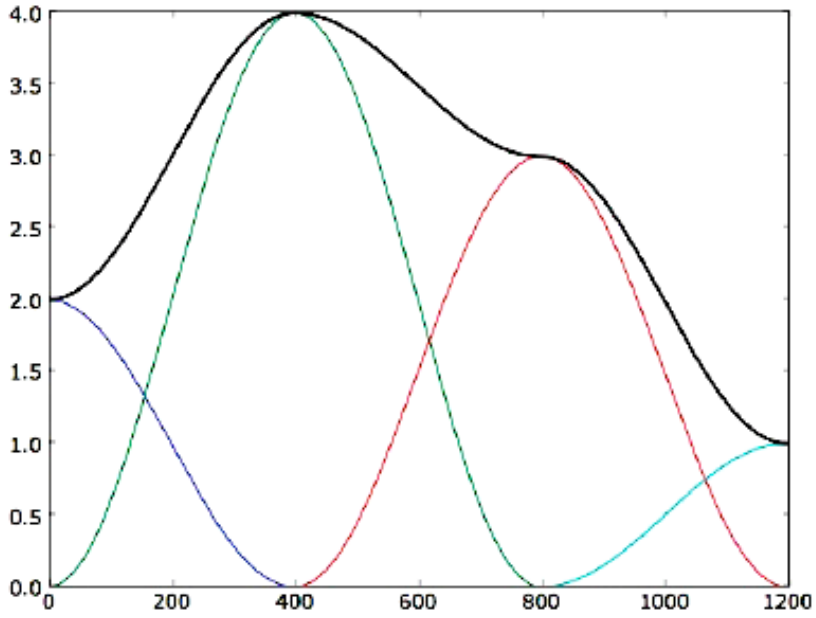

Figure 2.12: An illustration of time-varying amplitude of a harmonic component modeled as a superposition of four bases functions spanning the duration of the frame.

Note, this is different from the harmonic model represented previously in Equation 2.1. Unlike the previous model whose harmonic coefficients are fixed, in the time-varying harmonic model(TV-HM), as the name implies, the coefficients are allowed to vary $a_h(t)$ and $b_h(t)$ over time. Thus, this model is capable of capturing sample to sample variation in harmonic amplitude within a frame [24]. It is reasonable to assume that the sample to sample variation is smooth. We represent the amplitudes of the individual harmonics as a linear combination of a few local basis functions [24], as formulated in Equation 2.30 and shown in Figure 2.12.

$$\begin{aligned} a_h(t) &= \sum_{i=1}^{I} \alpha_{i,h}\psi_i(t) \\ b_h(t) &= \sum_{i=1}^{I} \beta_{i,h}\psi_i(t) \end{aligned} \tag{2.29}$$

The basis functions could be any convenient functions with a limited support. We represent this smoothness constraints within a frame using four ($I = 4$) Hanning windows

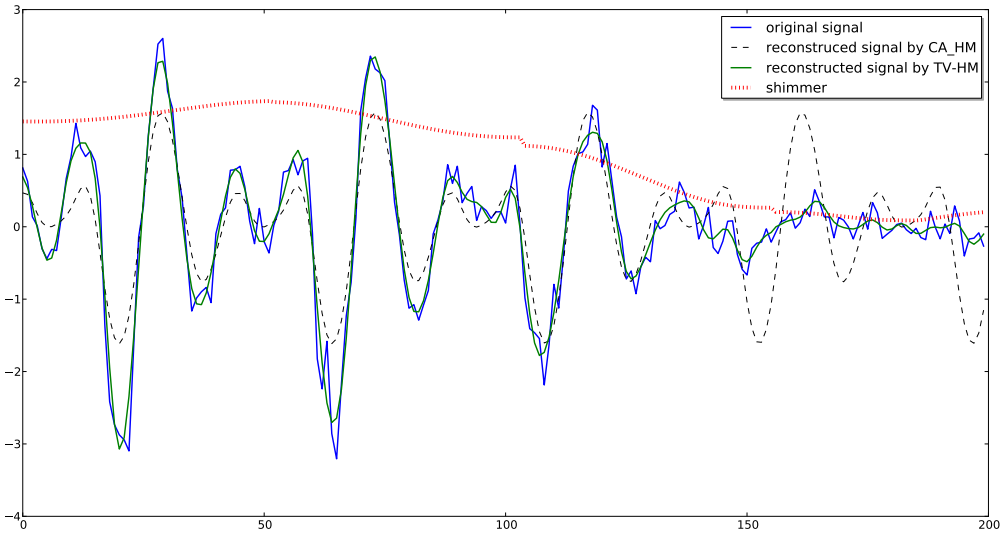

Figure 2.13: An example speech frame (blue), estimated signal from harmonic model with time-varying amplitude (green), estimated signal from harmonic model with constant amplitude (black), and estimated shimmer (red).

as basis functions. For a frame of length M, the windows are centered at 0, $M/3$, $2M/3$, and M. Each basis function is $2M/3$ samples long and has an overlap of $M/3$ with immediate adjacent window. The parameters of this model can be expressed, once again, as a linear model, similar to Equation 2.3, but this time the A and m have dimensions four times the original dimensions. Given the fundamental frequency from Equation 2.25, we compute $a_h(t)$ and $b_h(t)$ using a maximum likelihood framework.

Figure 2.13 illustrates an example frame, the signal estimated using the harmonic model with constant amplitude and with time-varying amplitudes. The signal estimated with the time-varying harmonic amplitudes is able to follow variations not only in amplitude but also variation in pitch to a certain extent.

Now, the cycle-to-cycle variation associated with shimmer can be computed from the estimated parameters of the two models, with constant amplitudes, a_h and b_h, and the time-varying amplitudes, $a_h(t)$ and $b_h(t)$, of the harmonics. Shimmer can be considered as a function $f(t)$ that scales the amplitudes of all the harmonics in the time-varying model.

$$c_h(t) = c_h f(t) + e(t), \quad t = 1, \ldots, T \ , h = 1, \ldots, H \tag{2.30}$$

where $c_h = \sqrt{\sum_{h=1}^{H}(a_h^2 + b_h^2)}$ denotes the amplitude of the harmonic components in harmonic model with constant amplitudes and $c_h(t)$ is the counterpart from the time-varying model. Once again, assuming uncorrelated noise, $f(t)$ can be estimated using maximum likelihood criterion.

$$\hat{f}(t) = \frac{\sum_{h=1}^{H} c_h c_h(t)}{\sum_{h=1}^{H} c_h^2} \tag{2.31}$$

Figure 2.13 illustrates an example frame where the dotted red line shows the extracted AM component (the envelop of speech waveform). The larger the tremor in voice, the larger the variation in $\hat{f}(t)$. Hence, we use the standard deviation of $\hat{f}(t)$ as a summary statistics for shimmer to quantify the severity of tremor.

Jitter

Jitter is the counterpart of shimmer in time period, i.e., the cycle-to-cycle variation in pitch period. It effects the spectrum of a sustained vowel by reducing the amplitudes of harmonics and adding noise between harmonics [71]. Analysis of jitter is based on the accurate estimation of pitch period. Given an estimate pitch period of the frame, $T_0 = 1/f_0$, we first create a matched filter by excising a one pitch period long segment from the signal estimated with the harmonic model from the center of the frame as follow:

$$h(t) \quad = \quad \begin{cases} s(t) & \text{if} \quad t_{N/2-T_0/2} < t \leq t_{N/2+T_0/2} \\ 0 & \text{otherwise} \end{cases} \tag{2.32}$$

where $s(t)$ denotes the N speech samples in a reconstructed voiced frame using harmonic model, measured at times $t_1, t_2, \ldots, t_N$. This matched filter is then convolved with the estimated signal from time-varying harmonic model as follow:

$$\begin{aligned} \Delta &= \{\text{local maxima} \ \ h(t) * s_{tv}(t)\} \\ &= \{\delta_1, \ldots, \delta_m\} \end{aligned} \tag{2.33}$$

where $*$ denotes the convolution operator and m is the number of local maxima. The distance between the maxima defines the pitch periods in the frame. The perturbation in

period is normalized with respect to the given pitch period and its standard deviation is an estimate of jitter as follow.

$$p_j = \frac{\delta_i - \delta_{i+1}}{T_0} \quad , \quad j = 1, \ldots, m-1 \tag{2.34}$$

$$\text{Jitter} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m-1}(p_j - \mu)^2}, \qquad \text{where} \qquad \mu = \frac{1}{m-1}\sum_{j=1}^{m-1} p_j \tag{2.35}$$

2.5.3 Harmonic-to-noise ratio

Related research

An accurate estimate of harmonic-to-noise ratio (HNR) provides useful information about the amount of aperiodicity in the speech signal. Acoustic properties of the speech signal such as period-to-period frequency perturbation, amplitude variation, voiced consonants, stops, fricatives, and aspiration noise are the sources of speech aperiodicity. Researchers have used the HNR in the acoustic studies for the evaluation and management of voice disorders. HNR seems to be the most applicable measure at clinics as a quantitative index to measure the degree of hoarseness. Hoarseness is an important symptom of most laryngeal disorders and speech pathologists rate the degree of hoarseness to assess voice disorders [72]. Generally, we expect the lower HNR in disordered voices rather than healthy voices [73]. A variety of HNR estimation methods in the studies can be classified into two types: time-domain methods, in which HNR is directly computed from the speech waveform, and frequency-domain methods, in which HNR is computed from the transformed version of speech waveform.

A representative time domain approach for measuring the HNR was introduced by Yumoto and his colleagues [72]. They assume that voiced speech is a sum of two parts: a periodic component, and an additive noise component. To estimate the HNR, they first compute an *average waveform* for a single period by calculating the mean of successive periods. The energy of this *average waveform* defines the harmonic energy. Assuming the noise is a stationary process across the frame, noise energy is then calculated using the mean squared difference between the *average waveform* and the individual periods. However, because of the cycle-to-cycle pitch period perturbations, the periods are not

necessarily aligned. Therefore, zero padding is used for time-normalization of the periods prior to computation of the mean and variance. However, this simple time-normalization technique significantly amplifies the computed noise energy when the speech signal has pitch period perturbations in its waveform, such as in disordered voices.

To overcome these limitations, Qi [74] proposed a time-normalization process using dynamic time warping (DTW) aiming to minimize the effects of f_0 perturbations. DTW is a non-linear time-normalization method that minimizes the mismatch between the two input frames. It optimally aligns speech waveforms prior to computation of the HNR. However, this time domain method for computing the HNR requires an accurate estimation of pitch period. Further, pitch boundaries are very sensitive to the phase distortion and cause inaccurate HNR estimation. Qi and his colleagues later [75] proposed another appropriate time-normalization technique using zero-phase transformation to minimize the influence of shimmer and jitter on the computation of the HNR.

Recently, Boersma [76] proposed a straightforward time domain technique based on short-term autocorrelation function (ACF). The key idea of this method is defining the degree of periodicity as the relative value of ACF at lag T_0 $(1/f_0)$ to the power of non-harmonic component. This allows defining the logarithmic HNR for a voiced frame as:

$$HNR = 10 \log \frac{ACF(T_0)}{ACF(0) - ACF(T_0)} \tag{2.36}$$

where, $ACF(0)$ denotes the power of speech waveform. In this equation, the power of harmonic component is calculated based on the value of ACF at lag T_0 and subsequently, the power of non-harmonic (noise) components is obtained by subtracting the the power of harmonic part from the power of signal.

A number of techniques have been proposed for computing the HNR in the frequency domain. The main advantage of those methods is less dependency on the accurate estimate of pitch period [77]. Krom [78] proposed a technique, in which the harmonic and noise components are discriminated in the cepstrum domain using a comb-liftering operation. However, cepstral analysis assumes that the process is stationary across the frame and waveform variations may leads to spectral leakage leading to the attenuation in magnitude of harmonics.

2.5.4 Estimating harmonic-to-noise ratio using harmonic model

In this section, we express a time domain approach given the parameters of a harmonic model. This method focuses on decomposition of voiced speech into a periodic and a non-periodic component generated from glottal excitation and random noise sources, respectively. Individuals with voice disorders usually cannot seamlessly switch between the two sources of voice production and therefore, excitation pulses are contaminated by the noise signal. As such, the goal of our approach is to separate the contribution of the two sources in order to quantify the degradation in voice quality. This method assumes that a harmonic model approximates the harmonic part of voiced speech and the non-harmonic part is obtained by subtracting the harmonic part from the original speech signal. The noisy part encompasses everything in the signal that is not described by harmonic components including the friction noise, the waveform fluctuations, and etc. One drawback of HM with constant amplitudes is its limitation to follow within the frame amplitude and frequency variations. In order to overcome this issue, we employ a HM with time-varying amplitudes providing more flexibility in capturing sample to sample variations across the frame.

HNR and the ratio of energy in first and second harmonics (H12) can be computed from the HM-VA as follow.

$$\begin{aligned} c_h(t) &= \sqrt{\sum_{i=1}^{I} a_h(t)^2 + b_h(t)^2} \\ HNR &= \log \sum_{t=1}^{N} \sum_{h=1}^{H} c_h(t)^2 - \log \sum_{t=1}^{N} (y(t) - s(t))^2 \\ H12 &= \log \sum_{t=1}^{N} c_1(t)^2 - \log \sum_{t=1}^{N} c_2(t)^2 \end{aligned} \tag{2.37}$$

Thus, we compute jitter, shimmer, harmonic-to-noise ratio and the ratio of energy in first and second harmonics using reconstructed signal that is less prone to noise related errors. The effectiveness of these two measures are evaluated in the experiments in the next chapter.

2.5.5 Evaluation

Straightforward comparison of proposed quantities with other algorithms employed in popular tools is difficult due to the lack of *gold standard*. We prefer to verify the accuracy of developed measures on a task of classifying subjects with voice impairments due to Parkinson's disease (PD) from normal subjects. Other researchers have also relied on such tasks to verify their proposed algorithms. For instance, Vasilakis et al. [79] evaluated their proposed method of estimating jitter with that of implemented on *Praat* [51] on the task of detecting sustained phonation recordings of disordered voices.

Corpus

We conducted our empirical evaluation on a sub-sample of data collected from 189 subjects, 168 of them were diagnosed with PD, and 21 controls. Subjects were asked to perform several speech tasks designed to exercise different aspects of speech and non-speech motor control. In order to examine voice quality measures, we only focused on sustained phonation task, in which subjects were instructed to phonate the vowel /a/ in a clear and steady voice as long as possible.

As a clinical reference, the severity of subjects' condition were measured by clinicians using the Unified Parkinson's Disease Rating Scale (UPDRS), the current gold standard. We focus on the motor sub-scale of the UPDRS (mUPDRS), which spans from 0 for healthy individual to 108 for extreme disability. The severity of the diseases ranged from 0 (control) to 55 on the UPDRS scale in our subjects. We will describe data collocation in more detail in Section 3.1.3 of Chapter 3. For the purpose of verifying developed voice quality measures, we created a sub-sample of subjects diagnosed with PD by estimating the distribution of subjects' mUPDRS, and picking the top 25%-ile. This gave us an evaluation set including 46 subjects with mUPDRS more than 28, with a mean of 38.2, and 21 control subjects.

Speech features

We extracted jitter, shimmer, and HNR employing both our proposed method and *Praat*, a standard feature extraction tool. *Praat* follows Equations 2.26, 2.27, and 2.36 for computing jitter, shimmer, and HNR, respectively. We then summarized frame-level features into a global feature vector of fixed dimension for each subject. Each feature was summarized across all frames in terms of standard distribution statistics such as, mean, median, variance, minimum, and maximum.

Classification model

We classify subjects with PD from controls using voice quality features extracted from both HM and *Praat* and report the performance in terms of *classification accuracy*. We used support vector machine (SVM) with several kernel functions including polynomial, radial basis function (RBF), and linear kernels employed from open-source Scikit-learn toolkit [80]. We used a twenty-fold cross validation scheme, setting all model parameter using four of the five sets as training set, and using the tenth ones only for reporting the performance estimates below. Parameters of the optimal SVM model were determined on the training set separately for each fold, via grid search and cross-validation.

Results

Table 2.3 indicates that voice quality features, extracted from both methods, can detect disordered voices belong to subjects with PD significantly better than chance with p-value of less than 0.001, according to cross-validated paired t-test [81]. Results show that HNR is more effective than shimmer and jitter in this task and features exacted from HM, except for jitter, perform better than features extracted by *Praat*.

2.6 Summary

In summary, this chapter describes a computational approach for quantifying perceptual voice qualities such as breathy and hoarseness. We focus on developing robust and accurate algorithms for estimating speech features. Starting with review of traditional acoustic

Model	Speech Features	*Praat*	HM
Chance	–	45.4	45.4
SVR	jitter	69.1	69.1
	shimmer	67.9	70.6
	HNR	70.56	72.4
	jitter + shimmer + HNR	71.6	73.5

Table 2.3: *Classification accuracy* measured on a 20-fold cross-validation for detecting subjects with PD from speech

feature extraction techniques, we illustrate a model-based approach based on a computational model of speech production. We develop a speech detection algorithm that exploits harmonic rich nature of voiced speech. For this purpose, we adopt the harmonic model of speech. We show how the parameters of the model can be computed for each frame using maximum *a posteriori* (MAP) estimate. In our MAP estimation, we modeled a prior distribution over coefficients of harmonic using a multivariate Gaussian distribution. Note, Bayesian estimate of all the model parameters is significantly more complex and requires expensive numerical approximations compared to our simpler MAP estimate with its closed form analytical solution [24]. Also, it is different from a related previous work that MAP smooths the likelihood using a first order HMM transition model [25]. We then use the likelihood of voicing under the harmonic model as the observation probability of our HMMs for detecting speech. We also overcome the weakness of harmonic model in differentiating speech from stationary harmonic noise using the non-stationary property of speech. Then, we address two outstanding problems related to harmonic models in the context of pitch estimation. Like other pitch estimation algorithms, the harmonic model suffers from pitch halving and doubling. We propose a local smoothing function that exploits the fact that there is more energy in the harmonics near the true pitch than at the corresponding neighborhoods of half or double the pitch. We utilize a local smoothing function to include this energy and improve the robustness of the pitch candidates in each frame. The harmonic model requires specification of the number of harmonics. Although the optimal choice depends on the noise conditions, we show how the local smoothing

can avoid per-frame model order optimization. We adopt a BIC criterion and define a model complexity that allows us to estimate the number of harmonics. We estimate the optimal number of harmonics using the average BIC per frame over a small subset of data. We then utilize the signal reconstructed from the harmonic model to robustly estimate harmonic-to-noise ratio (HNR), jitter and shimmer. We address the limitation of HM in capturing within the frame amplitude and frequency variations and employ a HM with time-varying amplitudes in order to tackle this issue. Finally, we evaluate the performance of our model in the context of voicing detection and pitch estimation on the Keele dataset under different noise types and levels. Taken together, improvements on HM provide substantial gain over other popular methods for all types and levels of additive noise.

Chapter 3

Acoustic Features in Clinical Speech Applications

In the previous chapter, we described a computational approach based on harmonic model of speech for reliably extracting acoustic features. In this chapter, we investigate the utility of these developed features in clinical speech applications. We employ supervised machine learning algorithms, such as support vector regression (SVR), and learn probabilistic speech models for automatically characterizing speech impairments in Parkinson's disease, autism spectrum disorder (ASD), and clinical depression. We briefly describe their symptoms on speech production system and previous works on speech-based assessment of these diseases. Through empirical experiments on a relatively large cohort of 168 subjects diagnosed with PD, we demonstrate that the severity of the disease can be predicted from extracted acoustic features well above chance. Then, we report our experiments on the *Interspeech 2013 Autism sub-challenge* where we verify the performance of our features in the task of detecting 35 children with ASD from 54 controls. We also classify the disorder into four categories. In addition to PD and ASD, we investigate the ability of machine learning and speech processing on the problem of detecting depressed adolescents from recordings collected during their family interactions. We employ both prosodic and context features, and learn probabilistic models for detecting 71 clinically depressed adolescents from 77 healthy controls. In the last section of this chapter, we investigate the problem of detecting social contexts from the audio recordings of everyday life such as in life-logs. In this task, we introduce a data learning framework employing deep neural networks (DNNs) and take potential advantages of available unlabeled data using sparse auto-encoders through an unsupervised learning scheme. We describe the data collection followed by a brief description of auto-encoders. We then propose a multi-label classification scheme and investigate different strategies for training this model. We report results of our both single-label and multi-label classifiers on the task of detecting social contexts such as speakers' location and activity. Finally, we conclude with summary of our key results.

3.1 Assessing the severity of Parkinson's disease

3.1.1 Introduction

Parkinson's disease (PD), is a progressive degenerative neurological disorder characterized by muscle rigidity, tremor, a loss and slowing of physical movement. A number of studies have shown a variety of symptoms in patient with PD affecting their quality of life. PD can affect all components of speech production including breathing, laryngeal function, and also their coordination for the production of smooth speech. Resulting dysarthric speech often exhibits monotonous pitch, slurring, reduced stress, inappropriate pauses, variable speech rate, short rushes of speech, harsh and breathy voices, and imprecise consonant production [2]. The severity of Parkinson's disease is typically assessed clinically using a widely accepted metric, the Unified Parkinson's Disease Rating Scale (UPDRS). The metric consists of clinician-scored motor evaluations and self evaluation of the activities of daily life. The UPDRS score ranges from 0 to 176, with 0 corresponding to a healthy state and 176 to a severe affliction; the range of target motor sub-scale is 0 to 108 [82]. The assessment is time-consuming and is performed by trained medical personnel, and can be burdensome in some situations, for example, when a patient must undergo frequent reassessments. Moreover, UPDRS is a subjective score and rater's bias can affect the assessment.

Since speech production involves complex motor coordination, the disease exhibits symptoms which are well-known to speech pathologists, although the exact pathological cause remains unclear. For several decades now, researchers have been interested in measuring these symptoms in speech more objectively with the hope of augmenting or simplifying the assessment. Speech tasks can be administered remotely, avoiding the need for driving to the clinic, which can be challenging for those with severe PD-related motor tremors. It is often cheaper and easier to automatically elicit, record, and analyze speech than conducting in-person clinical assessment.

There has been a growing interest in creating tools and methods for alternative home-based assessments of this disease. Easier methods of assessment could potentially be an important screening tool for a wider population as PD is the second most common

neurodegenerative disease in the United State after Alzheimer's disease. Since speech can be easily collected remotely across large distance using hand-held devices, it is an appealing source of evidence for telemonitoring PD. In this study, we investigate the accuracy of automatically inferring the severity of PD from speech samples in a relatively large cohort collected from multiple clinics. We extract a number of potential speech features using standard speech processing algorithms and apply several machine learning algorithms to predict the clinical ratings from the speech features.

3.1.2 Speech-based PD diagnosis

There have been extensive studies on employing speech processing and machine learning techniques for assessing voice disorders [83, 84, 85, 86] and in particular, classifying PD subjects or inferring the severity of the diseases [87, 14, 88]. Here, we describe a few studies. Guerra and Lovely attempted to create hand-crafted rules to mimic perceptual ratings [89]. Specifically, they created linear regression with automatically extracted measures such as harsh voice, breathy voice, and audible inspirations. Gil and his colleagues [90] proposed a hybrid classifier combining artificial neural network (ANN) and support vector machine (SVM) classifiers. Their experiments carried out using a range of speech measurements on a relatively small corpus of 31 subjects composed of 23 subjects diagnosed with PD and achieved a high accuracy of around 90%. Another study by Dus [91] attempted to compare different types of classification methods for diagnosis of PD on the same dataset. He employed different classifiers including neural network, regression, and decision tree where the neural network classifier yielded the best score at 92.9% classification accuracy. More recently, Bocklet et al. applied a more rigorous machine learning approach to classify PD subjects from control [92]. They extracted 292 prosodic features, adapted a 128 component Gaussian mixture model or universal background model using a *maximum a posteriori* criterion and found that they were able to perform the classification with good accuracy. However, their sample size contained only 46 Czech subjects of which 23 were diagnosed with PD. The severity of the disease in their subjects was relatively low, with a score of 17.5 on the UPDRS scale. Taken together, there has been continuous interest spanning several decades in characterizing the speech abnormalities in PD. However, most studies

were focused on measuring group differences of speech features or have been performed on small samples.

3.1.3 Corpus

Empirical evaluation reported in this study were performed on data collected from 168 subjects, all of whom were diagnosed with PD. Using a portable device, the data was collected through 3 clinics to alleviate potential bias due to clinic-specific practices [93]. Subjects were asked to perform several tasks designed to exercise different aspects of speech and non-speech motor control. The tasks were administered on a portable computer under the supervision of a clinician who was familiar with the computerized tests. As a clinical reference, the severity of subjects' condition were measured by clinicians using the Unified Parkinson's Disease Rating Scale (UPDRS), the current gold standard. In this study, we focus on the motor sub-scale of the UPDRS (mUPDRS), which spans from 0 for healthy individual to 108 for extreme disability. The severity of the diseases ranged from 0 (control) to 55 on the UPDRS scale in our subjects, with a mean of 22.9 and standard deviation of 9.3.

3.1.4 Speech elicitation tasks

Speech was elicited from subject's under 3 different conditions to obtain evidence of hypokinetic dysarthria.

1. Sustained phonation task: Subjects were instructed to phonate the vowel /ah/ in a clear and steady voice as long as possible. Speech pathologists rate voice quality during this task.

2. Diadochokinetic (DDK) task: DDK task is often used as a clinical test to assess the functional capacities of the articulatory system. Subjects were asked to repeat the sequence of syllables /pa/, /ta/ and /ka/ continuously for about 10 seconds as fast and as clearly as they possibly can. This task is often employed by speech pathologists to judge articulatory precision, control and speed.

3. Reading task: Subjects were asked to read three passages that are often employed in speech pathology and are referred to as The Rainbow Passage, The North Wind and The Sun, and The Grandfather Passage. Reading task imposes an additional cognitive processes during speech production and allows measurement of vocal intensity, voice quality, and speaking rate.

3.1.5 Experiments

Speech features

Criteria used by clinicians to rate hypokinetic dysarthria are often difficult to quantify. As mentioned earlier, we sidestep the difficult task of quantifying perceptual cues, and instead focus on extracting a large number of surface features. Classic perceptual characteristics associated with PD are reduced loudness; monotonous pitch; monotonous loudness; reduced stress; breathy, hoarse voice quality; imprecise articulation; and short rushes of speech. In general, we can divide the aforementioned problems into three major categories: loudness related problems, pitch related problems and articulatory related problem. In our experiment, we used a broad range of features to capture features associated with these categories.

- **Baseline features** : For our baseline system, we adopted the baseline features defined in INTERSPEECH 2010 Paralinguistic Challenge [94] using openSMILE toolkit [95]. The features, comprised of 1582 components, can be broadly categorized into three groups: 1) loudness related features such as RMS energy and PCM loudness, 2) voicing related features like pitch frequency, jitter, and shimmer., and 3) articulatory related features such as mel-frequency cepstral coefficients and line spectral frequencies. The above configuration provides 38 features along with their derivatives to form the frame-level acoustic features. The derivatives allows us to capture local dynamics of pitch and other features. The features computed at the frame-level were summarized into a global feature vector of fixed dimension for each recording using 21 standard statistical functions including min, max, mean, skewness, quartiles and percentile.

For our experiments, as in most speech processing systems, we extract 25 millisecond long frames using a Hanning window at a rate of 100 frames per second before computing the following features.

1. **Pitch frequency**: One of the key features in speech analysis is pitch frequency. Even though there are a large number of pitch estimators in the literature [54, 51, 52, 55, 56, 57, 58], most of them are not well-suited for measuring the small tremors, observed in Parkinson's disease that can be as low as ten hertz [61]. The auto-correlation based methods wrongly assume the pitch is constant over the duration of the frame [54, 51, 52]. Methods that locate peaks in frequency domain, power spectrum or cepstral domain also suffer from similar drawbacks [55, 56, 57, 58]. For example, at 16KHz sampling, a 25 millisecond frame would correspond to 400 sample points and a frequency resolution of about twenty hertz or more. Increasing the resolution with longer frames violates the stationarity assumption as the frame includes sounds corresponding to different phones. In other words, the pitch estimator needs to measure tremors of the order of 10 hertz using standard 25 millisecond time frames, which most current methods cannot do. One notable exception is the harmonic model of speech where the harmonic coefficients are allowed to vary in time within the frame. This model takes into account the harmonic nature of voiced speech and can be formulated to estimate pitch candidates with maximum likelihood criterion [11].

2. **Spectral entropy**:

   In information theory, the entropy is often referred as a metric of uncertainty, disorder, or unpredictability of random variables. The more a random event is predictable the lower the entropy is. In signal processing, this concept has been extended to the spectral entropy where we consider that the random variable is the spectral energies of a signal. Similarly, the more the spectrum is disordered, the higher the spectral entropy is. Therefore, an appropriate metric to measure the organization of the speech signal is spectral entropy. Properties of the spectrum serve as a useful proxy for cues related to voicing and quality. Spectral entropy can be used to characterize

speechiness of the signal and has been widely employed to discriminate speech from noise. As such, we compute the entropy of the log power spectrum for each frame, where the log domain was chosen to mirror perception.

3. **Mel frequency cepstral coefficients (MFCC)**: Shape of the spectral envelop is extracted from cepstral coefficients. Thirteen mel frequency cepstral coefficients (MFCCs) of each frame was augmented with their first- and second-order time derivatives.

4. **Segmental duration**: In the time-domain, apart from the energy at each frame, we compute the number and duration of voiced and unvoiced segments, which provides useful cues about speaking rate.

5. **Harmonicity, shimmer, and jitter**: Laryngologists often rate the degree of hoarseness (or harshness) to assess the functioning of the larynx. Spectrograms and perceptual studies reveal that this perceived abnormality of the voice is related to loss of harmonic components [72]. As the degree of perceived hoarseness increases, more noise appears to replace the harmonic structure and as a result harmonic to noise ratio (HNR) decreases. Irregular vocal fold vibration causes random modulation of the source signal and affects the amplitude (shimmer) distribution of harmonics throughout the spectrum and its time period (jitter). In addition, the ratio of energy between first and second harmonics (H1/H2) for each voiced frame has been found to be useful for characterizing breathy voice resulting from incomplete closure of vocal folds. Using the estimate of pitch and adopting a time-varying harmonic model of voiced signal, outlined at Chapter 2, we compute four quantities related to harmonic content. Note that this model allows the amplitude of the harmonics to vary smoothly over the duration of the frame and thus it is able to follow perturbations associated with shimmer and jitter.

The features computed at the frame-level needs to be summarized into a global feature vector of fixed dimension for each subject before we can apply models for predicting clinical ratings. Features extracted from voiced regions tend to differ in nature compared to those

from unvoiced regions. These differences were preserved and features were summarized in voiced and unvoiced regions separately. Each feature was summarized across all frames from the both voiced and unvoiced segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. The resulting features were computed separately for the three elicitation tasks (phonation, DDK and reading) and augmented into one vector, up to 925 long, for each subject.

Regression models

The clinical rating of severity of Parkinson's disease as measured by the motor sub-scale of UPDRS (mUPDRS) was predicted from extracted speech features using several regression models estimated by support vector machines using Scikit-learn toolkit [80]. We investigated three forms of regularization, L2-norm in ridge regression, L1-norm in lasso, and hinge loss function in support vector machine. The models were evaluated using 5-fold cross-validation, 30-fold cross-validation, and leave-one-out cross-validation. The results were measured using mean absolute error (MAE). Globally, across all three regressions, the performance of the leave-one-out cross-validation was better than 5-fold or 30-fold cross-validation and we adopt that for subsequent experiments. This is not only true for average absolute error but also for the standard deviation of the errors. This is not surprising since the leave-one-out cross-validation makes most of the data available for learning the parameters in each fold.

The performance of different learning strategies for leave-one-out cross-validation using baseline features (*openSMILE*) and our features (HM) are shown in Table 3.1. For this experiment, we considered all the speech elicited from each battery of speech tasks on a subject as one sample for our cross-validation evaluation.

The ridge regression and the support vector regression with both baseline features and those extracted from harmonic model perform significantly better than chance with a p-value of less than 0.001, according to cross-validated paired t-test, and is denoted by (†) in the table. Although there is not sufficient data to demonstrate statistical significance, the HM features consistently outperform the baseline features extracted by *openSMILE*. Ridge regression is much faster to learn than the support vector regression and we adopt

that for subsequent experiments in this study. The explained variance for the regression using features extracted from the harmonic model is about 61% when averaged across the cross-validation folds.

Model	Chance	Speech Features	
		openSMILE	HM
Lasso	8.0	6.9	6.5
Ridge	8.0	$5.9^{\dagger}$	$5.5^{\dagger}$
Linear SVR	8.0	$5.9^{\dagger}$	$5.8^{\dagger}$

Table 3.1: Comparison of different learning strategies on the accuracy of inferring severity of PD measured in terms of mean absolute error for leave-one-out cross validation.

Effectiveness of speech elicitation tasks

In this section, we separately examine the influence of each speech elicitation task in inferring the the severity of speech and report our results in the Table 3.2. As we mentioned earlier, speech elicitation task is consist of three tasks: 1) sustained phonation task, 2) diadochokinetic task, and 3) the reading task. These tasks influence differently on the speech production system and can reveal different clues for example, speech rate in reading task. We examined the benefits of the additional tasks in inferring the severity from speech. The results show that the sustained phonation task by itself is not particularly effective at this task. In contrast, the diadochokinetic task is a simple task and the speech features extracted from it are better at assessing the severity of the disease. The features extracted from the reading task are most effective at this task. This is an interesting result in particular, because Little and his colleagues [96] claim that phonation task is sufficient to obtain highly accurate PD diagnosis. It is very likely that their results are highly biased due to the the independence assumption of speech frames from each session. So, they model not just the difference between speakers due to PD but also due to normal variations in speaker traits. The performance of regressions with features from DDK and reading tasks are statistically better than chance, while the regression with features from phonation task is not.

	Elicitation Tasks	Mean of Absolute Errors
(a)	Phonation Task	7.1
(b)	DDK Task	$6.1^{\dagger}$
(c)	Reading Task	$5.6^{\dagger}$
(d)	All (a+b+c)	$5.5^{\dagger}$

Table 3.2: Effectiveness of elicitation tasks in capturing the severity of PD using features extracted by HM from speech, elicited by different tasks.

Influence of control

Apart from data described in the corpus, speech recordings were obtained from 21 controls from one clinic, where they were assigned a UPDRS motor score of zero without assessments. Thus the scale has a discontinuity close to zero which makes it difficult to learn a good fit for the controls. We investigate this effect by learning two different regressions, with and without the controls. The results for leave-one-out cross-validation are reported in Table 3.3 for ridge regression using features from harmonic model. The mean absolute error for chance or the best guess increases to 9.0 when the controls are included. The ridge regression improves the inferred severity in both cases. In Figure 3.1, we illustrate the correlation between the inferred severity and the clinical reference. The overall correlation in both cases is about the same, at 0.66. However, there is a large variance in inferred severity for the controls, as represented by the points on the y-axis.

Subjects	Mean of Absolute Errors	
	Chance	HM features
PD	8.0	$5.5^{\dagger}$
PD + control	9.3	$6.6^{\dagger}$

Table 3.3: Effect of including controls in the training data for inferring severity of PD using features extracted by HM from all speech tasks

Clinic-specific influence

One problem with clinical studies where the gold standard itself has a subjective component is the bias introduced by the data collected in each clinic. The bias could be due to numerous factors including the severity of the disease in the patient population or the

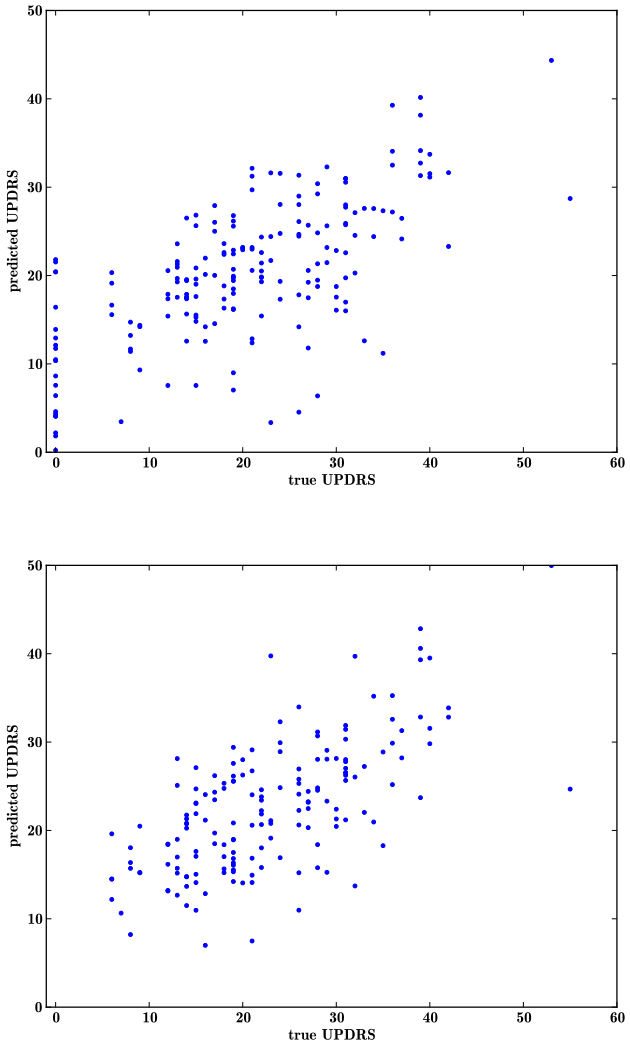

Figure 3.1: Plot of reference UPDRS vs. predicted UPDRS to illustrate how controls (assigned a reference UPDRS score of zero) skew the performance of the inference of severity of PD.

training of those administering the assessments. Our multisite study affords an opportunity to check this variability. We separated the data from the three clinics and Figure 3.2 illustrates the difference in distribution of severity of PD patients observed in the three clinics. The patients seen at clinic 1 have a wider distribution of severity of the disease than the other two clinics. Incidentally, more patients (99) were seen at clinic 1 than in

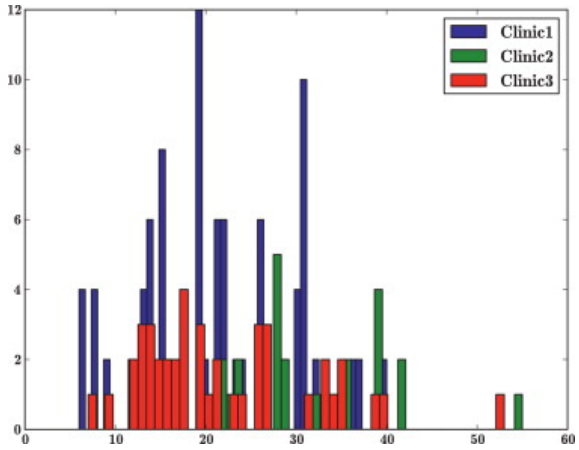

Figure 3.2: Plot illustrates the difference in the frequency (y-axis) of severity of PD across the three clinics in terms of UPDRS motor scale (x-axis).

the other two clinics (43 and 26). The median UPDRS motor score per clinic ranged from 6.9 to 7.4.

Training data	# of Subjects	Mean of Absolute Errors	
		Chance	HM features
Clinic 1	99	6.9	4.6†
Clinic 2	26	7.4	6.1
Clinic 3	43	7.3	5.7
All	168	8.0	5.5†

Table 3.4: Difference in inferring the severity of PD using speech collected from different clinics.

For understanding how well our models can generalize across clinics, we learned a separate model for each clinic and evaluated the model on the data from the other two clinics. Our results based on the features from our proposed harmonic model are reported in Table 3.4. The regression learned on clinic 1, which has the most diverse as well as the most number of patients, generalizes better than those trained on the other two clinics. While this is not a surprising result, it underlines the need for samples that are more diverse in severity and larger in number than in previous studies (Tsanas et al., 2010 and Bocklet et al., 2011).

3.1.6 Conclusions

In summary, we have reported our experiments on inferring the severity of PD from speech recorded from a relatively large sample of 168 subjects, from multiple clinics, using three elicitation tasks: the sustained phonation task, the diadochokinetic task and the reading task. These tasks can be administered automatically remotely and our results show that the severity can be inferred with a mean absolute error of 5.5, and consistently well-above chance across all clinics. In the framework described in this study, the errors will be lower in applications where the progression of the disease needs to be monitored over time and there is an opportunity to learn the regression for each subject via better priors. We found that our pitch related features are consistently better than alternative features across different test conditions. Our analysis of the results show that the phonation task is a poor predictor of the severity, unlike widely publicized claims by Little and his colleagues. Diadokinetic and reading tasks are better predictors and the combination of all three tasks gives the best results.

3.2 Detecting and diagnosing autism spectrum disorder

3.2.1 Introduction

The term autism spectrum disorder (ASD) covers a range of developmental disabilities that can cause significant social, communication, and behavioral challenges. Children with ASD often are self-absorbed in their private world and they have difficulties in communicating and interacting with others. While not every child with ASD has a language problem, the majority have difficulty using language effectively, especially when conversing with others. Often they exhibit unusual pitch and intonation, for example, monotonous pitch, reduced stress, odd rhythm, large pitch range [6], and even differences in harmonic structure of their speech [7]. There has been continual interest in characterizing these variations in ASD and potentially exploit them in objectively quantify and categorizing the language impairments in ASD.

The range of disorders in ASD are categorized according to Diagnostic and Statistical Manual of Mental Disorders (DSM), published by American Psychiatric Association. Most clinicians follow the fourth edition (DSM-IV) [97]. The diagnostic category pervasive developmental disorders (PDD) refers to disorders characterized by delays in the development of multiple basic functions including socialization and communication. This category includes Asperger and Rett syndromes. Pervasive developmental disorder not otherwise specified (PDD-NOS) is one of the five ASDs, characterized as "severe and pervasive impairment in the development of reciprocal social interaction or verbal and nonverbal communication skills, or when stereotyped behavior, interests, and activities are present, but the criteria are not met for a specific PDD" [1] or for several other disorders. Unrelated to the above conditions, a child could suffer from limited ability to socialize and communicate, not because of general developmental disorders, but due to specific language impairments such as dysphagia. In all these cases, prosody and intonation are compromised perhaps in different ways, and that is a topic of considerable research interest currently especially for developing useful intervention strategies.

[1] www.firstsigns.org/screening/DSM4.htm#pdd-nos

In this study, we report experiments on the Interspeech 2013 Autism sub-challenge task. The challenge consists of two tasks: 1) a binary 'Typicality' classification task with classes – TYPically developing (TYP) and ATYpically developing (ATY), and a four-way 'Diagnosis' task for classifying children into 4 categories – TYP, PDD, PDD-NOS, and specific language impairment such as DYSphasia (DYS). We had no input in the study design or data collection and use this data as provided by the organizers with all its caveats. Here, we aim to compare our developed acoustic features without making any claims about the clinical significance of results.

3.2.2 Corpus

Class	Train	Dev	Test	Σ
Typically developing				
TYP	566	543	542	1651
Atypically developing				
PDD	104	104	99	307
NOS	104	68	75	247
DYS	129	104	104	337
Σ	903	819	820	2542

Table 3.5: Test, train, and development sets

Empirical evaluation reported in this study were performed on Child Pathological Speech Database (CPSD) [98] collected from 99 children, age 9 to 18, through two hospitals located in Paris, France. This dataset provides 2542 short speech utterances collected for assessing children's abilities in imitation of different types of prosody contours. Based on the prosodic dependencies of French language, sentences carry out 4 intonation types, including *descending*, *falling*, *floating*, and *rising*. Subjects, were asked to read 26 phonetically easy sentences and they were recorded in separate files. As a clinical reference, the severity of subjects condition were measured by clinicians using the DSM-IV criteria [97], where 35 of these children showed PDD either of Autism Spectrum Condition (ASC, 12 children), specific language impairment (SLI, 13 children) or PDD non-otherwise specified (PDD-NOS, 10 children). The corpora includes rich annotation such as speaker metadata, orthographic transcript, phonemic transcript, and segmentation. Also, the corpus

treats sentences read by the same speaker as independent samples partitioned randomly in test, development, and training sets shown in Table 3.5.

Speech Features	TYP vs. ATY	
	Dev set	Test set
Baseline	92.8	90.7
HM Features	**98.12**	**93.58**
Chance	50.0	50.0

Speech Features	4-subtypes of ASD	
	Dev set	Test set
Baseline	52.4	67.1
HM Features	**57.38**	**69.42**
Chance	25.0	25.0

Table 3.6: Unweighted average recall (UAR) for detecting ASD kids from typically developing (TYP) kids (top), and for classifying the ASD kids into four sub-types (bottom).

3.2.3 Experiments

Baseline features

The provided corpus consists of 6125 acoustic features, per audio recording, extracted by TUM's open-source *openSMILE* feature extractor [95]. Despite the high dimension, the feature itself derived from a relatively small pool of acoustic phenomena including energy, spectral, cepstral, and voicing related low-level descriptors (LLDs). On the other hand, the features computed at the low-level were summarized into a global feature vector of fixed dimension for each recording using a variety of functionals related to local extrema, such as mean and standard deviation of inter-maxima distances, and quantiles. For more detail regarding the baseline low-level features and also functionals that are applied to those features, we refer the reader to the challenge paper [69].

Proposed features

We employ our developed features from harmonic model of speech as described earlier at Chapter 2. Briefly, we extract 25 millisecond long frames using a Hanning window at a rate of 100 frames per second before computing the frame-level features. Voicing related features including pitch, HNR, the ratio of energy in first to second harmonics (H1/H2), jitter, and shimmer are derived from the harmonic analysis over the voiced frames. The features computed at the frame-level are then summarized into a global feature vector of fixed dimension for each read sentence. Each feature was summarized across all frames from the voiced segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. We also computed the covariance matrix (upper triangular elements) of frame-level feature vectors over voiced segments to capture interaction between features. The resulting per-sentence voice quality feature vector was later augmented by per-sentence energy, spectral, and cepstral related features provided from baseline.

Regression and classification models

Typically, in clinical applications, the class distributions are highly unbalanced, as it is in the four subtypes within this corpus. The challenge evaluation metric of unweighted average recall attempts to normalize the influence of the highly skewed classes. We employed a support vector classifier and a support vector regression respectively to detect ASD cases and to identify the subtypes. Both the regression and classifier were learned from the data using open-source WEKA toolkit [99]. For the training the regression and classifier, we retained the hyper parameters from the baseline system, $C = 0.001$. For the test set, all labeled data from train and developing sets were pooled for training and a new model learned using parameters reported in the baseline. Since the class distribution in the training data was skewed, we upsampled instances in atypicality categories (PDD, NOS, and DYS) by a factor of five. Table 3.6 reports UAR evaluated from baseline feature vectors and proposed feature vector on detecting ASD and classifying the sub-types.

From the results, it is clear that our voice quality related features (derived by harmonic analysis) significantly improve UAR in both tasks.

3.2.4 Conclusions

In summary, we considered several speech measures to detect children with ASD and to classify them into four subtypes. For both tasks, our features can be categorized into four groups – voice quality features (estimated from harmonic analysis), energy-related features, spectral features, and cepstral features. We found that our features, specifically the voice quality features, improve the performance of both tasks in terms of unweighted average recall (UAR), of detecting autism spectrum disorder by 2.3% and diagnosing the disorder into four categories by 2.8% over the baseline in test set of this challenge.

3.3 Inferring clinical depression from speech and spoken utterances

3.3.1 Introduction

Clinical depression is a common mental disorder that negatively affects person's health, mood, thoughts, behavior, work, family, and ability to function in everyday life [100, 101]. It is one of the basic forms of mood disorders and signaled by a series of abnormalities such as loss of interests, avoidance of social interactions, and changes of mood, feeling, eating, and sleeping. It is a disease that is often unrecognized though common, threatening a large group of people each year, for example, over 19 million American adults [102]. The diagnosis is subjectively performed by an expert practitioner based on patient's mental state driven from interviews and self-report experiences. This assessment is costly, time-consuming, and often requires patent's presence at the clinic. Recent studies have explored the influence of emotional changes on phonatory and articulatory characteristics of speech production system [5]. These observations have motivated researchers to explore alternative approaches based on speech processing techniques, which can be used in real applications such as automatically screening and telemonitoring of depressive disorders. Since then, a number of studies have attempted to find potential clues in subject's speech that reflect influences of mood disorders [103]. Acoustic features of speech signal including pitch, formants, harmonic-to-noise Ratio (HNR), shimmer, jitter, speech rate, energy, and glottal features have been used to analyze of depressed voices [104, 12, 13, 105].

Brief review of previous works

Speech pathologist have characterized depressed speech as monotone, mono-loud, and lifeless [106]. There has been a considerable interest on analyzing acoustic properties of speech for the hope of quantitative assessment of clinical depression [107, 108, 109, 3]. Here, we describe a few studies. Moore and his colleagues [110] attempted to employ prosodic (pitch, energy, speech rate), vocal tract (first, second, and third formant frequencies and their bandwidths), and glottal features (starting points of glottal opening and closing, minimum point in glottal derivative, maximum glottal opening) for classifying

15 depressed (six males and nine females) from 18 control (nine males and nine females) subjects. They achieved the best classification accuracy, up to 96%, in a leave-on-out cross validation strategy by combination of vocal and glottal features. However, this study has been conducted on a relatively small dataset and results cannot be generalized to a larger population. Recently, Low and his colleagues introduced an automatic approach to classify 68 clinically depressed adolescents from 71 controls. They employed a range of speech features classified into five categories of spectral, prosodic, cepstral, glottal, and features derived from a non-linear operation, Teager energy operator (TEO). They evaluated the performance of different combination of features by GMM and SVM classifiers using cross validation and reported the classification accuracy of 81%-87% for males and 72%-79% for females. In another study, Algowinem and his colleagues [111] applied several machine learning strategies including hierarchal fuzzy signature (HFZ) and multi layer perceptron (MLP) classifiers on a broad range of speech measures. They reported that loudness, intensity, and root mean square are strongest voice features for their task.

3.3.2 Corpus

Our corpus for this study was collected by Oregon Research Institute (ORI) and consists of video recordings of adolescents subjects' during their interaction with their family. Subjects were asked to participate in three different 20-minutes interactions with their parents: event-planning interaction (EPI), problem-solving interaction (PSI), family consensus interaction (FCI) [112]. All interaction were administrated by a trained interviewer in a quiet room at ORI. The recordings were collected from 148 subjects, including 98 females and 50 males, 14 to 18 year old. Of these subjects, based on clinical assessment, 71 adolescents (50 females and 21 males) were diagnosed depressed and 77 individuals (48 females and 29 males) were healthy controls.

As a clinical reference, the severity of subject's condition were coded by living-in-family-environment (LIFE) coding system [112]. The LIFE coding system was developed for assessment of behavioral characteristic of individuals with depressive disorders. In

LIFE coding system, behavior is coded based on verbal, nonverbal, and paraverbal behavior and codes do not necessarily imply whether subject is speaking. A group of psychologists at ORI coded subjects' verbal content and emotional state in terms of 27 contents code and 10 affect code available on the LIFE coding system. Subjects' behavior were marked by six categories of *angry*, *dysphoric*, *happy*, *neutral*, *end*, and *other* per each second in family interaction sessions. For the purpose of our experiments, we extracted audio from the video recordings and converted them from stereo to mono channel format. Then, speech segments annotated with *angry*, *dysphoric*, and *happy* tags where chopped and concatenated for the manual transcription. There was noise in the annotations as they were not always completely aligned with the utterances. We asked annotators to manually identify the speaker at each segment to alleviate noise in the labels.

3.3.3 Textual features

In order to gauge the effect of speech contents in the clinical depression, we extracted textual features from manual transcripts. To extract features from text, we used a published table of valence and arousal ratings by Warriner et al. [17] to tag each word in an utterance with an arousal and a valence rating and computing their per-utterance mean, standard deviation, minimum, and maximum. For missing words we imputed valence and arousal by randomly drawing 5 words from the table and computing their average.

3.3.4 Speech features

In our experiment, we used a broad range of features to capture speech clues associated with clinical depression. For our baseline system, we adopted the baseline features defined in Interspeech 2010 Paralinguistic Challenge [94] using *openSMILE* toolkit [95].

The features, comprised of 1582 components, can be broadly categorized into three groups: 1) loudness related features such as RMS energy and PCM loudness, 2) voicing related features like pitch frequency, jitter, and shimmer., and 3) articulatory related features such as mel-frequency cepstral coefficients and line spectral frequencies. The above configuration provides 38 features along with their derivatives to form the frame-level acoustic features. The derivatives allow us to capture local dynamics of pitch and

other features. The features computed at the frame-level were summarized into a global feature vector of fixed dimension for each recording using 21 standard statistical functions including min, max, mean, skewness, quartiles and percentile.

Speech features from harmonic model

Alternatively, we extracted prosodic features using the harmonic model of speech as we described earlier at Chapter 2.

Briefly, we extract 25 ms long frames using a Hanning window with a 10 ms shift. We first detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level is computed by formulating a one-state hidden Markov model (HMM) [9]. The state could either be voiced or unvoiced, with likelihood given by the per-frame harmonic model. The transition model consists of a simple zero-mean Gaussian. We compute voicing decision over the utterance using Viterbi alignment. Subsequently, we compute various voicing related features for voiced frames, including pitch frequency, HNR, H12, jitter, shimmer, and harmonic coefficients. These pitch-related features are combined with standard features including energy, spectral entropy, and MFCCs. For unvoiced frames, we just compute energy, spectral entropy, and MFCCs features. Features extracted from voiced regions tend to differ in nature compared to those from unvoiced regions. These differences were preserved and features were summarized in voiced and unvoiced regions separately. Per-utterance features are computed by applying standard summary statistics such as mean, median, variance, minimum and maximum to the per-frame voiced (unvoiced) features, generating a 192-dimensional per-utterance feature vector.

3.3.5 Experiments

We compared the performance of a SVM classifier on our data using 30-fold cross-validation for classifying depressed from control subjects. Table 3.7 reports the performance of classifiers trained on different feature sets. The SVM classifier with several kernel functions including linear, polynomial, and radial basis function were employed from open-source Scikit-learn toolkit [80]. Parameters of the optimal classifier were determined via grid

search and cross validation on training set. The best performance among all models was obtained with the linear kernel except for openSMILE feature set with RBF kernel. As it is shown in Table 3.7, speech features, extracted from both openSMILE and harmonic model, perform significantly better than chance with a p-value of less than 0.01, according to cross-validated paired t-test [81] and denoted by (†) in the table.

Features	Classification Accuracy
Chance	52.4
openSMILE	$64.7^{\dagger}$
Harmonic Model	$\mathbf{68.7}^{\dagger}$

Table 3.7: Comparison of performance of SVM classifier using different acoustic features for classifying clinical depression of adolesents.

Also, Table 3.8 indicates that incorporating textual features result in an additional improvement. Spoken words contain information for detecting depression as expected and solely use of textual features performed significantly better than chance.

Features	Classification Accuracy
Chance	52.4
Text	65.4
openSMILE+Text	$68.0^{\dagger}$
Harmonic Model+Text	$\mathbf{74.0}^{\dagger}$

Table 3.8: Comparison of performance of SVM classifier using different combination of textual and acoustic features for classifying clinical depression of adolesents.

Effectiveness of family interaction sessions

In this section, we separately examine the influence of each family interaction session in classifying the the depressed speech. As we mentioned earlier, there are three types of family interactions: 1) EPI, 2) PSI, and 3) FCI. These tasks differently evoke emotional state of adolescents during the family interaction and potentially reveal different aspects of their behavior. For instance, PSI tends to elicit the conflictual behavior of adolescents when interact with their parents. Table 3.9 reports the performance of SVM classifiers. The results indicate that features extracted from PSI are most effective in this task. The

performance of classifier trained on features extracted from harmonic model is significantly better than chance with p-value of less than 0.01. Our analysis of the results show that the FCI is not an effective task for evoking clues associated with clinical depression of adolescents and features extracted from this family session are poor predictor of the severity.

Speech Features	Classification Accuracy		
	EPI	PSI	FCI
Chance	49.2	49.2	49.2
openSMILE	60.0	64.7	56.0
Harmonic Model	66.1	**71.4**†	57.6

Table 3.9: Effect of family interaction sessions on classifying clinical depression in adolesents.

3.3.6 Conclusions

This study investigates the problem of detecting depression from recordings of subjects' spoken utterances. Given the scarcity of the text data for training models with n-grams, we explore an alternative method to extract content information related to affect by encoding words in terms of valence and arousal, using a look up table that has been compiled by averaging responses from large number of raters. We extract novel acoustic and prosodic features from harmonic models and find that they outperform standard features such as those computed from openSMILE. The textual features provide additional gain, achieving a classification accuracy of about 75%.

Chapter 4

Acoustic Features for Inferring Social Contexts

In the previous chapter, we investigated the utility of the speech features developed in Chapter 2 in clinical applications employing supervised machine learning techniques. Supervised methods cannot learn from unlabeled data and do not benefit from the large amount of unlabeled data available in some applications. In this chapter, we introduce a data learning framework employing deep neural networks (DNNs) to take potential advantages of unlabeled data. In particular, we investigate the problem of detecting social contexts from the audio recordings of everyday life such as in life-logs. We describe the data collection followed by a brief description of auto-encoders. We then propose a multi-label classification scheme and investigate different strategies for training this model. We report results of detecting social contexts specifically on two tasks related to speakers' location and activity. Finally, we conclude with a summary of our key results.

4.1 Introduction

Low power devices and smartphone apps have made it possible to record clips of everyday life with relative ease for sharing on social media or for archiving. The recordings such as those on YouTube have a wide range of background noise and the task of transcribing the spoken utterances present in them is challenging. Social context is an alternative layer of information that can be inferred from the audio and used to annotate and index these clips. Furthermore, real-time inference of social context would be highly useful in providing personalized services on smartphones. For social scientists, psychologists and gerontologists, the ability to infer social context from audio life logs provides a convenient way to study social behaviors without perturbing the behavior itself, unlike previous methods of measurements that relied on sampling or recollections in journal entries. Through a number of studies, ranging from depression to adolescent behavior, Pennebaker and Mehl have already illustrated the value of inferring social contexts from audio life logs even when they were severely handicapped by the need to listen to audio and manually annotate them [20]. They demonstrated that social context and other information from the audio life logs can be used to quantify subjects' social life (interaction and engagement), cognitive function, emotional conditions, or even health status [113]. In this study, our

focus is on automatically inferring social contexts from life logs specifically those collected by Mehl and colleagues [114].

The audio life logs can be easily collected in large amounts, however, annotating them is relatively expensive. Supervised classifiers such as support vector machines do not have the ability of using unlabeled data. On the other hand, deep neural networks, specifically, auto-encoders can extract potentially useful features in an unsupervised manner. The layers of the network that extract these features can then be modified in a supervised manner to fine-tune the network for a classification task with limited amounts of labeled data. We investigate the use of such a framework to detect social contexts [115, 116], such as speakers' location (e.g. in transit) or activity (e.g. watching TV or eating) in audio life logs. Moreover, the early layers that extract potentially useful features may be shared across related tasks such as the two classification tasks. In certain applications, such sharing have been shown to be beneficial and so we investigate multi-label learning on audio life logs. Most previous work on the general audio classification task have employed MFCCs. Experiments on speech recognition show that neural networks provide better performance with filter bank features. We investigate an alternative feature representation using harmonic model that nicely captures the harmonic nature of natural sounds. We report experiments on real world samples of life logs with manual annotations.

4.2 Corpus

Our corpus for this study consists of samples of snippets of audio recordings from everyday life of university students. The corpus was collected using the Electronically Activated Recorder (EAR), which records 30-seconds snippets, every 12 minutes. This sampling scheme was chosen to provide sufficient information about social lives of the students in a way that doesn't allow full reconstruction of their private lives [117]. The corpus was collected from 96 student volunteers, who were asked to wear a lapel microphone connected to the EAR device for 4 consecutive days during their waking hours. The resulting 22,140 audio snippets or files were manually annotated by 8 trained annotators. There were 4 annotation tasks: the speaker's current location (e.g. in-transit and outdoor), activity

(e.g. socializing and eating), mood (e.g. arguing and laughing), and interaction (e.g. alone and talking). For details, see the Social Environment Coding of Sound Inventory (SECSI) [114]. The inter-rater agreement was assessed on a set of 392 recordings, where the intraclass correlations (ICCs) based on two-way random effects model [118] was found to exceed 0.71 for all the categories.

4.3 Deep neural network

Deep neural networks have been demonstrated to outperform other machine learning techniques in a variety of tasks ranging from speech recognition to computer vision. In brief, a deep neural network (DNN) is a feed-forward network comprised of several layers of hidden units. These networks are trained using back-propagation, which was introduced by Rumelhart and his colleagues [119]. The problems encountered in training large number of layers using back-propagation is better understood now. Improvement in training techniques and availability of large amounts of data and compute power, have helped resurrect them.

The DNN architecture for classifying the speaker's location and activity is shown in Figure 4.1. It consists of several hidden units and two separate softmax units, in the output layer, to classify two independent classes of events (i.e. location and activity). Two steps of learning the network is described as follow: 1) Unsupervised learning of hidden layers, known as *pre-training*, performed in a greedy layer-wise fashion introduced by Hinton [120]. First, a sparse autoencoder is trained in the first layer in an unsupervised manner. Then, the output of this layer (feature activations in hidden layer) is fed to the higher layer as input features. Similarly, the second sparse autoencoder and eventually higher layers are trained. 2) Supervised *fine-tuning* performed by the back-propagation algorithm, in which estimated weights of softmax and hidden units are adjusted.

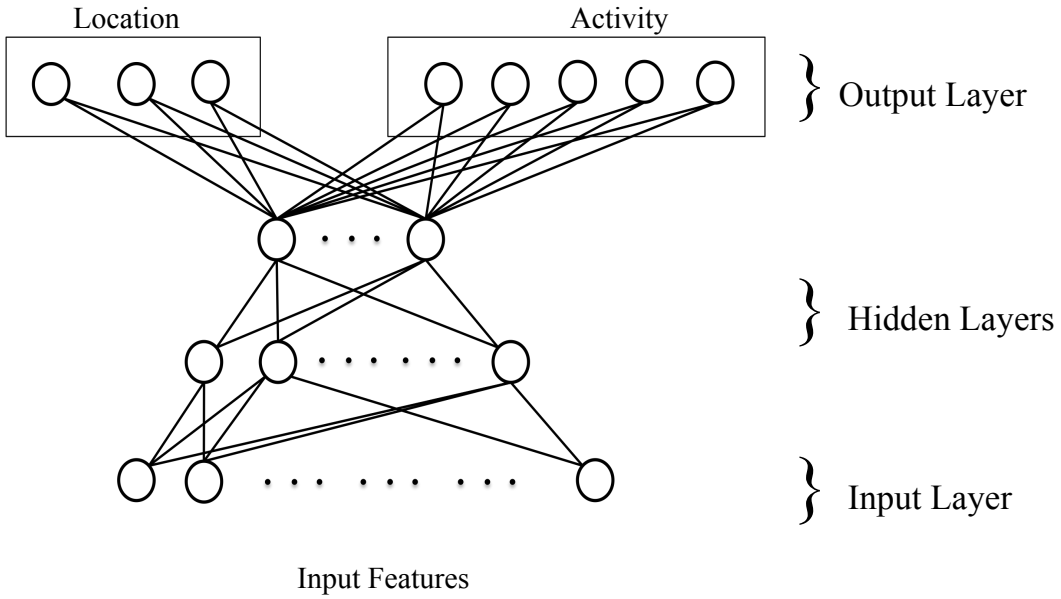

Figure 4.1: DNN for classifying speakers' location and activity

4.3.1 Sparse autoencoder

Autoencoder

The Autoencoder (AE) is a neural network used for compact representation of data. As a generative model with one layer of hidden units, the autoencoder explores the underlying structure of input features by learning an identity function that maps the input vector of $\mathbf{x}$ into the output vector of $\hat{\mathbf{x}}$, which is similar to the $\mathbf{x}$. Suppose we have D-dimensional unlabeled speech features (like MFCCs) set $\{x_1, x_2, x_3, \cdots, x_n\}$. The learning process begins with an encoding function f that maps input vectors to the hidden representation as shown in Equation 4.3.1.

$$h = f(x) \tag{4.1}$$

$$= s(W_1\mathbf{x} + b_1) \tag{4.2}$$

where s is a non-linear activation function such as logistic sigmoid, and b_1 is a bias vector. Then, a decoder function g maps the hidden representation to the output layer

for reconstructing the vector $\mathbf{x}$.

$$\hat{x} = g(h) \tag{4.3}$$

$$= s(W_2\mathbf{x} + b_2) \tag{4.4}$$

The goal of learning is to estimate unknown parameters $\theta = \{W_1, W_2, b_1, b_2\}$ such that the reconstruction error on the training set is minimized. We define the following objective function where the cost function, L, is the traditional squared error as follows:

$$J(\theta) = \sum_{t=1}^{n} L(x_t, g(f(x_t))) \tag{4.5}$$

Sparse autoencoder

When the number of units in the hidden layer is large, the autoencoder is still able to capture structure of data by imposing a sparsity constraint over hidden units. This can be viewed as adding a sparsity penalty term to the objective function. The sparsity term imposes a constraint on units of hidden layer to having sparse average activation. Following an approach proposed by Andrew Ng [121], we add the penalty term to the objective function as follow:

$$J(\theta) = \sum_{i=1}^{n} L(x_i, g(f(x_i))) + \beta \sum_{j=1}^{n_h} \mathrm{KL}(\rho||\hat{\rho}_j) \tag{4.6}$$

where β controls the weight of the sparsity term, $\hat{\rho}_j$ is average activation of hidden unit j over whole training set, and ρ is a constant sparsity term close to zero (typically $\rho = 0.05$) that defines the level of the sparsity. $\mathrm{KL}(\rho||\hat{\rho}_j)$ is the Kullback-Leibler (KL) divergence between a Bernoulli random variable with the mean of ρ and a Bernoulli random variable with the mean of $\hat{\rho}_j$.

4.3.2 Softmax classifier

The output layer is designed to encode each instance of input feature vector, $\mathbf{x}$, into one of the K categories. This is typically conducted by adopting a softmax classifier with number of nodes equal to the number of categories. Computationally, it is convenient to

convert labels into *1-of-K* coding scheme [122]. The conditional probability $p(c_k|\mathbf{x})$ of a softmax classifier is given by:

$$p(c_k|\phi) = \frac{\exp(D_k^T \phi)}{\sum_{j=1}^{K} \exp(D_j^T \phi)} \tag{4.7}$$

where c_k is the class identifier, D is the weight matrix of softmax classifier and ϕ is a feature vector (output of the last hidden layer). Given the feature vector, ϕ, conditional probability for all classes, $c_1, \ldots, c_k$, is evaluated for predicting the target category employing an *argmax* function as follow.

$$\hat{c} = \underset{c_1,\ldots,c_k}{\arg\max} \; p(c_k|\phi) \tag{4.8}$$

Unknown weights of the softmax layer is typically estimated by minimizing the cross-entropy error function given by:

$$L(\mathbf{D}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log p(c_k|\phi_n) \tag{4.9}$$

where N is the total number of training instances and t_n stands for a vector of binary label associated with the $n_t h$ data instance.

4.3.3 Multi-label classifier

In multi-label classification, contrary to the single-label classification, instances of data are associated with a set of class labels. Medical diagnosis, text documents, and movie genres are a few examples that often can be categorized into more than on category. Multi-label classification problem can be cast into a set of independent single-label classification problems [123], in which each classifier is independently trained over training instances of the class. For example in our classification task, one can independently learn two DNNs for separate classification of speaker's current *activity* and *location*. However, class labels are not necessarily independent of each other and thus, correlated information is ignored in this approach. As an alternative approach, instead of employing two independent DNNs for multi-label classification, we propose to employ a single DNN with two independent softmax units in its output layer as shown in Figure 4.1. In this architecture, the network

is learned to minimize the cross-entropy error function given by:

$$L(\mathbf{D_a}, \mathbf{D_l}) = -[\sum_{n=1}^{N}\sum_{k=1}^{K_a} t_{nk} \log p(c_k^a|\phi_n) + \sum_{n=1}^{N}\sum_{k=1}^{K_l} t_{nk} \log p(c_k^l|\phi_n)] \quad (4.10)$$

where D_a and D_l are the weight matrices, K_a and K_l are the number of categories, and c^a and c^l are class identifiers of *activity* and *location* softmax units, respectively.

4.3.4 Fine-tuning

After pre-training of hidden layers, weights of hidden layers, θ, and softmax units, D_a and D_l, are fine-tuned in a supervised fashion using back-propagation algorithm. The gradient of the cross-entropy error function is back-propagated and the the gradient with respect to the weights are computed at each layer, which are used to update the respective weights.

4.4 Speech Features

In our experiment, we used a broad range of features to capture features associated with these categories. For our baseline system, we adopted the baseline features defined in Interspeech 2010 Paralinguistic Challenge [94] using *openSMILE* toolkit [95]. The features, comprised of 1582 components, can be broadly categorized into three groups: 1) loudness related features such as RMS energy and PCM loudness, 2) voicing related features like pitch frequency, jitter, and shimmer., and 3) articulatory related features such as mel-frequency cepstral coefficients and line spectral frequencies. The above configuration provides 38 features along with their derivatives to form the frame-level acoustic features. The derivatives allow us to capture local dynamics of pitch and other features. The features computed at the frame-level were summarized into a global feature vector of fixed dimension for each recording using 21 standard statistical functions including min, max, mean, skewness, quartiles and percentile.

Alternatively, we employ two sets of acoustic features for the comparison: 1) Mel frequency cepstral coefficients (MFCCs), which are widely used in automatic speech and

speaker recognition systems, and 2) pitch-related features extracted from harmonic model of speech.

4.4.1 Speech features from harmonic model

We extract 25 millisecond long frames using a Hanning window at a rate of 100 frames per second before computing the frame-level features using harmonic model (HM) described at Chapter 2. Briefly, voicing related features including pitch, HNR, the ratio of energy in first to second harmonics (H1/H2), jitter, and shimmer are derived from the harmonic analysis over the voiced frames. Moreover, we add the vector of harmonic coefficients to the set of frame-level features. Given the constant number of harmonics obtained from the model order selection (Chapter 2, Section 2.2.4), $H = 7$ in this task, we estimate the coefficient of harmonics and introduce a 15-dimensional feature vector, $[a_0, a_1, \ldots, a_7, b_1, \ldots, b_7]^T$. We then transform the feature vector to the log-domain followed by taking its absolute value. After extracting frame-level features, we summarize them into a global feature vector of fixed dimension for each 30-second long recordings. Features extracted from voiced regions tend to differ in nature compared to those from unvoiced regions. These differences were preserved and features were summarized in voiced and unvoiced regions separately. Each feature was summarized across all frames from the voiced (unvoiced) segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. Ultimately, segment-level summary features of voiced and unvoiced regions were concatenated into a global vector of 182 features for each recording .

4.5 Experiments

We evaluate the effectiveness of the proposed method on learning single-label and multi-label classification models. For our experiments, we chose to automatically identify the speaker's current location and activity on a subsample of recordings in the corpus. In particular, we chose recordings annotated with in-apartment, in-transit, or in-restaurant from the *Location* class, as well as sleeping, eating, watching TV, studying, or working from the *Activity* class. This gave us an evaluation set containing of 1470 recordings with

two-label annotations. Training and test set of recordings on this evaluation set were then defined by speaker independent subdivisions as shown in Table 4.1.

Class		Train	Test	Total
Location	in-restaurant	384	126	510
	in-apartment	477	189	666
	in-transit	189	105	284
	total	1050	420	1470
Activity	computer	321	135	456
	social	207	69	276
	eat	333	117	450
	work	189	99	288
	total	1050	420	1470

Table 4.1: Test and training sets for *location* and *activity*.

4.5.1 Evaluation metrics

In order to evaluate the performance of proposed multi-label classifier, we adopt evaluation metrics that are frequently used in the literature [124, 125, 126]. These metrics are different from conventional metrics adopt for single-label classification problems. To formulate the evaluation metrics, let Γ be our multi-label dataset containing N multi-label instances, $(\mathbf{x_i}, \mathbf{y_i})$, $i = 1, \ldots, N$. Also, $|L|$ stands for the number of labels in the target vector $\mathbf{y_i}$, and $h(\mathbf{x_i}) = \mathbf{z_i}$ denotes predicted labels for the instance $\mathbf{x_i}$. We assess the performance of our multi-label classification task using accuracy (ACU) and Hamming-Loss (HL) described at bellow.

Accuracy

For a multi-label data instance, accuracy (ACU) is the portion of labels that are correctly identified in all classes divided by the total number of labels [124]. The main drawback of this measure is that it ignores partially correct labels. Higher accuracy corresponds to better performance. For the entire dataset, accuracy is computed as follow.

$$Accuracy(\mathbf{z}, \mathbf{y}) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{z_i} = \mathbf{y_i}) \tag{4.11}$$

where I is the indicator function.

Hamming-Loss (HL)

HL is a widely used criterion in multi-label classification that counts the number of instances that are incorrectly predicted [124]. Lower the value of HL corresponds to the better performance and defined as follow.

$$HL(\mathbf{z}, \mathbf{y}) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{x_i} \oplus \mathbf{y_i}}{|L|} \tag{4.12}$$

where $\oplus$ stands for symmetric difference (XOR operation) between two $h(\mathbf{x_i})$ and $\mathbf{y_i}$ sets.

4.5.2 Multi-label classification

For the evaluation of multi-label classification task, we learn a DNN with two hidden layers, each containing 1,024 sigmoid hiden units. We also stack two independent softmax units representing *Location* and *Activity* classes on top of the DNN as depicted in Figure 4.1. First, we pool all available data in the corpus for pre-training of weights of hidden layers in the DNN using the sparse autoencoder. Note that this step of training is conducted in an unsupervised fashion on unlabeled data. Next, we feed training data into the pre-trained network and extract activation features from the last hidden layer. This step is known as feed-forwarding. Activation features are then used for independently training of two softmax units with the number of units equal to 4 and 3 corresponding to number of categories in *activity* and *location* classes, respectively. In order to estimate unknown weights of softmax units, we define a cross-entropy error function described in Equation 4.9 and minimize it in a supervised manner. Estimated weights of softmax units are then stacked with pre-trained weights of the DNN to construct a multi-label classifier. The last step of training is *fine-tuning*, in which weights of hidden layers and softmax units are jointly adjusted in order to minimize the cross-entropy error function given by Equation 4.10. The length of feature vector input into the first layer of the DNN is 182, 1582, and 39 for HM, MFCCs, and *openSMILE* (OS) feature sets, respectively. the In order to highlight the effect of *fine-tuning*, we separately evaluated our multi-label classifier in two scenarios of *before* and *after* step of *fine-tuning*, and independently reported their performances in the Table 4.2. The table shows the accuracy and hamming-loss of the

DNN-based classifier trained on three types of speech features on our multi-label evaluation set. As it is seen in the results, features from harmonic model perform better than MFCCs and *openSMILE* features, but the improvement may not be statistically significant in this evaluation set. Also, it is seen that that *fine-tuning* of the model parameters significantly improve the classification performance in terms of all evaluation metrics.

Metric	Fine-tuning	Speech Features		
		MFCC	OS	HM
Accuracy	No	68.4	70.3	70.0
	Yes	69.7	72.3	**80.1**
Hamming-Loss	No	14.8	11.8	12.5
	Yes	12.5	10.9	**8.57**

Table 4.2: Effect of *fine-tuning* DNNs with different features.

In order to gauge the relative importance of number of hidden units, we repeated this experiment with models learned on HM features with varying number of hidden units and reported their performance in the Table 4.3. Note that we retained the number of hidden units equal in each hidden layer. The results show that increasing the number of hidden units improve the performance by reducing the Hamming-Loss, though accuracy does not change. Also, adding more than 1,024 hidden units does not add to the performance.

Metric	Number of hidden units				
	128	256	512	1,024	2,048
Accuracy	80.1	80.1	80.1	80.1	78.5
Hamming-Loss	10.8	9.57	9.28	8.57	9.86

Table 4.3: Effect of the size of hidden layer with HM features.

4.5.3 Single-label classification

In this experiment, we conduct a single-label classification problem for independently predicting of categories in both *Activity* and *Location* classes. For the comparison, we employ several approaches and report their performances in terms of classification accuracy, the percentage of correctly identified labels. Two strategies for learning the DNN classifier is approached. First, we independently learn two DNN classifiers using data instances of each classes. Note that these DNNs have a single softmax unit representing the class

categories. Second, we employ the exact DNN trained in multi-label classification task and independently test on data instances of each classes. In this scenario, parameters of the DNN is retained unchanged from the multi-label training. For the completeness, we included a support vector classifier (SVM) using radial basis function (RBF) and linear kernels implemented in *scikit-learn* toolkit [80]. Parameters of the optimal SVM classifiers were determined on the training set separately for each fold, via grid search and cross-validation. Table 4.5.3 reports the performance of different classifiers measured in terms of classification accuracy for *Activity* and *Location* classes, respectively. In the table, DNN-SSU and DNN-DSO denote the DNN classifiers with single softmax unit (SSU) and double softmax units (DSU) in the output layer, respectively. Models were trained on three sets of speech features. From the results, it is clear that both versions of the DNN classifier outperform the SVM classifiers. The one exception is the MFCC features, which performed better with SVMs for *location* than DNNs, though well-below the performance of our *HM* features. DNN-DSU slightly improves the performance compare to the DNN-SSU. This might be due to the fact that learning with multi-label data captures correlated information and that improves the performance in single-label classification problem.

Class	Medels	Speech Features		
		MFCC	OS	HM
Location	Chance	45.0	45.0	45.0
	SVM	83.5	71.0	85.0
	DNN-SSU	82.2	78.2	87.1
	DNN-DSU	82.2	79.6	**87.3**
Activity	Chance	31.4	31.4	31.4
	SVM	60.0	54.2	72.6
	DNN-SSU	80.0	78.2	85.7
	DNN-DSU	80.1	81.5	**86.8**

Table 4.4: Comparison of classification accuracy of DNNs with SVM using different features.

4.6 Summary

In this work, we find that DNNs can be employed effectively to infer social contexts from audio snippets of everyday life, achieving classification accuracy as high as about 87.7%

and 86.8% for speakers' *location* and *activity*. We also find that the features extracted using harmonic models are significantly better than MFCC and *OpenSmile* features in these tasks.

Chapter 5

Summary and future work

Numerous attempts have been undertaken during last decades to analyze and characterize speech samples in a range of applications such as, clinical assessment of voice disorders and inferring social context. Accuracy and reliability of acoustic features that are extracted from speech samples play an important role in overall performance of speech systems in these applications. In this thesis, we first developed algorithms for robust and accurate estimation of speech features. Then, these features were employed to build probabilistic speech models for analyzing speech in clinical applications. Note that models were not inspired by the biological process associated with clinical conditions. Instead, we extracted robust and accurate features from speech that is broadly applicable to a variety of clinical problems.

5.1 Summary and contributions of the thesis

In Chapter 2, we described the harmonic model of speech and formulated approaches for accurate and reliable estimation of voiced segments, fundamental frequency, harmonic-to-noise ratio (HNR), jitter, and shimmer.

As the first aim of this thesis, we developed a robust algorithm for detecting voiced segments in adverse noisy conditions. Toward this goal, we adopted the harmonic model of speech to exploit the rich harmonic structure of voiced speech. We overcame the weakness of harmonic model in differentiating speech from stationary harmonic noise using the non-stationary property of speech. Unlike the previous works where model parameters were estimated independently using maximum likelihood framework, we estimated

model parameters more robustly using a maximum *a-posteriori* (MAP) criterion. Given the estimates of model parameters, we detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level was computed by formulating a two-state hidden Markov model (HMM) followed by Viterbi algorithm. We empirically demonstrated the advantage of our algorithm on two data sets. On the Keele dataset under different types and levels of additive noise and on a large collection of 30-seconds samples of ambient recordings of everyday life, we showed that our algorithm performs better than other popular alternatives.

The next aim of this thesis was achieving an accurate and robust estimation of pitch frequency. For this, we adopted the harmonic model of speech and addressed two major problems of this model in the context of pitch estimation: 1) the problem of pitch halving and doubling, and 2) the need to specify the number of harmonics. Like other pitch estimation algorithms, the harmonic model suffers from pitch halving and doubling. We proposed a local smoothing function that exploits the fact that there is more energy in the harmonics near the true pitch than at the corresponding neighborhoods of half or double the pitch. We used a local smoothing function to include this energy and improved the robustness of the pitch candidates in each frame. Also, harmonic model requires specification of the number of harmonics. Although the optimal choice depends on the noise conditions, we showed how the local smoothing could avoid per-frame model order optimization. We adopted a BIC criterion and defined a model complexity that allowed us to estimate the number of harmonics. We estimated the optimal number of harmonics using the average BIC per frame over a small subset of data. We evaluated our proposed pitch estimation method with other state-of-the-art techniques on the Keele dataset in terms of gross pitch error and fine pitch error. Through extensive experiments on several noisy conditions, we demonstrated that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments.

We then introduced alternative techniques for estimating jitter, shimmer, and harmonic-to-noise ratio (HNR) as key features in quantifying voice quality. We employed two versions of harmonic models – a version where amplitude of harmonics are constant, as in

Equation 2.1 and another without that assumption (harmonic model with varying amplitude), as in Equation 2.28. The latter is capable of capturing cycle-to-cycle variations within the frame. We utilized the parameters of both models for estimating jitter, shimmer, and HNR. In order to verify the effectiveness of our voice quality measures, we empirically evaluated them in the context of detecting impaired voices due to Parkinson's disease. For the comparison, we repeated the same experiments with alternative measures extracted by *Praat*, a popular feature extraction tool. Our results showed that voice quality features, extracted by both methods, could detect disordered voices significantly better than chance. We found that HNR is more effective than shimmer and jitter in this task and achieved the best classification accuracy of 72.4%. Also, features exacted by harmonic models, except for jitter with similar performance, outperformed features extracted by *Praat*.

In Chapter 3, we focused on investigating the utility of developed measures in clinical applications. One of the main aims of this thesis is automatically characterizing the severity of Parkinson's disease from speech samples. Toward this aim, we performed empirical studies using speech features and machine learning algorithms on a relatively large sample of 168 subjects, collected from three clinics. We elicited speech using three tasks – the sustained phonation task, the diadochokinetic task, and a reading task prompted by a portable device. From these recordings, we extracted speech features for each subject using *openSMILE*, a standard feature extraction tool. We refined the feature extraction to capture pitch-related cues, including jitter and shimmer, more accurately using harmonic models of speech as described in Chapter 2. We compared the effectiveness of three strategies for learning a regularized regression and found that ridge regression performs better than lasso and support vector regression for our task. Our results showed that the severity of the disease can be inferred from speech with a mean absolute error of about 5.5, explaining 61% of the variance and consistently well-above chance across all clinics. Of the three speech elicitation tasks, we found that the reading task is significantly better at capturing cues than diadochokinetic or sustained phonation task and the combination of all three tasks gives the best results. In all, we demonstrated that the data collection and

inference can be fully automated, and the results showed that speech-based assessment has promising practical application in PD.

Leveraging the same mechanisms developed for inferring the severity of PD, we then focused on detecting children with autism spectrum disorder (ASD) from speech samples. We conducted our experiments on a children's speech corpus provided in *Autism Sub-Challenge* of *Interspeech 2013* including 35 children with ASD and 54 normal subjects. We employed speech measures extracted by harmonic model, along with standard features such as energy, cepstral, and spectral features. Employing a support vector classifier and a support vector regression, we conducted a binary classification task to detect ASD subjects and a 4-way classification task to diagnosis the disease into four sub-categories, respectively. We found that our proposed features improve the performance, measured in terms of unweighted average recall (UAR), of detecting ASD by 2.3% and diagnosing the disorder into four sub-categories by 2.8% compare to a state-of-the-art baseline performance, leading us to win the challenge.

In addition to PD and ASD, we also investigated the use of our features in detection of clinical depression in adolescents. We conducted our empirical experiment on a large corpus of 20-minutes recordings collected from 148 adolescents, including 71 subjects diagnosed with clinical depression and 77 healthy controls, during their family interactions. We employed acoustic features computed from *openSMILE* as well as our developed features to learn SVM classifiers. Beside the acoustic features, we also extracted content information related to affect from manual transcripts using a lookup table. Our experiments demonstrated that our features extracted by the harmonic model outperform features computed by *openSMILE*, on detecting depression in terms of the classification accuracy from 64.7% to 68.7%. Also, combining the textual features with acoustic features provided additional gain, achieving a classification accuracy of 74%.

In Chapter 4, we investigated the problem of detecting social contexts from the audio recordings of everyday life such as in life-logs. Unlike previous clinical applications described in this chapter where we only employed supervised machine learning techniques, in this task we used deep neural networks (DNNs) that would benefit from both supervised and unsupervised learning. The choice of DNNs enabled us to take potential advantages

of available unlabeled data using sparse auto-encoders through an unsupervised learning scheme. We conducted our empirical experiments on a corpus of 30-seconds long recordings of university students recorded every 12 minutes in their daily life. We proposed a multi-label classification scheme using the DNN to detect social contexts, such as speakers location (e.g., in transit) or activity (e.g., watching TV or eating) in this recordings. We first pre-trained the layers of the network in an unsupervised manner using sparse autoencoder. We then modified them in a supervised manner to fine-tune the network for a classification task with limited amounts of labeled data. As other works in this chapter, we employed acoustic features computed by *openSMILE* and our developed features from harmonic model to train the models. From the experiments, we demonstrated that DNNs can be employed effectively to infer social contexts from audio snippets of everyday life, achieving classification accuracy as high as 87.7% and 86.8% for speakers' location and activity. We also showed that features extracted using harmonic model are better than *openSMILE* features in these tasks.

5.2 Future work

This thesis has introduced accurate and robust algorithms for estimating speech features and has explored the use of developed features in a range of clinical applications. Many aspects of this work such as feature extraction and machine learning approaches can be improved. In the following, we describe some suggestions and directions for the future work.

We introduced a Bayesian framework for estimation of harmonic model parameters to exploit the fact that model parameters depend on articulation and cannot vary arbitrarily. The choice of Bayesian priors in our work was limited to the amplitude of harmonics that distributed with a multivariate Gaussian distribution. For future work, we aim to expand Bayesian priors to other parameters of the harmonic model, such as the number of harmonics. We also aim to choose more informative amplitude priors to better fit the empirical distribution of amplitude of harmonics compare to a Gaussian distribution.

The proposed voiced activity detection (VAD) algorithm only employs a single acoustic feature, the likelihood of voicing computed by harmonic model, as an observation probability of the HMM to detect voiced frames. Incorporating other acoustic features, such as spectral entropy, can potentially improve the discrimination power of the VAD. Thus, our future work includes investigating the combination of more VAD features, particularly, those that are robust against stationary harmonic noises.

Our empirical experiments on the Keele dataset verified that our proposed pitch detection method outperforms other state-of-the-art methods. One drawback of our experimental setup is its limitation in amount of speech samples in the Keele dataset – less than six minutes. For the future work, we aim to examine our pitch detector on a larger dataset to understand how well our algorithm can generalize on other datasets.

Moving beyond improving pitch-related features, we investigated the use of developed features in clinical applications. Considerable work still remains for our future work to improve the accuracy of inference. In our studies, information from unvoiced segments have not been fully exploited so far. Similarly, cues from formant trajectories can be useful in quantifying the versatility of speech production, especially function of muscles involved in shaping the oral cavity. In our speech-based assessment of Parkinson's disease (PD), the perceptual characteristics of PD such as imprecise articulation, short rushes of speech, and language impairment are still not modeled in the literature on this topic.

One potential limitation of our methodology in detecting clinical depression of kids is that the speech models can contain parents' speech features due to the poor alignments. In the future work, we aim to employ a speaker identification algorithm to tackle this issue and exclude speech segments belong to parents from speech models.

Finally, our experiments on detecting social contexts using DNNs is far from fully explored yet. There are many questions left in this study and we wish to address them in our future work. Here, we mention a few of these questions: 1) can we improve the performance of pre-training using stacks of restricted Boltzmann machines (RBMs) [120] instead of sparse autoencoders? 2) can other types of non-linearity functions, such as rectified linear units (ReLUs) [127] improve to the performance? and 3) can other methods,

such as *drop-out* [128], reduce the amount of overfitting and lead us to the optimal network architecture?

Bibliography

[1] J. Duffy, *Motor speech disorders: Clues to neurologic diagnosis.* Totowa, NJ: Mayo Foundation for Medical Education and Research, 2000.

[2] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J Speech Hear Res*, vol. 12, no. 2, pp. 246–249, 1969.

[3] L.-S. Low, M. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents speech during family interactions," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 3, pp. 574–586, 2011.

[4] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 17–20.

[5] P. J. Moses, "The voice of neurosis." 1954.

[6] K. Hubbard and D. A. Trauner, "Intonation and emotion in autistic spectrum disorders," *Journal of psycholinguistic research*, vol. 36, no. 2, pp. 159–173, 2007.

[7] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal speech spectrum and increased pitch variability in young autistic children," *Frontiers in human neuroscience*, vol. 4, 2010.

[8] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[9] M. Asgari, I. Shafran, and A. Bayestehtashk, "Robust detection of voiced segments in samples of everyday conversations using unsupervised hmms," in *Spoken Language Technology Workshop (SLT), 2012 IEEE.* IEEE, 2012, pp. 438–442.

[10] M. Asgari and I. Shafran, "Predicting severity of parkinson's disease from speech," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE.* IEEE, 2010, pp. 5201–5204.

[11] M. Asgari and I. Shafran, "Improving the accuracy and the robustness of harmonic model for pitch estimation." in *INTERSPEECH*, 2013, pp. 1936–1940.

[12] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *The Journal of the Acoustical Society of America*, vol. 83, p. 716, 1988.

[13] F. Tolkmitt, H. Helfrich, R. Standke, and K. R. Scherer, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," *Journal of communication disorders*, vol. 15, no. 3, pp. 209–222, 1982.

[14] M. Asgari and I. Shafran, "Extracting cues from speech for predicting severity of parkinson's disease," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 462–467.

[15] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer Speech & Language*, 2013.

[16] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *INTERSPEECH*, 2013, pp. 191–194.

[17] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[18] M. Asgari, G. Kiss, J. van Santen, I. Shafran, and X. Song, "Automatic measurement of affective valence and arousal in speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 965–969.

[19] M. Asgari, I. Shafran, and L. Sheeber, "Inferring clinical depression from speech and spoken utterances," in *IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2014.

[20] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[21] M. Asgari, I. Shafran, and A. Bayestehtashk, "Inferring social contexts from audio recordings using deep neural networks," in *IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2014.

[22] D. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. on Biomedical Engineering*, vol. 41, pp. 663–671, 1994.

[23] I. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, 1996.

[24] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 1769–72.

[25] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76 – 87, 2004.

[26] V. Mahadevan and C. Espy-Wilson, "Maximum likelihood pitch estimation using sinusoidal modeling," in *International Conference on Communications and Signal Processing (ICCSP),*. IEEE, 2011, pp. 310–314.

[27] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 279–290, 2004.

[28] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[29] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering." *EURASIP J. Adv. Sig. Proc.*, vol. 2011, p. 13, 2011.

[30] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[31] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions." in *INTERSPEECH*, 2001, pp. 1887–1890.

[32] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on.* IEEE, 1993, pp. 321–324.

[33] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 139, no. 4, pp. 377–380, 1992.

[34] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[35] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[36] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *Signal Processing Letters, IEEE*, vol. 8, no. 10, pp. 276–278, 2001.

[37] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 498–505, 2003.

[38] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *Signal Processing Letters, IEEE*, vol. 12, no. 10, pp. 689–692, 2005.

[39] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–253.

[40] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 502–510, 2006.

[41] Y. Kida and T. Kawahara, "Voice activity detection based on optimally weighted combination of multiple features." in *INTERSPEECH*, 2005, pp. 2621–2624.

[42] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2624–2633, 2011.

[43] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.

[44] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 845–848.

[45] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, vol. 2, 2007, pp. 556–561.

[46] A. Temko, D. Macho, and C. Nadeu, "Enhanced svm training for robust speech activity detection," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1025.

[47] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, 2013.

[48] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks," in *Proc. Interspeech*, 2013.

[49] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. EUROSPEECH*, 1995, pp. 837–840.

[50] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life," *Journal of Personality and Social Psychology*, vol. 90, no. 5, pp. 862–877, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16737378

[51] P. Boersma and D. Weenink, "Praat speech processing software," *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org.*

[52] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.

[53] K. Sjölander and J. Beskow, "Wavesurfer-an open source speech tool," in *Proceedings of ICSLP*, vol. 4, 2000, pp. 464–467.

[54] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[55] D. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, p. 257, 1988.

[56] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–333.

[57] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," *Proc. Interspeech, Florence, Italy*, pp. 1973–1976, 2011.

[58] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP.)*. IEEE, 2008, pp. 3933–3936.

[59] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America*, vol. 111, p. 1399, 2002.

[60] P. Lieberman, "Some acoustic measures of the fundamental frequency of normal and pathological larnges," *J. Acoust. Soc. Am.*, vol. 35, pp. 344–353, 1963.

[61] I. R. Titze, *Principles of voice production*. Prentice Hall Englewood Cliffs, 1994.

[62] N. Pinto and I. Titze, "Clinical measurement of speech and voice," *J. Acoust. Soc. Am.*, vol. 87, pp. 1278–1289, 1990.

[63] M. Rosa, J. Pereira, and M. Grellet, "Adaptive estimation of residual signal for voice pathology diagnosis," *IEEE Trans. on Biomed. Eng.*, vol. 47, pp. 96–104, 2000.

[64] M. Vasilakis and Y. Stylianou, "A mathematical model for accurate measurement of jitter," in *5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*. Firenze, Italy: Firenze University Press., December 2007.

[65] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 9, 2009.

[66] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech and Hearing Research*, vol. 39, no. 1, p. 126, 1996.

[67] T. Haderlein, E. Nöth, A. Batliner, U. Eysholdt, and F. Rosanowski, "Automatic intelligibility assessment of pathologic speech over the telephone," *Logopedics Phoniatrics Vocology*, vol. 36, no. 4, pp. 175–181, 2011.

[68] E. Nöth, A. Maier, A. Gebhard, T. Bocklet, W. Schupp, M. Schuster, and T. Haderlein, "Automatic evaluation of dysarthric speech and telemedical use in the therapy," *Secretary General: Honorary President*, p. 75, 2011.

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

[70] I. Titz, *Motor and Sensory components of a feedback control model of fundamental frequency.* producting speech: contemporaty issues, 1995.

[71] E. Yumoto, Y. Sasaki, and H. Okamura, "Harmonics-to-noise ratio and psychophysical measurement of the de gree of hoarseness," *J. Speech and Hearing Research*, vol. 27, no. 6, pp. 2–6, 1984.

[72] E. Yumoto, W. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1544–1549, 1982.

[73] C. Ferrer, E. González, M. a. E. Hernández-Díaz, D. Torres, and A. del Toro, "Removing the influence of shimmer in the calculation of harmonics-to-noi se ratios using ensemble-averages in voice signals," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 4:1–4:7, January 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/784379

[74] Y. Qi, "Time normalization in voice analysis," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 5269–5276, 1992.

[75] Y. Qi, B. Weinberg, and N. Bi, "Minimizing the effect of period determination on the computation of amplitude perturbation in voice," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2525–2523, 1995.

[76] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[77] Y. Qi and R. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 537–543, 1997.

[78] D. Krom, "A cepstrum-based technique for determining a harmonics-to-noise r atio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, 1993.

[79] M. Vasilakis and Y. Stylianou, "Spectral jitter modeling and estimation," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 183–193, 2009.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[81] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[82] C. Goetz, W. Poewe, O. Rascol, C. Sampaio, G. Stebbins, S. Fahn, A. Lang, P. Martinez-Martin, B. Tilley, B. Van Hilten *et al.*, "The unified parkinson's disease rating scale (updrs): status and recommendations," *Movement Disorders*, vol. 18, no. 7, pp. 738–750, 2003.

[83] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 16, no. 4, pp. 74–82, 1997.

[84] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 2, pp. 380–384, 2004.

[85] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 4, no. 1, pp. 68–73, 2000.

[86] J. H. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *Biomedical Engineering, IEEE Transactions on*, vol. 45, no. 3, pp. 300–313, 1998.

[87] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, and F. Grenez, "Low-frequency vocal modulations in vowels produced by parkinsonian subjects," *Speech communication*, vol. 50, no. 4, pp. 288–300, 2008.

[88] J. J. Jiang and Y. Zhang, "Chaotic vibration induced by turbulent noise in a two-mass model of vocal folds," *The Journal of the Acoustical Society of America*, vol. 112, p. 2127, 2002.

[89] E. C. Guerra and D. F. Lovey, "A modern approach to dysarthria classification," in *IEEE Conference on Engineering in Medicine and Biology Society (EMBS)*, 2003, pp. 2257–2260.

[90] D. Gil and D. J. Manuel, "Diagnosing parkinson by using artificial neural networks and support vector machines," *Global Journal of Computer Science and Technology*, vol. 9, no. 4, 2009.

[91] R. Das, "A comparison of multiple classification methods for diagnosis of parkinson disease," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, 2010.

[92] T. Bocklet, E. Noth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with parkinson's disease by acoustic, vocal, and prosodic analysis," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 478–483.

[93] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger *et al.*, "Testing objective measures of motor impairment in early parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, 2009.

[94] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge." in *INTERSPEECH*, 2010, pp. 2794–2797.

[95] 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*. ACM, 2010, pp. 1459–1462.

[96] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[97] S. B. Guze, "Diagnostic and statistical manual of mental disorders, (dsm-iv)," *American Journal of Psychiatry*, vol. 152, no. 8, pp. 1228–1228, 1995.

[98] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1328–1342, 2011.

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

[100] D. S. Pine, E. Cohen, P. Cohen, and J. Brook, "Adolescent depressive symptoms as predictors of adult depression: moodiness or mood disorder?" *American Journal of Psychiatry*, vol. 156, no. 1, pp. 133–135, 1999.

[101] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite *et al.*, "Influence of life stress on depression: moderation by a polymorphism in the 5-htt gene," *Science Signaling*, vol. 301, no. 5631, p. 386, 2003.

[102] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 62, no. 6, p. 617, 2005.

[103] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patients speech," *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.

[104] W. A. Hargreaves and J. A. Starkweather, "Voice quality changes in depression," *Language and Speech*, vol. 7, no. 2, pp. 84–88, 1964.

[105] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 9, pp. 1530–1540, 2004.

[106] K. R. Scherer, "Expression of emotion in voice and music," *Journal of voice*, vol. 9, no. 3, pp. 235–248, 1995.

[107] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry*, vol. 4, no. 6, p. 561, 1961.

[108] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *Biomedical Engineering, IEEE Transactions on*, vol. 47, no. 7, pp. 829–837, 2000.

[109] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Content based clinical depression detection in adolescents," *17th EUSIPCO*, pp. 24–28, 2009.

[110] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *Biomedical Engineering, IEEE Transactions on*, vol. 55, no. 1, pp. 96–107, 2008.

[111] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, and G. Parker, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 8022–8026.

[112] H. Hops, *Living in Family Environments (LIFE) coding system: Reference manual for coders.* Oregon Research Institute, 1995.

[113] M. Al Masum Shaikh, M. K. I. Molla, and K. Hirose, "Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense," in *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on.* IEEE, 2008, pp. 294–299.

[114] M. R. Mehl and J. W. Pennebaker, "The social dynamics of a cultural upheaval social interactions surrounding september 11, 2001," *Psychological Science*, vol. 14, no. 6, pp. 579–585, 2003.

[115] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[116] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

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

[118] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[119] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[120] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, p. 1, 2010.

[121] A. Ng, "Cs294a lecture notes: Sparse autoencoder," 2010.

[122] C. M. Bishop *et al.*, *Pattern recognition and machine learning.* springer New York, 2006, vol. 1.

[123] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2004, pp. 22–30.

[124] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[125] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Machine Learning*, vol. 76, no. 2-3, pp. 211–225, 2009.

[126] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[127] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8609–8613.

[128] N. Srivastava, "Improving neural networks with dropout," Ph.D. dissertation, University of Toronto, 2013.