2-2014

Accurate and Robust Models for Clinical Speech Processing

Meysam Asgari

Accurate and Robust Models for Clinical Speech Processing

Meysam Asgari
B.Sc., Amirkabir University of Technology (Tehran Polytechnic), Tehran, 2005
M.Sc., Amirkabir University of Technology (Tehran Polytechnic), Tehran, 2009

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
Master of Science
in
Computer Science & Engineering

February 2014

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

CERTIFICATE OF APPROVAL

This is to certify that the M.Sc. dissertation of
Meysam Asgari
has been approved.

Izhak Shafran, Thesis Advisor
Associate Professor, OHSU

Brian Roark,
Research Scientist, Google Inc.

Alexander Kain
Associate Professor, OHSU

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

Acknowledgment

First of all, my thanks go to my advisor , Dr. Izhak Shafran, for his continuous support and encouragement. I would like to thank all the members of thesis committee, Dr. Brian Roark, Dr. Alexander Kain, and Dr. John Hansen, 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.

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

Accurate and Robust Models for Clinical Speech Processing

Meysam Asgari

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

February 2014
Thesis advisor: Izhak Shafran

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 to develop robust and accurate algorithms for estimating speech features. Researchers have employed a number of techniques in time and frequency domains to estimate, for example, fundamental frequency and harmonic-to-noise ratio (HNR). However, their limitations hinder applications in clinical assessments. Time domain methods often ignore the 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).

The purpose of this study is to achieve accurate and reliable estimation of fundamental frequency, HNR, jitter, and shimmer for clinical speech analysis. Adopting a time-varying harmonic model (TVHM) for representing speech, we quantify hoarseness, a salient feature of PD, as well as jitter and shimmer. We verify our implementation of TVHM and pitch estimation on Keele data set. Results show that pitch detected using TVHM outperforms those from *get-f0* , an algorithm employed in many popular tools (wavesurfer, praat,etc). Further, we demonstrated the utility of our measures for hoarseness, jitter and shimmer in predicting clinical rating of severity of Parkinson's disease.

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 deficit [10]. PD can affect all of the components of speech production including breathing, laryngeal function, articulator movement as well as their coordination for smooth and fluent speech. Resulting dysarthric speech often exhibits monotonous pitch, slurring, reduced stress, inappropriate pauses, variable speech rate, short rushes of speech, harsh voice, imprecise consonant production and breathy voice [8]. Researchers have shown the effects of psychological disorders such as depression in patient's voices [20]. Moreover, a number of studies have shown that the level of emotional excitement changes during speech production [28]. These observations encourage researchers to study about objective measurements using speech parameters that reflect the effects of such disorders. Acoustic features of speech signal including fundamental frequency (f_0), Harmonic-to-Noise Ratio (HNR), shimmer, jitter, and speech rate are used to analyze of pathological voices. These measures can be used to quantify the voice quality, for example, in predicting the severity of PD. Also, a number of techniques use these measures for automatically screening for many neurodegenerative diseases such as Parkinson and Alzheimer.

HNR is a quantity to measure the amount of noise in voice to assess the degree of hoarseness. Jitter and Shimmer refer to a short-term (cycle-to-cycle) perturbation in the f_0 and the amplitude of voice waveform respectively. Perturbation analysis is based on the fact that small fluctuations in frequency, and amplitude of waveform reflect the inherent

noise of voice. However, acoustic analysis of perturbation and HNR is usually dependent on the accurate estimation of f_0.

The main focus of our study is to robustly estimate acoustic features for clinical speech analysis. There are a large number of approaches in time and frequency domain to estimate f_0 and HNR. However, they face limitations for the analysis of disordered voices. Time domain methods often ignore the 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, Parkinsons disease (PD).Adopting a time-varying harmonic model (TVHM) for representing speech, we quantify hoarseness, a salient feature of PD, as well as jitter and shimmer. TVHM exploits the underlying structure of speech production and aims to decompose the speech signal into a harmonic and a non-harmonic component.

Starting with review of traditional acoustic feature extraction techniques in Chapter 2, we will illustrate a model-based approach to quantify voice quality in Chapter3. The model allows robust estimation of HNR, jitter and shimmer. Since, these quantities are difficult to evaluate independently, we evaluate them in the context of predicting clinical assessment of Parkinson's disease as described in **??**. The machine learning experiments and the results are reported and discussed in Chapter 5.

Chapter 2

Review of Traditional Approaches for Acoustic Feature Extraction

2.1 Fundamental Frequency Estimation

Fundamental frequency, also referred as pitch period, is a key feature in speech analysis. Due to important effect 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: (1) pitch candidate generation, in which local pitch candidates are selected from a correlation function that measures the self-similarity, such as autocorrelation function and normalized cross-correlation function; and (2) performing a dynamic programming algorithm, such as Viterbi algorithm to obtain the most probable trajectory of pitch periods among all the candidates. Such methods have been used in standard pitch detector tools such as WaveSurfer [4] and Praat [33]. However, they are sensitive to background noise and their performance significantly drop at low signal-to-noise ratios (SNRs). Tabrikian and his colleagues [32] integrated a Harmonic model with MAP framework, to robustly estimate pitch period at low SNR situations. However, the proposed harmonic model is not able to follow small waveform variations, especially in disordered voices. Adopting a MAP framework, we will modify the introduced harmonic model of Tabrikian [32] to robustly estimate the pitch period for pathological voice analysis.

2.2 Harmonic-to-Noise Ratio Estimation

An accurate estimate of the 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, 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 in the clinic 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 the voice disorders [37]. Generally, we expect the lower HNR in disordered voices rather than the healthy voices [11]. A variety of HNR estimation methods in the studies can be classified into two types: (1) time-domain methods, in which HNR is directly computed from the speech waveform; and (2) 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 [37]. They assume that the 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 large waveform variations, such as in disordered voices.

To overcome these limitations, Qi [24] proposed a time-normalization process using Dynamic Time Warping (DTW), which aims to minimize the effects of f_0 perturbations. DTW is a non-linear time-normalization method, which minimizes the mismatch between the two input frames. It optimally aligns the waveforms prior to computation of the

HNR. However, the time domain HNR estimation requires accurate pitch period estimation. Further, the pitch boundaries are very sensitive to the phase distortion and cause inaccurate HNR estimation. Qi and his colleagues later [26] proposed another appropriate time-normalization technique using zero-phase transformation to minimize the influence of shimmer and jitter on the computation of the HNR.

A number of techniques have been proposed for HNR estimation in the frequency-domain. The main advantage of those methods is less dependency on the accurate estimate of pitch period [25]. Krom [18] 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, which causes the reduction in magnitude of harmonics.

Recently, Asgari and Shafran [1] introduced a model-based framework for HNR estimation. This method focuses on decomposition of voiced speech into a periodic and a non-periodic component. It assumes that a harmonic model approximates the harmonic part of the voiced speech and the the non-harmonic part is obtained by subtracting the harmonic part from the original speech signal. Tabrikian and his colleagues [32] introduced a harmonic model, in which the amplitudes are assumed to be constant. However, this model is not able to follow the amplitude variations within the frame. Asgari and Shafran [1] improved the proposed harmonic model using time-varying amplitudes, which provides more flexibility in capturing sample to sample variations in harmonic amplitudes across the frame. We elaborate this further in Section 3.

2.3 Jitter and Shimmer Estimation

Jitter and shimmer are the prominent acoustic measures that can be 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 [19]. A number of methods have been proposed for the computation of the jitter and shimmer [23, 27]. They

usually employ relative frequency and amplitude differences between consecutive pitch periods for jitter and shimmer estimation. However, these approaches are sensitive to pitch period estimation and their accuracy is a function of the accuracy of the pitch period estimators. Vasilakis and Stylianou [36] proposed a mathematical model for estimation of jitter in frequency domain. Assuming that the magnitude spectrum can be separated into a harmonic part and a sub-harmonic part, they showed that the jitter could be estimated by counting the number of intersections between harmonic and sub-harmonic spectra.

In this study, we will illustrate a model-based approach for jitter and shimmer estimation proposed by Asgari and Shafran [1].

Chapter 3

Model-based Acoustic Feature Extraction

3.1 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. 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. From a signal processing point of view, speech production process can be modeled by a linear system as shown in figure 3.1. 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 account by $\mathbf{L(z)}$. 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 have been successfully employed for periodic signal [6] and in the next subsection, we develop the model for our context.

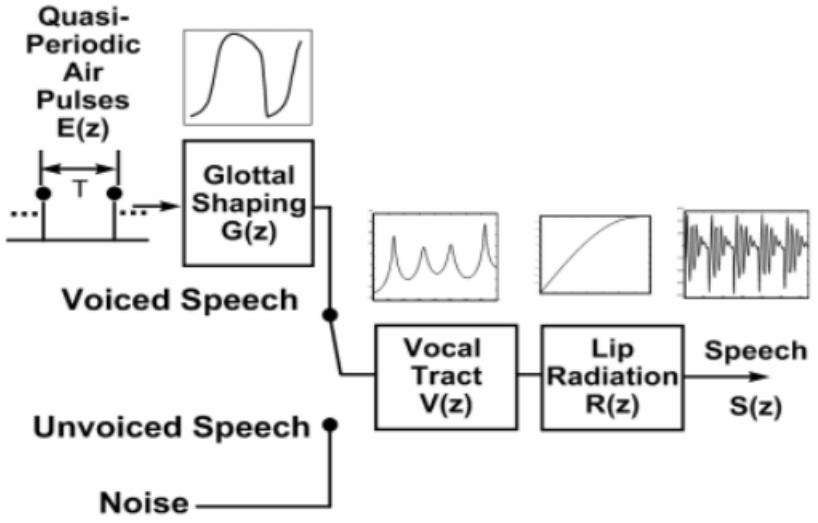

Figure 3.1: A computational model of speech production.

3.2 Time-Varying Harmonic Model

The Harmonic Model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related, i.e., the frequencies of the sinusoids are multiples of the fundamental frequency, f_0. This assumption arises from the harmonic nature of the speech signal and reduces the number of parameters in general sinusoidal model. Stylianou [31] introduced a Harmonic plus Noise Model (HNM) for speech analysis and synthesis, in which speech signals are represented as a time-varying harmonic component plus a modulated noise component. The harmonic part accounts for the periodic component of the speech signal while the noise part accounts for its non-periodic components. Speech decomposition using a HNM is useful for applications in speech synthesis, voice conversion, speech enhancement, and speech coding.

3.2.1 Model Description

Let $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_N)]^T$ denote the speech samples in a voiced frame, measured at times $t_1, t_2, \ldots, t_T$. 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:

$$s(t) = a_0 + \sum_{h=1}^{H} a_h(t)cos(2\pi f_0 ht) + b_h(t)sin(2\pi f_0 ht) \tag{3.1}$$

$$y(t) = s(t) + n(t) \tag{3.2}$$

where H denotes the number of harmonics and $2\pi f_0$ stands for the fundamental angular frequency. The amplitude of cosine components, $a_h(t)$, and sine components, $b_h(t)$ are not constant across the analysis frame. Due to the fact that vocal tract and lip radiation transfer functions vary much slower than the frequency of glottal pulse excitation, $a_h(t)$ and $b_h(t)$ can effectively capture sample to sample variation in harmonic amplitude within the frame. We represent the amplitudes of the sinusoidal components as a linear combination of a few local basis functions as follow:

$$a_h(t) = \sum_{i=1}^{I} \alpha_{i,h}\psi_i(t) \ , \ b_h(t) = \sum_{i=1}^{I} \beta_{i,h}\psi_i(t) \tag{3.3}$$

where $\psi_i(t)$, $i = 1, \ldots, I$ are set of smooth basis functions that can be obtained by translating in time a prototype of any convenient function $\psi(t)$. In this work, we use four ($I = 4$) Hanning windows as basis functions, which were centered on 0, $M/3$, $2M/3$, and M with an overlap of $M/3$ with adjacent basis functions and length of $2M/3$ where M is the analysis window length. Figure 3.2 shows a representation of amplitude of a harmonic component obtained by combination of four basis functions.

The signal $s(t)$ can be expressed as a linear combination of harmonic components and coefficients of basis functions.

$$s(t) = a_0 + \sum_{h=1}^{H} [\psi(t)cos(2\pi f_0 ht) \quad \psi(t)sin(2\pi f_0 ht)] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{3.4}$$

$$\alpha = [\ \alpha_{1,1} \quad \cdots \quad \alpha_{I,H}\ ]^T \ , \ \beta = [\ \beta_{1,1} \quad \cdots \quad \beta_{I,H}\ ]^T$$

$$\psi(t) = [\ \psi_1(t) \quad \cdots \quad \psi_I(t)\ ]^T$$

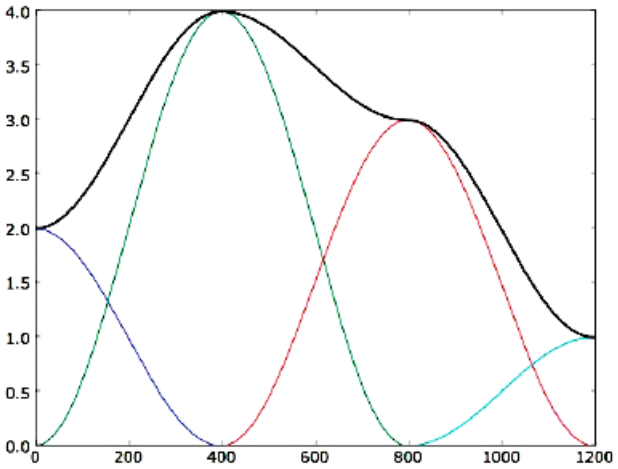

Figure 3.2: An illustration of time-varying amplitude of a harmonic component modeled as a superposition of four bases functions spanning the duration of the frame.

The harmonic signal can be factored into coefficients of basis functions, α, β, and the harmonic components which are determined solely by the given angular frequency $2\pi f_0$ and the choice of the basis function $\psi(t)$.

$$s(t) = [1 \quad A_c(t) \quad A_s(t)] \begin{bmatrix} a_0 \\ \alpha \\ \beta \end{bmatrix} \tag{3.5}$$

$$A_c(t) = [\psi(t)cos(2\pi f_0 t) \quad \cdots \quad \psi(t)cos(2\pi f_0 Ht)]$$

$$A_s(t) = [\psi(t)sin(2\pi f_0 t) \quad \cdots \quad \psi(t)sin(2\pi f_0 Ht)]$$

Stacking rows of $[1 \quad A_c(t) \quad A_s(t)]$ at $t = 1, \cdots, T$ into a matrix $\mathbf{A}$, equation (2) can compactly represented in matrix notation as:

$$\mathbf{y} = \mathbf{A}\,\mathbf{b} + \mathbf{n} \tag{3.6}$$

where $\mathbf{y} = \mathbf{A}\,\mathbf{b}$ corresponds to a basis function expansion of the harmonic part of voiced frame in terms of windowed sinusoidal components. We assume the distribution of noise is constant during the frame and can be modeled by a zero-mean Gaussian noise with unknown variance σ^2), $\mathbf{n} \sim N(0, \sigma^2)$, which is required to be estimated.

3.2.2 Parameter Estimation

The model analysis consists of estimation of the parameters of harmonic and noise part. The unknown parameters in the model described in (3.2) are: fundamental frequency, f_0, the vector of coefficients of basis functions,$\mathbf{b}$, and the noise variance, σ^2. The number of basis functions and harmonics are assumed to be known. Assuming a voiced frame, we first estimate the f_0 and noise intensity.

Frame Level Maximum Likelihood Estimation: Assuming the noise samples $\mathbf{n}$ in equation (3.2) are independent and identically distributed random variables, with zero-mean Gaussian distribution the likelihood function of the observed vector, $\mathbf{y}$, given the model parameters is as follow:

$$p(\mathbf{y} \mid f_0, \mathbf{b}, \sigma^2) = (2\pi\sigma^2)^{-D/2} exp(-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{Ab}||^2) \tag{3.7}$$

where D stands for dimension of observed vector, $\mathbf{y}$. Note that the only unknown parameter of matrix $\mathbf{A}$ is fundamental frequency, f_0.

The objective is to estimate the unknown parameters. We employ Maximum Likelihood estimator (ML) to maximize the log-likelihood function with respect to unknown parameters [32]. The log-likelihood function is defined as:

$$\mathbf{L}(f_0, \mathbf{b}, \sigma^2) = -\frac{D}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{Ab}||^2 \tag{3.8}$$

Note that the vector of harmonic coefficients, $\mathbf{b}$, and matrix $\mathbf{A}$, are independent from each other. We first maximize the log-likelihood function with respect to $\mathbf{b}$ and one obtains:

$$\hat{\mathbf{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \tag{3.9}$$

We then substitute $\hat{b}$ into equation (3.8) and the log-likelihood function can be written as follow:

$$\mathbf{L}(f_0, \ \hat{\mathbf{b}}, \sigma^2) = -\frac{D}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{P})^T(\mathbf{y} - \mathbf{P}) \tag{3.10}$$

where $\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the function of f_0. To maximize equation (3.10) with respect to f_0, we can rewrite equation (3.10) as:

$$\mathbf{L}(f_0) = const + \mathbf{y}^T\mathbf{P}\mathbf{y} \tag{3.11}$$

$$\hat{f}_0 = \arg\max_{f_0}\{\mathbf{y}^T\mathbf{P}\mathbf{y}\} \tag{3.12}$$

Maximizing f_0 in equation (3.12) requires a search over the coordinates of f_0 to find the global maximum.

To estimate unknown noise variance, the log-likelihood function in equation (3.8) needs to be maximized also with respect to σ^2. Taking derivative with respect to σ^2 and making it zero leads to:

$$\hat{\sigma^2} = \frac{1}{D}\mathbf{y}^T(\mathbf{I} - \hat{\mathbf{P}})\mathbf{y} \tag{3.13}$$

where $\mathbf{I}$ is a identity matrix.

Multiple Frame Pitch Tracking: Single frame estimation of f_0 sometimes leads to f_0 halving and doubling estimates. To better estimate the f_0 trajectory, a common approach is to use dynamic programming to find an optimal f_0 trajectory for sequence of frames [32]. Lets define $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_M\}$ and $\mathbf{F}_0 = \{{f_0}^1, {f_0}^2, \cdots, {f_0}^M\}$ as a sequence of M consecutive voiced frames and their corresponding f_0 trajectory respectively. Assuming that $\mathbf{y}_i$ are independent of each other, the conditional probability of data vector, $\mathbf{Y}$, given the vector of $\mathbf{F}_0$ can be expressed as:

$$p(\mathbf{Y}|\mathbf{F}_0) = \prod_{i=1}^{M} p(\mathbf{y}_i|f_0^i) \tag{3.14}$$

According to the Bayes rule, we can drive the posterior probability as:

$$p(\mathbf{F}_0|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{F}_0)p(\mathbf{F}_0)}{p(\mathbf{Y})} \tag{3.15}$$

The Maximum *A Posteriori* (MAP) estimation of f_0 is then obtained by maximizing the following equation:

$$\hat{F}_0 = \arg\max_{\mathbf{F}_0}\{p(\mathbf{Y}|\mathbf{F}_0)p(\mathbf{F}_0)\} \tag{3.16}$$

The vector of fundamental frequency, $\mathbf{F}_0$ can be treated as a first order Markov process by assuming that the probability of the f_0 at a frame depends only on the f_0 in the previous frame, and it can be approximated using a Gaussian distribution.

$$p(\mathbf{F}_0) = p({f_0}^{(1)}, {f_0}^{(2)}, \cdots, {f_0}^{(M)}) = p({f_0}^{(1)}) \prod_{m=2}^{M} p({f_0}^{(m)}|{f_0}^{(m-1)}) \tag{3.17}$$

$$p({f_0}^{(m)}|{f_0}^{(m-1)}) \sim N({f_0}^{(m-1)}, \sigma_t) \tag{3.18}$$

where $p({f_0}^{1})$ is the prior probability function of f_0 at the first frame. Substituting (3.10) and (3.18) into (3.16) and taking the logarithm leads to:

$$\hat{F_0} = \arg\max_{\mathbf{F}_0} \sum_{m=1}^{M} [\mathbf{L}({f_0}^{(m)}, \ \hat{\mathbf{b}}_m, \hat{\sigma^2}) + \log p({f_0}^{(m)}|{f_0}^{(m-1)})] \tag{3.19}$$

Maximizing $\hat{\mathbf{F}}_0$ requires a multidimensional search over the possible f_0 values across the whole frames, which is not a computationally feasible task. As it can be seen in (3.19), f_0 trajectory estimation consists of simultaneously maximizing the likelihood function, and the log of the transition probability function between the states. So, we can employ a Hidden Markov Model (HMM), in which the log of observation and emission probabilities are computed by $\mathbf{L}({f_0}^{(m)}, \ \hat{\mathbf{b}}_m, \hat{\sigma^2})$, and $\log p({f_0}^{(m)}|{f_0}^{(m-1)})$ respectively.

$$\log p({f_0}^{(m)}|{f_0}^{(m-1)}) = -\frac{1}{2} log(2\pi\sigma_t^2) - \frac{1}{2\sigma_t^2}({f_0}^{(m)} - {f_0}^{(m-1)})^2 \tag{3.20}$$

The states in the HMM represent the possible discrete values of f_0 ranging from 50 Hz to 500 Hz. Finally, we use a Viterbi algorithm to find the optimal state sequence through this trellis of states. The Viterbi path is most likely hidden states, which in our case is fundamental frequency.

3.2.3 Pitch Tracking Evaluation

To verify our implementation of TVHM for pitch estimation, the performance of the proposed algorithm is evaluated and compared to *get-f0* , an algorithm employed in many popular tools (wavesurfer, praat,etc). For all the experiments, we used Keel pitch reference database [21], which is available online. It contains 10 files from 10 speakers (five males and five females), each 35s long. It provides a reference pitch, which is obtained from a

recorded laryngograph. The performance of each estimators in terms of mean absolute error (MAE) and gross error rate (GER) are reported in table 3.1. GER is defined as percentage of pitch estimates that deviate more than 20% of the ground truth. The voiced boundaries are assumed to be known and all the comparisons are performed on voiced frames. The audio files are contaminated with the additive white Gaussian noise (AWGN) at different SNRs. The mean of computed errors for all the files are reported in the table 3.1. The results proves the robustness of the proposed method for severe noise conditions.

Table 3.1: Mean absolute error (MAE) and gross error rate (GER) in estimated pitch

SNR (dB)	$\text{MAE}_{(Hz)}$ ($\text{GER}_{(\%)}$)	
	get-f$_0$	TVHM
0	11.12(17.20)	8.45(5.23)
5	7.33(14.44)	4.12(4.53)
10	4.65(10.23)	3.73(4.23))
15	3.07(5.56)	3.70(3.12))
No additive noise	3.12(2.8)	2.67 (2.21)

3.2.4 Harmonic-to-Noise Ratio

Estimating the unknown parameters of TVHM in previous subsection enables us to compute the Harmonic-to-Noise Ratio (HNR). Given an estimate of fundamental frequency, the vector $\mathbf{b}$ that contains all the coefficients of basis functions can be estimated as $\hat{\mathbf{b}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$. However, ML estimation of parameter vector $\mathbf{b}$ may leads to overfitting. The robustness of the estimates can be improved using the prior statistical information regarding to the shape of vocal tract, in which the amplitudes of the harmonics are not allowed to vary in an arbitrary subspace.

We can integrate this additional knowledge by adding a regularization term to equation (3.10), which restricts the parameters into a limited subspace. For computational convenience, we chose the L2 regularization term, $||\mathbf{b}||_2$, to obtain the closed form solution $\mathbf{b} = (\mathbf{A^T A} + \lambda\mathbf{I})^{-1}\mathbf{A^T y}$ where higher λ increases the weight on the regularization term. From a Bayesian point of view, adding a penalty term is equivalent to imposing a

prior distribution on model parameters. As we pointed out in the introduction section, noise presence in a voiced utterance can be regarded as incomplete physiological closure of vocal folds. As such, our focus is to separate the contribution of the two noisy and harmonic sources in order to quantify the degradation in voice quality. Given the estimate of fundamental frequency at m-th frame, $f_0^{(m)}$, and the corresponding vector of basis functions, $\mathbf{b}_m$, we can reconstruct the signal in equation (3.1) by

$$\hat{\mathbf{s}_m} = \mathbf{A}(f_0^{(m)})\hat{\mathbf{b}}_m \ , \quad m = 1, \ldots, M \tag{3.21}$$

where $\hat{\mathbf{s}_m}$ denotes the reconstructed signal at m-th frame.

Given the reconstructed signal as the harmonic source of vocal tract, the noisy part is obtained by subtracting the reconstructed signal from the original speech signal. The noisy part encompasses everything in the signal that is not described by harmonic components including the frication noise, the waveform fluctuations, etc. Figure 3.3 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 more flexible and it is able to follow sample-to-sample variations not only in amplitude but also variation in pitch to a certain extent.

According to Parseval's theorem, HNR and the ratio of energy in first and second harmonics can be computed from the time-varying amplitudes of the harmonic components.

$$c_h(t) = \sqrt{\sum_{i=1}^{I} a_h(t)^2 + b_h(t)^2} \tag{3.22}$$

$$HNR = \log \sum_{t=1}^{N} \sum_{h=1}^{H} c_h(t)^2 - \log \sum_{t=1}^{N} (y(t) - s(t))^2 \tag{3.23}$$

$$H12 = \log \sum_{t=1}^{N} c_1(t)^2 - \log \sum_{t=1}^{N} c_2(t)^2 \tag{3.24}$$

3.2.5 Shimmer

Shimmer is defined as the variation in amplitude between the adjacent cycles of the glottal waveform. From a point of view, it can be referred as a slow amplitude modulation

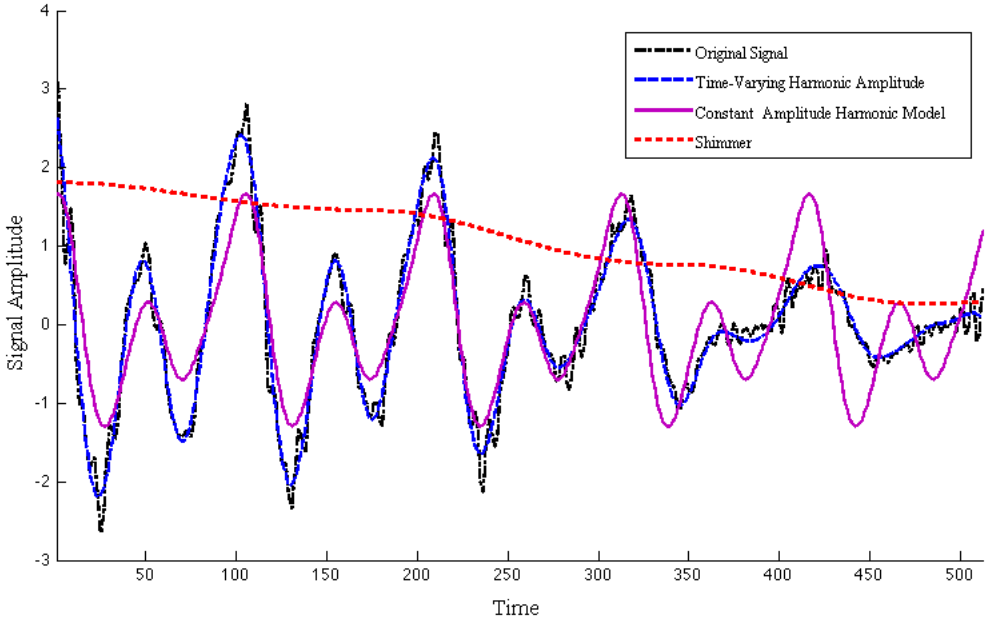

Figure 3.3: An example speech frame (black), estimated signal from harmonic model with time-varying amplitude (blue), estimated signal from harmonic model with constant amplitude (purple), estimated shimmer (red).

(AM) of glottal waveform due to the inability of humans to keep constant the tension of their vocal folds [34]. In order to compute shimmer in output waveform, we first estimate a prototype waveform using all the observed signals in the frame. This can be easily computed from the harmonic model by assuming the amplitudes of the harmonic components are constant across the frame.

$$s(t) = a_0 + \sum_{h=1}^{H} a_h cos(2\pi f_0 ht) + b_h sin(2\pi f_0 ht),\ c_h = \sqrt{\sum_{i=1}^{H} a_h^2 + b_h^2} \tag{3.25}$$

where c_h denotes the amplitude of the harmonic components obtained using a maximum likelihood framework. Now, shimmer can be considered as a function $f(t)$ that scales the amplitudes of all the harmonics in the time-varying model. From another point of view, $f(t)$ can be regarded as the envelop of speech waveform extracted by AM demodulation.

$$c_h(t) = c_h f(t) + e(t), \quad t = 1, \ldots, T\ , h = 1, \ldots, H \tag{3.26}$$

where e(t) is assumed to be uncorrelated noise. We estimate $f(t)$ using maximum likelihood criterion as follow:

$$\hat{f}(t) = \frac{\sum_{h=1}^{H} c_h c_h(t)}{\sum_{h=1}^{H} c_h^2} \tag{3.27}$$

Figure 3.3 illustrates an example frame where the solid red line shows the estimated shimmer and blue line is the 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.

3.2.6 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 them [38]. Analysis of jitter is based on the accurate estimation of pitch period. Given an estimate of the average pitch period of the frame $(1/f_0)$, 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. This matched filter is then convolved with the estimated signal and the distance between the maximas 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. Thus, we compute jitter quantitatively on any voiced signal, unlike many previous techniques were jitter could be computed only in specially elicited signals (e. g. phonation task).

Chapter 4

Assessing the severity of Parkinson's disease

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 of the components of speech production including breathing, laryngeal function, articulator movement, and also their coordination for smooth speech. Resulting dysarthric speech often exhibits monotonous pitch, slurring, reduced stress, inappropriate pauses, variable speech rate, short rushes of speech, harsh voice, imprecise consonant production, and breathy voice [8]. 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.

4.1 speech-based PD diagnosis

There have been extensive studies on employing automatic speech processing for assessing voice disorders [5, 13, 15, 16] and in particular, classifying PD subjects from controls or inferring the severity of the diseases [7, 2, 17]. Researcher have used a range of machine learning techniques for diagnosing PD. Gil and his colleagues [12] 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

dataset composed of 31 people, 23 with Parkinson's disease and achieved a high accuracy of around 90%. Another study by Dus [9] attempted to compare different types of classification methods for diagnosis of PD on the same dataset. He employed four different classifiers: neural networks, DMneural, regression, and decision tree where the neural network classifier yielded the best score at 92.9% classification accuracy. More recently, Bocklet and colleagues applied a more rigorous machine learning approach to classify PD subjects from control [3]. 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 fairly 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.

Chapter 5

Experimental Paradigm

Empirical evaluation reported in this study were performed on data collected from 116 clinical assessment from 82 subjects, including 21 controls, through two clinics, namely OHSU and Parkinson's Institute. Subjects were asked to perform 3 tasks designed to exercise different aspects of speech and non-speech motor control: (1) sustained phonation task where subjects were instructed to phonate the vowel /a/ in a clear and steady voice as long as possible; (2) Diadochokinetic (DDK) task where subjects are 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; and (3) Reading task where subjects are asked to read standard passages.

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 [22]. 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. Most clinical ratings of speech pathologies such as hypokinetic dysarthria in PD are based on perceptions of trained clinicians. For automating assessment, we adopt a machine learning approach, where we define a large number of features that can be reliably extracted from speech and let the learning algorithm pick out the features that are most useful in predicting the clinical rating.

5.1 Features

As in most speech processing systems, we extract 32 millisecond long frames using a Hanning window at a rate of 100 frames per second before computing the following features.

1. **Pitch:** One of the key features in frequency domain is pitch, which can be extracted using a standard pitch tracking algorithm such as *get-f0* . The estimated pitch is also used to estimate the harmonic model mentioned earlier.

2. **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. **Cepstral Coefficients:** Shape of the spectral envelope is extracted from cepstral coefficients. Thirteen cepstral coefficients of each frame were augmented with their first- and second-order time derivatives.

4. **Segmental Duration and Frequency:** 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:** We compute HNR, the ratio of energy in first to second harmonics, jitter and shimmer, as described earlier.

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 voiced (unvoiced) segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. Speech pathologists often plot and examine the inter- action between quantities such as pitch and energy to fully understand

the capacity of speech production [34]. We capture such interactions by computing the covariance matrix (upper triangular elements) of frame-level feature vectors over voiced (unvoiced) segments. The segment-level duration statistics including mean, median, variance, minimum and maximum were computed for both voiced and unvoiced regions. The three kinds of summary features were concatenated into a global feature vector for each subject. There has been suggestion that many speech features are better represented in log domain. So, we performed experiments by augmenting the global feature vector with its mirror in log domain. The resulting features were computed separately for the three elicitation tasks (phonation, DDK and reading) and augmented into one vector, up to 17K long, for each subject.

5.2 Regression Model

The motor sub-scale of UPDRS (mUPDRS) was predicted from extracted speech features using several regression models estimated by support vector machines. Epsilon-SVR and nu-SVR were employed using several kernel functions including polynomial, radial basis function and sigmoid kernels [29]. The models were evaluated using a 20-fold cross-validation and the results were measured using mean absolute error (MAE).Not all the features extracted from speech are expected to be useful and in fact many are likely to be noisy. We apply standard feature selection algorithm over training folds and evaluate several models using cross-validation to pick the one with optimal performance. One weakness of most feature selection algorithm is that they compute the utility of each element separately and not over subsets. For understanding the contribution of the different features, we introduced them incrementally and measured performance, as reported in Table 5.1. The first regression model was estimated with frequency-domain, temporal-domain and cepstral-domain features. Subsequently, log space features, segmental durations and harmonicity were introduced.

The baseline system contains features related to pitch, spectral entropy and cepstral coefficients, in all about 7K features per subject. From among these features, automatic feature selection picks about 800 features to predict the UPDRS scores with an MAE of

Table 5.1: Mean absolute error (MAE) measured on a 20-fold cross-validation for predicting severity of Parkinsons disease (mUPDRS) from speech

	Speech Features	Features	MAE
(a)	Baseline	7K	6.14
(b)	(a) + log-space	14K	6.06
(c)	(b) + duration	14K	5.85
(d)	(c) + HNR + $H1/H2$	15K	5.81
(e)	(d) + jitter + shimmer	15K	5.66

about 6.14 and a standard deviation of about 2.63. Recall that guessing the mean UPDRS score on this data incurs an MAE of about 9.0. As a check for overfitting, we shuffled the labels, selected features and then learned the regression using the same algorithms. The resulting models performed significantly worse, at about 8.5 MAE. To put the reported results in the right perspective, studies show that the clinicians do not agree with each other completely and attain a correlation of about 0.82 and commit an error of about 2 points. The improvement in prediction with the baseline model is statistically significant. The mapping of features in the log-space provides a small and consistent gain, but not as large as the ones reported in [35] whose experimental setup (utterance- level test vs. train split, not subject-level), number of subjects (only 42), features and models are significantly different from ours. The frequency and duration of voiced segments proved to be useful cues in predicting mUPDRS as expected from clinical observations [30]. Finally, the HNR and the ratio of energy in first to second harmonic estimated using the algorithm proposed in this paper provides further improvement in predicting mUPDRS. The gains from harmonicity are consistent with previous studies on classification of dysarthria [14]. Among all combination of features listed in the table, the size of the optimal feature set was about 550 features for model (e). The best performance was consistently obtained with epsilon SVR using 3rd degree polynomial kernel functions.

Chapter 6

Conclusion

This study describes a computational approach for quantifying perceptual voice qualities such as breathy and hoarseness. We focused to develop robust and accurate algorithms for estimating speech features. Starting with review of traditional acoustic feature extraction techniques, we illustrated a model-based approach based on a computational model of speech production. We solved the problem of parameter estimation using a maximum likelihood framework for voiced speech. We then employed this model to robustly estimate fundamental frequency, harmonic-to- noise ratio (HNR), jitter and shimmer. We evaluated the performance of estimated pitch using Keel Pitch Reference database at different noisy conditions. We evaluated other estimated quantities in the context of predicting clinical assessment of Parkinsons disease. These features are exploited along with energy, spectrum, cepstrum and segmental features in a support vector machine based regression model. The epsilon support vector machines with polynomial kernel of degree 3 was found to be most effective, whose performance was about 5.66 mean absolute error as measured on a 20-fold cross-validation.

For the future works, we will integrate pitch estimates of proposed model into the feature set and use it to estimate other parameters of harmonic model. Also, the experiments will be performed on more subjects.

6.1 Acknowledgement

We would like to thank L. Holmstrom, K. Kubota and J. McNames for facilitating the study, designing the data collection and making the data available to us. We are extremely

grateful to our clinical collaborators M. Aminoff, C. Cristine, J. Tetrud, G. Liang, F. Horak, S. Gunzler, and B. Marks for performing the clinical assessments and collecting the speech data from the subjects.

Bibliography

[1] Asgari, M., and Shafran, I. Extracting cues from speech for predicting severity of parkinsonś disease. In *Machine Learning for Signal Processing (MLSP)* (2010), pp. 462–467.

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

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

[4] Boersma, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Institute of Phonetic Sciences* (1993), pp. 97–110.

[5] Boyanov, B., and Hadjitodorov, S. Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases. *Engineering in Medicine and Biology Magazine, IEEE 16*, 4 (1997), 74–82.

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

[7] Cnockaert, L., Schoentgen, J., Auzou, P., Ozsancak, C., Defebvre, L., and Grenez, F. Low-frequency vocal modulations in vowels produced by parkinsonian subjects. *Speech communication 50*, 4 (2008), 288–300.

[8] Darley, F. L., Aronson, A. E., and Brown, J. R. Differential diagnostic patterns of dysarthria. *J Speech Hear Res 12*, 2 (1969), 246–249.

[9] Das, R. A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications 37*, 2 (2010), 1568–1572.

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

[11] Ferrer, C., González, E., Hernández-Díaz, M. E., Torres, D., and del Toro, A. Removing the influence of shimmer in the calculation of harmonics-to-noise ratios using ensemble-averages in voice signals. *EURASIP J. Adv. Signal Process 2009* (January 2009), 4:1–4:7.

[12] Gil, D., and Manuel, D. J. Diagnosing parkinson by using artificial neural networks and support vector machines. *Global Journal of Computer Science and Technology 9*, 4 (2009).

[13] Godino-Llorente, J. I., and Gomez-Vilda, P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *Biomedical Engineering, IEEE Transactions on 51*, 2 (2004), 380–384.

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

[15] Hadjitodorov, S., Boyanov, B., and Teston, B. Laryngeal pathology detection by means of class-specific neural maps. *Information Technology in Biomedicine, IEEE Transactions on 4*, 1 (2000), 68–73.

[16] Hansen, J. H., Gavidia-Ceballos, L., and Kaiser, J. F. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. *Biomedical Engineering, IEEE Transactions on 45*, 3 (1998), 300–313.

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

[18] Krom, D. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research 36* (1993), 254–266.

[19] Lieberman, P. Some acoustic measures of the fundamental frequency of normal and pathological larnges. *J. Acoust. Soc. Am. 35* (1963), 344–353.

[20] Low, L., Maddage, M., Lech, M., Sheeber, L., and Allen, N. Detection of clinical depression in adolescents speech during family interactions. *IEEE Trans. on Biomedical Engineering 58*, 3 (2011), 547–586.

[21] Meyer, G. F. Keele pitch database.

[22] Movement Disorder Society Task Force on Rating Scales for Parkinson Disease. The Unified Parkinson Disease Rating scale (UPDRS): status and recommendations. *Mov, Disord 18*, 7 (Jul 2003), 738–750.

[23] Pinto, N., and Titze, I. Clinical measurement of speech and voice. *J. Acoust. Soc. Am. 87* (1990), 1278–1289.

[24] Qi, Y. Time normalization in voice analysis. *J. Acoust. Soc. Am. 92*, 5 (1992), 5269–5276.

[25] Qi, Y., and Hillman, R. Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *J. Acoust. Soc. Am. 102*, 1 (1997), 537–543.

[26] Qi, Y., Weinberg, B., and Bi, N. Minimizing the effect of period determination on the computation of amplitude perturbation in voice. *J. Acoust. Soc. Am. 97*, 4 (1995), 2525–2523.

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

[28] Scherer, K. Expression of emotion in voice and music. *J. Voice 9*, 3 (1995), 235–248.

[29] Scholkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. New support vector algorithms. *Neural Comput 12*, 5 (2000), 1207–1245.

[30] Skodda, S., Flasskamp, A., and Schlegel, U. Instability of syllable repetition as a model for impaired motor processing: is parkinson disease a rhythm disorder? *J Neural Transm,* (Mar 2010).

[31] Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. on Speech and Audio Processing 9*, 1 (2001), 21–29.

[32] Tabrikian, J., Dubnov, S., and Dickalov, Y. Maximum a posteriori probability pitch tracking in noisy environments using harmonic model. *IEEE Trans. on Speech and Audio Proc 12*, 1 (1991), 76–87.

[33] Talkin, D. A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis, 87* (1995), 495–518.

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

[35] Tsanas, A., Little, M., McSharry, P., and Ramig, L. O. Enhanced classical dysphonia measures and sparse regression for telemonitoring of parkinson disease progression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2010), pp. 594–597.

[36] Vasilakis, M., and Stylianou, Y. A mathematical model for accurate measurement of jitter. In *5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (December 2007).

[37] Yumoto, E., Gould, W., and Baer, T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am. 71*, 6 (1982), 1544–1549.

[38] Yumoto, E., Sasaki, Y., and Okamura, H. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *J. Speech and Hearing Research 27*, 6 (1984), 2–6.