36th Asilomar Conf. Signals, Systems & Computers, 2002

A System for Automatic Detection of Pathological Speech

Alireza Afshordi Dibazar[1], Shikanth Narayanan[2]

[1]Biomedical Engineering. Department, University of Southern California, CA, USA
[2] Electrical Engineering Department, University of Southern California, CA, USA
Email: dibazar@usc.edu, shri@sipi.usc.edu

ABSTRACT

This study focuses on a robust, rapid and accurate system for automatic detection of normal voice and speech pathologies. This system employs non-invasive, non-expensive and fully automated measures of vocal tract characteristics and excitation information. Mel-frequency filterbank cepstral coefficients and measures of pitch dynamics were modeled by Gaussian mixtures in a Hidden Markov Model (HMM) classifier. The method was evaluated using the sustained phoneme /a/ data obtained from over 700 subjects of normal and different pathological cases from the Massachusetts Eye and Ear Infirmary (MEEI) database. This method attained 99.44% correct classification rates for discrimination of normal and pathological speech for sustained /a/. This represents 8% detection error rate improvement over the best performing classifier using carefully measured features prevalent in the state-of-the-art in pathological speech analysis. The result of this method also shows significant improvement in detection of different A-P-Squeezing pathology with respect to the other methods.

Keywords: pathological speech, HMM, Mel frequency filters

I. INTRODUCTION

There are numerous medical conditions that adversely affect the human voice. Many of these conditions have their origins primarily in the vocal system and the available tools for detection of speech pathologies are either invasive or require expert analysis of numerous human speech signal parameters. So, a reliable, accurate and non-invasive automatic system for recognizing and monitoring speech abnormalities is one of the necessary facilities in pathological speech assessment. Automatic voice analysis for pathological speech have several advantages: 1) It has quantitative and non-invasive nature. 2) Allows identification and monitoring of the onset of vocal system diseases. 3) Reduces analysis cost and time.

In previous studies, several methods for assessing speech pathologies have been introduced. In general, these methods, based on features they use, fall into two groups: Spectral envelope measures and temporal dynamic measures. In the spectral analysis methods, researchers have tried to keep track of the spectral variations of signal such as amplitude, bandwidth and frequency of formants including sub-band processing methods. In time domain, authors have employed two major methods: 1) Methods based on temporal measurements of signal and their statistics, such as average pitch variation, jitter, shimmer, etc 2) analysis of residual of inverse filtering [1] of the speech signal, which corresponds to an estimate of the source excitation, to distinguish between normal and pathological speech. However, there are some difficulties associated with these methods. As it has been reported in published articles, these methods have accuracies between 80 to 90% [2], typically with a small limited number of subjects. In addition, there are robustness, consistency and complexity difficulties for measuring those features including the degree of human intervention needed in the measurements.

It has been shown that spectral measurements can identify diseases like some kinds of Cysts, Polyps and other diseases, which indicate vocal fold malfunctioning. It is has also reported that one of the characteristics of speech signals under a non-healthy [3] condition is an irregular periodicity. Since this is essentially a function of excitation of the speech production system, it is necessary to incorporate measures that focus on excitation characteristics.

In the production of short vowels, the poor control of respiratory system is not significant [4]; hence, vowels phonated in a sustained fashion with comfortable levels of pitch and loudness are interesting and useful from a clinical point of view. The results of study, which conducted by Vieira et al [5], showed that there is consistency between electro glottal graph (EGG) parameters and acoustic signal features of sustained vowel /a/. According to their report, this is because of the larger and sharper peaks of time domain acoustic signal of /a/ with respect to the other vowels.

In this study, to successfully achieve the assessment of pathological speech, spectral envelope and pitch information have been considered. The first aim of this work is to classify speech signals in terms of being normal or pathological then, based on the results of this step, recognizing different abnormalities. We focus on two methods for pathological speech assessment. The first method is based on classification using an assortment of utterance level parameters, reflecting those used in the state of art in practice (as provided by the multi dimensional voice analysis – MVDP -- program), which have been derived from time domain analysis of speech signal. The second method is based on short-term analysis of spectral envelope and temporal dynamic measures. The results of these two methods and influence of different

classifiers in the classification rate of sustained vowel /a/ are also investigated.

This paper is organized as follows: in the next section, the employed method and database are discussed. In section three, the experimental results are described and finally, in section four, the conclusion of this study is stated.

II. MEHTOD AND DATABASE

In our implementation, two requirements were imposed. First, the features had to be efficient in terms of measurement cost and time. Second, both the vocal tract and excitation source information, had to be included. The block diagram of the proposed algorithm is shown in Fig. 1. The cepstral features of a mel frequency filter bank outputs were obtained by a standard short-term speech analysis along with frame-level pitch estimates. Then, a HMM based classifier was applied. The following section briefly describes the method of extracting features and in the next subsection, the MEEI voice disorders database is discussed.

A. METHOD

The normalized cross-correlation [6] based method is commonly used for pitch estimation. The algorithm assumes a monophonic signal. The method follows the assumption that the signal has a periodicity corresponding to the fundamental frequency or pitch. Starting with a signal $s(t)$ that is assumed to be periodic, or more precisely, quasi-periodic, it follows that $s(t) \approx \alpha . s(t+T)$ where T is the quasi-period of the signal and the scalar α accounts for amplitude variations. Considering a window of d samples of $s(t)$ taken with a sampling period, τ, one can define the vector:

$$v(t) = \left[s(t), s(t+\tau), ..., s(t+(d-1)\tau) \right]^t \qquad (1)$$

Two such windows, $v(t)$ and $v(t+T')$, are separated in time by T'.Using a Bayesian approach detailed in [6], the best estimate for the quasi-period is the T which maximizes the expression:

$$\lambda(t, T') = \frac{v(t)^t \, v(t+T')}{\|v(t)\| . \|v(t+T')\|} \qquad (2)$$

which is a normalized cross correlation between $v(t)$ and $v(t+T')$.

The main focus of this study is the binary classification of the speech signal. Therefore, this problem can be formulated as a two class system. Let each of subjects be represented by a sequence of feature vectors O ,which are the Mel frequency cepstral coefficients (MFCCs) and pitch. Pathological speech recognition then can be regarded as computing:

$$\arg \underset{i}{MAX} \left\{ P(Path_i \mid O) \right\} \qquad (3)$$

where, $Path_i$ = *{normal, pathology}* for the first step of classification. In practice, if a parametric production model such as the Markov model is assumed, then computing the joint probabilities, which are necessary for solving (3), can be replaced by estimating the Markov model parameters.

For training HMM, the hidden Markov [7] model toolkit (HTK) was modified to accommodate the fundamental frequency. Twelve Mel frequency Cepstral coefficients using 10 msec Hamming windowed frames were extracted. The Mel scaled defined as:

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (4)$$

The employed filters were triangular and equally spaced along the entire Mel-scale. The pitch frequency was also computed in the same window. The zero order energy and F0 were included with the above-mentioned features. In order to take the advantage of pitch and spectral dynamics, the velocity (delta) and acceleration parameters were also added to the feature space.

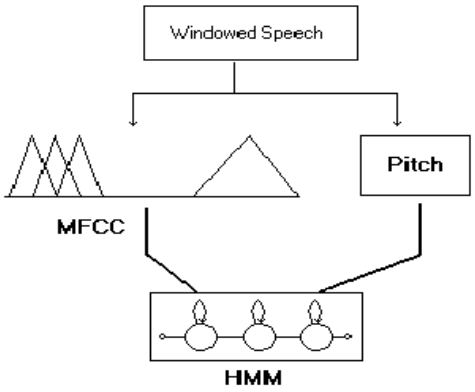

Figure 1: Block diagram of the method.

B. DATABASE

The database [8] developed by Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Laboratory was used. It contains voice samples of 710 subjects.

Disorders	No. of cases
Adductor spasmodic dysphonia	20
A-P squeezing	167
Paresis	22
Gastric reflux	54
Hyper-function	188
Paralysis	79
Keratosis / Leukoplakia	30
Normal	53
Others	97

Table 1: some subjects of database

Included are sustained phonation speech samples from patients with a wide variety of organic, neuralgic,

traumatic, and psychogenic voice disorders, as well as 53 normal subjects.

Along with sustained samples of the vowel /a/, the database also contains acoustic speech sample files: readings of "Rainbow passage". The Rainbow dataset consists of 662 pathological subjects from the same patients who provided samples of the sustain vowel /a/. There are also 53 recordings of up to 12 second of the Rainbow passage from the same normal subjects included in normal sustained /a/.

Utterance level results of analyzing each of the vowels by the Multi-Dimensional Voice Program (MDVP) were also included in the database [6]. These acoustic parameters, which include 34 time domain features of speech signal, represent a superset of the most popular measures employed in pathological speech analysis.

All patients had extensive vocal function testing including stroboscopic aerodynamic-phonatory, and acoustic analysis. The clinical and diagnostic information are also included. All acoustic files were recorded with condenser microphone in a soundproof booth. The distance from mouth to microphone was 15 cm. The signals were acquired at sampling rates of 25 kHz and 50 kHz after applying an appropriate anti-aliasing filter.

III. EXPERIMENTS AND RESULTS

In this section, experiments and classification results are discussed. As stated earlier, first we classify speech signals into two class of normal and abnormal, using different features and classifiers, then based on these results; recognition of A-P-squeezing abnormality will be performed.

The features derived from MDVP analysis of sustained vowel /a/, were classified using different classification methods. Some of these features were as follows: average fundamental frequency, average glottal period, highest fundamental frequency, lowest fundamental frequency, relative average perturbation, jitter, shimmer, etc where the method of their calculation has been documented in [8]. The linear discriminant analysis (LDC), nearest mean classifier (NMC) and HMM (GMM) with 3-mixtures were applied to these features to distinguish between normal and non-healthy subjects. A neural network (NN) with 34 input neurons, one hidden layers, and two output neurons was also trained using a back propagation algorithm.

In addition, two 3-states, 3-mixtures, left to right HMMs were formed based on 42 features obtained from spectral and fundamental frequency information. The EM algorithm was used to train the HMM and a series of experiments carried out with this HMM topology. In all of the experiments of this study, seven training iterations were enough for good convergence. The database was divided into two equal groups, in which training and testing took place.

In order to compare the results of each experiment, an improvement factor was defined as follows:

	Vowel /a/			
	Train		Test	
	CCR %	I %	CCR %	I %
MFCC	98.59	-	97.75	-0.85
MFCC+ Pitch	99.44	0.86	98.30	-0.29
	Rainbow Pasage			
	Train		Test	
	CCR %	I %	CCR %	I %
MFCC	98.03	-0.57	97.46	-1.15
MFCC+ Pitch	98.59	0.0	97.75	-0.85

Table 2: Correct classification rates (CCR) for training and test of HMMs using vowel /a/ and rainbow passage datasets, the improvement percentage (I) with respect to closed set performance is also included.

$$improvement = \frac{Accuracy}{Base\_Acuraccy} - 1 \qquad (5)$$

where, the $Base\_Acuraccy$ is the correct classification rate of HMM based classifier using MFCC features for vowel /a/ and $Accuracy$ is the correct classification rate of the data being processed.

In the first experiment HMMs were trained and tested using sustained vowel /a/ training data to discriminate healthy and pathological speech. In the next simulation, the HMMs were trained and tested with Rainbow training datasets. The trainings and testing for HMMs was performed with and without fundamental frequency. The results of these experiments are shown in Table 2. The effect of various classifiers on MDVP parameters was also investigated. The results of this study are shown in Table 3.

For the purpose of pathological speech assessment, 163 A-P-Squeezing subjects were selected from 657 non-healthy cases. The A-P-Squeezing data were further divided into four sub categories based on severity of A-P-Squeezing: 52 minor, 57 mild, 39 moderate, and 19 severe subjects. The data was divided into two equal groups for training and testing. In the third experiment HMMs were set up for discriminating of A-P-Squeezing and other abnormalities using the same structure of previous HMMs.

Method	Training		Test	
	CCR %	I %	CCR %	I %
LDC	95.64	-2.99	95.93	-2.07
NMC	67.15	-31.98	65.26	-33.81
NN	96.02	-2.61	96.15	-2.48
HMM (GMM)	97.97	-0.63	97.67	-0.93

Table 2: Correct classification rates (CCR) for normal/abnormal assessment using MDVP parameters and improvement percentage (I)

Method	Minor %	Mild %	Moderate %	Severe %
Training	65.38	75.00	73.68	88.88
Test	57.69	72.41	80.00	90.00
Total	61.53	73.68	76.92	89.47

Table 4: Correct classification rates for A-P-Squeezing analysis using spectral and pitch features

The models were trained using training data and tested with the test data. Table 4 shows results obtained from the HMM training and testing phases. The average correct classification rate for A-P-squeezing was 72.55%. As the results show, the correct classification rate for severe A-P-squeezing is higher than the other types of A-P-Squeezing. The results of this table also indicate that the correct classification percentage increases with the degree of pathology.

In order to compare the results of third experiment with traditional temporal dynamic measures, the MDVP parameters were analyzed using the same classifiers, which listed in Table 3. The classification results are shown in Table 5. The average correct classification rate for in the best case (HMM) was 57.99%.

IV. CONCLUSION

In this paper two methods for pathological speech assessment were discussed. The first method was based on classification of MDVP parameters, which were derived from time domain analysis of speech signal. As the results of table 3 show, using the GMM classifier, the correct classification rate was 97.97%. The best correct classification rate for sustained vowel /a/ was 99.40% using spectral and pitch features with an HMM. The improvement percentage indicated that the spectral and pitch features had better recognition rate and the MDVP parameters had lower classification rate than spectral features. In addition, the best correct classification rate for the Rainbow passage was 98.59% and the improvement was less than vowel /a/.

Method		Minor %	Mild %	Moderate %	Severe %
LDC	Training	43.08	37.86	35.79	48.50
	Test	31.76	32.50	27.89	46.32
NMC	Training	51.21	60.75	43.72	40.10
	Test	50.18	58.63	41.03	41.34
NN	Training	39.23	30.71	25.26	46.38
	Test	29.80	25.36	27.89	41.05
HMM (GMM)	Training	56.15	63.57	51.58	53.11
	Test	55.10	61.07	54.21	51.58

Table 5: Correct classification rates for A-P-Squeezing analysis using MDVP parameters

In the case of pathological assessment, using the MDVP parameters for sustained vowel /a/, the average result was 57.99% in the detection of A-P-Squeezing abnormality. The average correct classification rate with spectral features and pitch using HMM was 72.55%, which shows significant improvement with respect MDVP parameters.

As illustrated by the results of this study, using just a sustained vowel /a/ provides fairly reliable detection in terms of normal and pathology assessment. The results also indicated that the spectral and pitch features, which are low cost and fully automatic, better classification rate with respect MDVP features so it is possible to make a low cost, accurate, and automatic tool for pathological speech assessment using spectral and pitch information. In addition, the results showed that using of lexical utterances such as the rainbow passage are still reliable albeit with some performance degradation. The analysis of vowel /a/ using pitch and spectral information showed better results comparing with MDVP parameters. However, to achieve very high performance accuracy, the feature from other vowels may be added to the features, which derived from vowel /a/. Also, more research is needed to develop methods for automatic recognition of other specific speech pathological types.

V. REFRENCES

[1] Rosa, M. D. O. Pereira, J. C. Grellet, M., ``Adaptive Estimation of Residue Signal for Voice Pathology Diagnosis``, IEEE Trans. Biomedical Eng. Vol. 47, No. 1, Jan. 2000.

[2] Ceballos, L. G. and Hansen, H. L. ``Direct Speech Feature Estimation Using an Iterative EM Algoritm for Vocal Fold Pathology Detection``, IEEE Trans. Biomedical Eng. Vol. 43, No. 4, April. 1996.

[3] Iwata, S., ``Periodicities of pitch perturbation in normal and pathologic larynges``, Laryngoscope, vol. 82, pp. 87-96, 1972.

[4] Maurilio, N. Vieira, Fergus, R. Mclnnes, Mervyn, A. Jack, ``On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures``, JASA vol. 111, Feb. 2002.

[5] Maurilio, N. Vieira, Fergus, R. Mclnnes, Mervyn, A. Jack, ``comparative assessment of electroglottographic and acoustic measures of jitter in pathological voice``, J speech Lang. Hear. Res. 40, pp. 205-228, 1997.

[6] Talkin, D. Klejin W. B. and Paliwal, K. K., ``A Robust Algorithm for Pitch Tracking``, Speech coding and synthesis, Elsevier, New York, 1995.

[7] Young, S. Kershaw, D. Odell, J. Ollason, D. Valtchev, V. Woodland, P. ``The HTK book``, Microsoft Corporation, July 2000.

[8] ``Disorder Dtabase Model 4337`` Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA, Jan. 2002.