Vocal Features for Glottal Pathology Detection using BPNN

Ashwini Visave
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai - 400019, India

Pramod Kachare
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai - 400019, India

Amutha Jeyakumar
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai - 400019, India

Alice Cheeran
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai - 400019, India

Gurmit Bachher
Department of Speech Therapy,
TATA Memorial Hospital,
Parel, Mumbai - 400012, India

ABSTRACT

Development of low cost, non-invasive applications is one of the most challenging tasks in the field of biomedical signal processing. Present work focuses on detection of glottal pathology with the knowledge of prominent speech processing and machine learning techniques. This paper addresses the discriminative characteristics of speech signal like, pitch, jitter, linear prediction residual and cepstral source excitation to aid such an identification system. Back-propagation Neural Network model is developed for various feature combinations to classify the glottal pathologic voice from normal voice. Accuracy of the developed system is evaluated considering different feature sets. Work also concludes the efficiency of such acoustic features for various combinations using objective measures like confusion matrix, true positive rate i.e. sensitivity, specificity i.e. true negative rate and accuracy. The results show promising development in identification of pathological individual from normal person using voice samples.

General Terms:

Glottal pathology, glottal features, BPNN

Keywords:

Pitch, jitter, lpc residual, source excitation, short time energy, confusion matrix

1. INTRODUCTION

Speech is one of the fundamental mediums of communicating known to mankind since the beginning of mankind. Absence or deterioration of such a media will pose a great threat to proper understanding between individuals. Such a difficulty may arise due to the malfunction of the human speech production system or auditory system. In this work we focus on the problem caused in vocal folds, as their movement which contributes towards large, in the production of voice. Any pathology that occurs to alter the periodic movements of the vocal folds affect speech produced. Some of the various pathologies causing such a distress can be listed as cold, cough, vocal fold paralysis, papilloma, carcinoma, polyp, etc. The existing detection methods are both expensive and time consuming. Also, the invasive nature of the techniques restricts their ease of operation. Efforts are being made to design an automatic, rapid, cost effective and non-invasive recognition system which will discriminate such pathologic samples from those of the normal ones. Also various methods have been tried with distinct features of speech involving machine learning.

The main objective of this work is formulation of a robust system to detect vocal fold pathology at an early stage from set of features derived from simple voice sample. Significant changes in voice occur can be perceived by our ear, occur at a later stage. Many researchers have contributed in areas related to automatic feature extraction process like acoustic analysis, parametric and non-parametric feature extraction, pattern recognition and statistical analysis[3].

L. Gavidia-Ceballos et al.[3], conducted some research on vocal fold pathology using acoustic analysis of speech. Both acoustic analysis and voice production features are important in this detection. L. Gavidia-Ceballos et al.[3], used speech production parameters as complete glottal closure is very hard to obtain in vocal fold pathology. The method proves advantageous in the estimation of Enhanced Spectral Pathology Component (ESPC) instead of glottal flow waveform, which varies considerably between pathology and healthy conditions. The meaningful measures Mean-Area-Peak-Value (MAPV) and Weighted-Slope (WSLOPE) were derived from ESPC feature[3].

Among all the features, Harmonic to Noise Ratio (HNR) is a very popular measure among otolaryngologists in voice quality assessment. The experiments on voice quality assessment extracted short term and long term in time domain and frequency domain parameters from impedance (EGG) signals considered[12]. The long term features include the mean of fundamental frequency F0, the standard deviation of F0, and the percentage of the voiced part in 3 sec long speech signal, while the short-term features include parameters related to the spectral envelope of the first few glottal harmonics, and the glottal noise. In [1], a system was implemented using 12 Mel-Frequency Filter Bank Cepstral Coefficients (MFCC) and dynamic pitch measures. It comments that few features like pitch perturbation parameters, MFCC can deliver us with better accuracy in classification.

An acoustic analysis was done by S. Hadjitodorov et al.[6], in which jitter, shimmer, Harmonic to Noise Ratio (HNR) along with the new proposed parameters Turbulent Noise Index (TNI, for voiced signals) and Normalised First Harmonic Frequency (NFHF), for breathy voice characterization. The decisive support system built in[4], considered parameters like, pitch and amplitude perturbation measures, Frequency measures, MFCC, Autocorrelation, HNR in spectral domain and cepstrum domain, Linear Prediction Coefficients (LPC), Linear Prediction Cosine Transform Coefficients (LPCT), 23 different features used in commercial “Dr. Speech” software[4]. Formulation of the features in [7], based on the parameters mentioned as, pitch and Degree of Voicing (DOV), spectral envelope, harmonic frequency jitter, LPC and its residual, Glottal Source[7]. V. Uloza et al. [14], performed work based on the parameters such as pitch and amplitude perturbation measures (24), Frequency measurs (100), Mel-frequency (35), Cepstral energy features (100), MFCC (35), Autocorrelation (80), HNR in spectral domain (11), HNR in cepstrum domain (11), LPC (16), LPCT (16). In [15], X. Wang et al. proved that Mel-Frequency Cepstral Coeffients (MFCC) are effective to discriminate between pathological and healthy voice. However, MFCC suppresses the excitation parameter in filtering using Mel filters and the significant information about the source is lost. Glottal muscles cannot move at infinite rate. It remains quasi-stationary for about 30 ms. The glottis generates pitch spikes usually shorter than 30 ms. Therefore, much lower energy signal, residual, needs to be derived from LPC[8].

The paper mainly contributes to detection of glottal pathology using features like pitch, jitter, LPC residual and source parameter derived from cepstrum. Evaluation of the system performance was carried out for various features and for ANN classifier, using confusion matrix, true positive rate, false negative rate and overall accuracy.

The rest of the paper is organized as follows: section II explores proposed algorithm, section III explains the database, section IV gives idea of classifier used in the work, section V elaborates the results obtained.

2. PROPOSED ALGORITHM

The goal of this work is the extraction of glottal parameters from speech signal for the distinction between glottal pathological voice and normal voice. Block diagram representation is shown in fig. 1. The speech database follows pre-processing in which normalization, framing, windowing, overlapping and voiced/ unvoiced decision is made. Various acoustic features are then extracted from the voiced speech frames. A feature set is formed using the extracted acoustic parameters and a neural network is trained for the patterns of healthy and pathological voice feature patterns. The test speech sample is processed through the same procedure as that of speech corpus till feature extraction. The trained neural network examines the pattern of the test sample and classifies it according to the pattern matching technique. Next sessions describe the various blocks individually.

2.1 Speech Corpus

The initial step is to collect input speech data for the two directories, normal voice samples and glottal pathologic voice samples. This mainly includes speech database of patients suffering from glottal and supra-glottal cancer and from normal persons. The corpus is discussed thoroughly in section III.

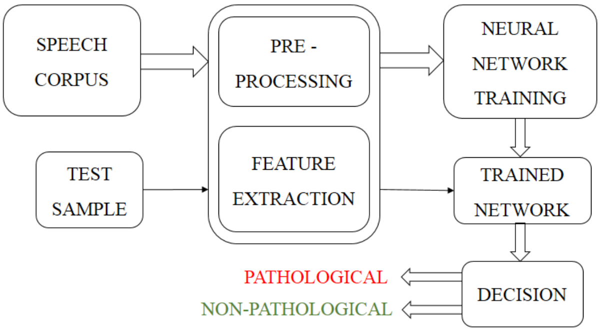

Fig. 1. Architecture

2.2 Preprocessing

2.2.1 Normalization. It involves zero mean and unity variance normalization of speech recordings of both normal persons and patients with pathology to remove dc offset.

2.2.2 Framing and Windowing. Due to the quasi-stationary nature of speech signal, its properties are stable for a very short time of 10-25 ms, entire length of speech signal is fragmented in smaller frames each of 20ms.

Windowing converts long length signal into finite length signal[13]. Hamming window is used in this work. The window function for hamming window is given as below:

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{1}$$

De-emphasis occurs at the ends of the applied window.

2.2.3 Overlapping. The solution to avoid the de-emphasis introduced because of windowing is overlapping of the frames i.e. the new frame contains some part of previous frame and the upcoming frame[8]. In this work, an overlap of 70% chosen.

2.2.4 Voiced/ Unvoiced Decision. Voiced/ unvoiced decision is made using zero crossing rate (ZCR) and short time energy (EN) to remove unvoiced part from continuous speech samples. ZCR and EN are found from following equation[8],

$$ZCR_j = \frac{1}{N}\sum_{n=0}^{N-1} |sign(s_j[n]) - sign(s_j[n-1])| \tag{2}$$

$$EN_j = \frac{1}{N}\sum_{n=0}^{N-1} |s_j[n]|^2 \tag{3}$$

For voiced sounds, ZCR is low and energy is higher than the threshold which is generally set as 10% of the total energy while for unvoiced sound, ZCR is high and energy is lower than the threshold.

2.3 Feature Extraction

The features namely pitch, pitch perturbation measure (jitter), source excitation parameter derived from cepstrum, LPC residual and short time energy are used in the proposed work.

2.3.1 Pitch. The fundamental frequency of speech signal is the pitch. It represents the rate of vibration of vocal folds during sounds such as voiced ones. Vocal folds vibrate at different rates because of in built tension in them and subglottal air pressure which forms pitch frequencies[13]. Pitch and its perturbation parameter are the most widely used features in the area of voice pathology as both define significant alterations in waveform patterns in pathological sound and normal sound. In this work, we have used autocorrelation method and center clipping to compute the pitch.

The straightforward method of selecting the highest peak of autocorrelation function fails when a situation comes into existence like the autocorrelation peaks due to periodic nature of vocal excitation are smaller than vocal tract response. To overcome this problem, techniques like "spectrum flatness" were introduced to discard the effects introduced due to vocal tract response[11]. Center clipping is one of the techniques used for spectrum flatness. The technique decides clipping level which is a fixed percentage (we have used 20%) of the maximum amplitude of the speech segment. Pickpeaks algorithm finds the peak locations and peak values of the speech signal. From this, the pitch of the signal is estimated.

2.3.2 Pitch perturbation measure (Jitter). Jitter is a measure of alterations in the pitch period of the speech signal. The cycle to cycle perturbations of the glottal cycles which lead to aperiodicity in the speech waveform is termed as jitter[10]. The absolute difference between the time variations of consecutive fundamental periods is known as absolute jitter.

$$absolute\ jitter = \left(\frac{1}{L-1}\right)\sum_{i=1}^{L-1}|T_i - T_{i+1}| \tag{4}$$

where,
T_i = Period of the i^{th} frame
L = Number of voiced segments

2.3.3 Source Excitation Feature. This is the cepstrum based approach to analyse the speech as shown in Fig. 2. Cepstrum is computed by taking Inverse Fourier Transform of log magnitude spectrum of input speech signal.

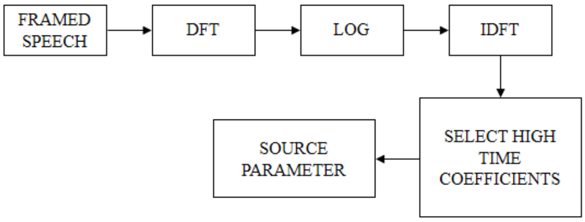

Fig. 2. Glottal parameter extraction

Speech signal s[n] can be considered as the linear convolution of source i.e. glottal excitation g[n] and impulse response of vocal tract v[n], which can be stated as below:

$$s[n] = g[n] * v[n] \tag{5}$$

By applying Discrete Fourier Transform (DFT) to the framed speech signal, we get,

$$S(w) = \sum_{n=-L}^{+L} s[n]e^{-jwn} \tag{6}$$

where, L is the order of the cepstrum which is the number of one sided frequencies[9]. DFT transforms the time domain convolution in simple multiplication in frequency domain as,

$$S(w) = G(w).V(w) \tag{7}$$

To separate the glottal excitation and vocal tract parameters from each other, we draw log transform of S(ω).

$$\hat{S}(w) = \log(S(w)) = \log(G(w)) + \log(V(w)) \tag{8}$$

The cepstrum is defined as,

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \hat{S}(w)e^{jwn}dw \tag{9}$$

The log spectral components which vary rapidly with frequency ω are stated as high time coefficients, log(G(ω)). High time coefficients can be obtained from high time lifter [9] given by as below:

$$l_h[n] = \begin{cases} 0, & \text{elsewhere} \\ 1, & L_c < \text{n} < \text{L} \end{cases} \tag{10}$$

where, L_c is chosen to be less than the pitch period[11].
The excitation component can be found from cepstrum and high time lifter as follows:

$$c_e[n] = c[n]l_h[n] \tag{11}$$

where, $l_h[n]$ is high time lifter, $c_e[n]$ is excitation parameter.

2.3.4 LPC Residual. The proposed algorithm then proceeds to find out LPC parameters and LPC residual from LPC parameters as shown in fig. 3. Applying inverse filter to the obtained LPC coefficients $a_i[n]$ of speech frames, ultimately provides us the noisy, time localised excitation called as LPC residual *e*[*n*][7][2].

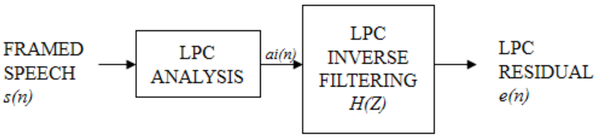

Fig. 3. LP residual extraction

$$e[n] = s[n] - \sum_{i=1}^{N} a_i s[n-i] \tag{12}$$

where, *i* is the number LPC coefficients, N is the filter order. We have used N = 13 in the work.

2.3.5 Short time energy. It can be described as speech activity detector as the energy for each speech segment is calculated and frame–to–frame change in energy can be easily noticed. High energy leads to the production of high volume and high amplitude speech. The amplitude variations are described by the short time energy of the voice signals[13]. Short time energy is given by the formula as below:

$$EN_i = \frac{1}{N}\sum_{n=0}^{N-1}|s_i[n]|^2 \tag{13}$$

3. DATABASE

The database consists of 65 recordings of the male patients who have glottal cancer and 27 recordings of male candidates who are free from any disorder. The average age of the persons whose voice samples are used is 60. The recording is done by an unidirectional microphone using “Dr. Speech” and “Goldwave” softwares at the sampling rate of 11025 Hz and 16 bit allocation for each sample. Average recording length of sustained phonations of vowels /*a*/, /*i*/, /*u*/ is 2 seconds while that of reading passage in Marathi and Hindi language is 25 seconds. The database is prepared at the “Speech Therapy Department” of “TATA Memorial Hospital, Parel, Mumbai” in a sound proof room. Clinical diagnosis of each of the patient is already assessed by the experts of the TATA Memorial Hospital i.e. the pathological state of the person is known at the time of recording. The factors which were considered while recording are mentioned as below:
1) Age
2) Gender
3) Distance of microphone from speaker’s mouth
4) Sampling rate and bit resolution used for each recording
5) Recording length
6) Software used for recording

4. CLASSIFIER

Artificial Neural Network based approach of classification is applied in this work. Fig. 4 shows a simple architecture of backpropagation neural network model. The model consists of interconnection of neurons and interconnection between the two neurons has weights associated with it. Backpropagation model with learning delta rule is used to update the interconnection weights to minimize the mean squared error between the actual output values and the desired ones.

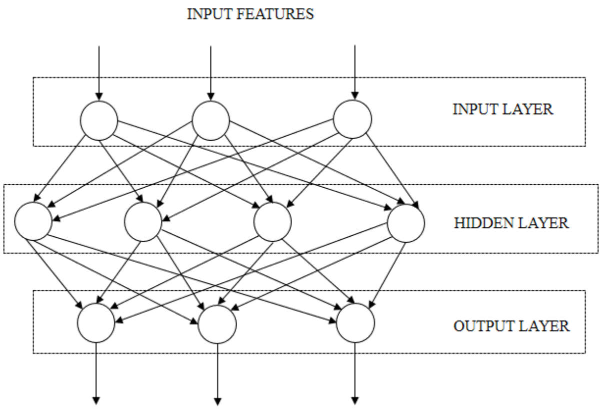

Fig. 4. Backpropagation neural network model

The process of weights updation is repeated untill the error becomes tolerable and sufficient gradient is reached. The training process was repeated several times with different combinations of the hidden layers and hidden neurons to get more accuracy in terms of the confusion matrix.

Table 1. Brief Understanding of TP, TN, FP, FN

CLASSIFICATION PREDICTIONS	PATHOLOGY	NORMAL
PATHOLOGY	TP	FN
NORMAL	FP	TN

Table 2. Feature Combinations and Accuracy Obtained

FEATURES	ACCURACY	TPR	TNR
PITCH, JITTER, LPC RESIDUAL,	65.65%	73.13%	50%
PITCH, JITTER, LPC RESIDUAL, SOURCE EXCITATION	67.39%	69.23%	62.96%
PITCH, JITTER, LPC RESIDUAL, SOURCE EXCITATION, SHORT TIME ENERGY	68.08%	61.53%	82.75%

The feature set is designed for clinical voice and normal voice using the parameters pitch, jitter, source component derived from cepstrum and LPC residual. Target set representing the feature set for both clinical voice and normal one is designed for two nodes. The network is trained using the feature set and the target set. Then the feature set of test sample is given as input to the trained network for pattern matching and a decision is made.

5. PERFORMANCE EVALUATION AND DISCUSSION

The total number of speech samples used are 92 for the evaluation purpose of which 65 are glottal cancerous while 27 are normal voices. The performance evaluation is given by the knowledge of actual and the future event discriminations in the classification system as included in the confusion matrix or contingency matrix. The terms used in the confusion matrix can briefly be described and are as shown in table I.

5.1 True Positive (TP)

True decisive system classified as true[5].

5.2 True Negative (TN)

False event detected as false[5].

5.3 False Positive (FP)

The event is false and discriminated as true [5].

5.4 False Negative (FN)

True event classified as false[5].
The combination of pitch, jitter, LPC residual, source excitation and short time energy provides better accuracy than for any other parametric combination (Table II). LPC residual delivers the noise and pitch information as the vocal tract parameters are eliminated by inverse filtering while the source excitation, from high time liftering of cepstrum, gives the glottal waveform. This comments that the source excitation together with LPC residual add value to the vocal fold pathology detection with just pitch and jitter. The short

Table 3. Confusion matrix

	class 1	class 2	
class 1	40 43.47%	5 5.43%	88.88% 11.11%
class 2	25 27.17%	24 26.08%	48.97% 51.02%
	61.53% 38.46%	82.75% 17.25%	68.08% 31.92%

time energy is less for clinical voice samples than that for healthy voice recordings. It increases the system accuracy with smaller percentage as amplitude variations are reflected by it.

5.5 Specificity (SP)

It is the probability that an event is absent and will be detected as absent. It can also be termed as true negative rate(TNR)[5].

$$SP = \frac{TN}{TN + FP} * 100 \tag{14}$$

5.6 Sensitivity (SE)

Sensitivity i.e. the true positive rate (TPR) is the probability that the event is present provided that it is present[5].

$$SE = \frac{TP}{TP + FN} * 100 \tag{15}$$

5.7 Accuracy (AC)

It is the probability that the classification by the system is correct[5].

$$AC = \frac{TP + TN}{TP + FP + TN + FN} * 100 \tag{16}$$

The true positive rate i.e. sensitivity obtained is 61.53% while TNR obtained is 82.75%. The positive predictive value (PPV) 88.88% is the proportion of true positive and test outcome positive while 11.11% is the complement of PPV known as false discovery rate (FDR). Similarly, the negative predictive value (NPV), 48.97%, is the ratio of true negative and test outcomes negative. While, 51.02% is the false omission ratio (FOR), which is the complement of NPV. The overall accuracy obtained with the help of the confusion matrix is 68.08%, while the percentage of events incorrectly classified is 31.92.

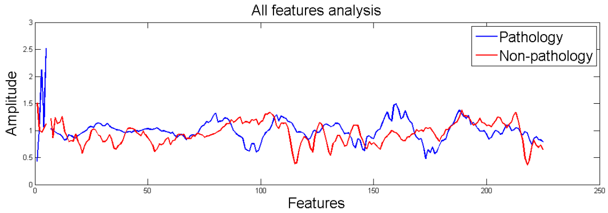

Fig. 5. All Features analysis

Fig. 5 shows the variations in features of a pathological and a non-pathological speech samples. The figure plots all of the features of single sample from both the categories viz., LPC residual, source excitation, pitch, jitter and short time energy, for both the samples. Very small variations are observed when analyzed using different combinations of features (combinations as shown in confusion matrix) which can be seen from figures 5, 6 and 7. Fig. 6 is a combination of pitch, jitter, LPC residual and excitation derived from cepstrum using high time liftering while, fig. 7 gives the information of pitch, jitter and LPC residual.

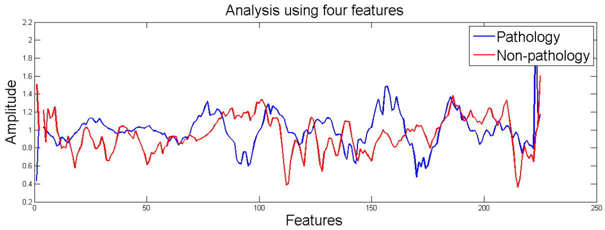

Fig. 6. Four Features analysis

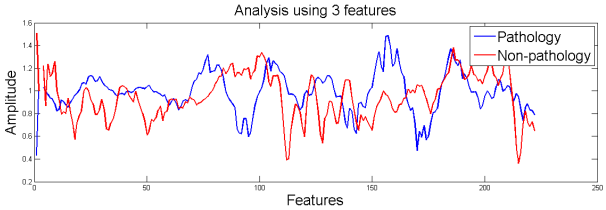

Fig. 7. Three Features analysis

6. CONCLUSION

The contribution of the features like pitch, jitter, LPC residual and excitation parameters is to the great extent than the parameters like MFCC, which are vocal tract parameters, in detection of the glottal pathologies. Accuracy in detection of glottal pathology increases impulsively with the addition of LPC residual and excitation parameters to the pitch and jitter. Short time energy adds considerable value in making correct decision as it is the reflection of amplitude variations. Among the five features considered, pitch and jitter are the most important parameters in discrimination of the sounds. The confusion matrix envisions the performance of the system from TPR, TNR, PPV, NPV, AC, TP, TN, FP, FN. The work presented can be improved by making the use of glottal to noise excitation ratio which reflects breathiness in voice production system. The system may give better results using support vector machines with non-linear kernel function for classification purpose.

7. ACKNOWLEDGEMENT

The authors would like to thank Dr. Samudravijaya from TATA Institute of Fundamental Research, Mumbai, Dr. P. C. Pandey from Electrical Engineering Department, IIT Bombay and Prof. J. H. Nirmal from K. J. Somaiya College of Engineering, Mumbai, for their guidance to complete the work in this area of research. Authors are very much thankful to Mrs. Swathi Shetty and Miss. Neha Patil, TATA Memorial Hospital, for their precious efforts in collecting the speech recordings of the patients suffering from glottal cancer.

8. REFERENCES

[1] A.A. Dibazar, S. Narayanan, and T.W. Berger. Feature analysis for automatic detection of pathological speech. *Proceedings of the Second Joint 24th Annual Conference and the*

Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology, 1:182–183, 2002.

[2] Carlo Drioli and Federico Avanzini. Hybrid parametric-physiological glottal modelling with application to voice quality assessment. *Medical Engineering and Physics*, 24(7-8):453–460, 2002.

[3] L Gavidia-Ceballos and J H Hansen. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *IEEE transactions on bio-medical engineering*, 43(4):373–83, April 1996.

[4] A Gelzinis, A Verikas, and M Bacauskiene. Automated speech analysis applied to laryngeal disease categorization. *Computer Methods and Programs in Biomedicine*, 91(1):36–47, July 2008.

[5] Juan Ignacio Godino-Llorente, Pedro Gómez-Vilda, and Manuel Blanco-Velasco. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *IEEE Transactions on Biomedical Engineering*, 53(10):1943–1953, 2006.

[6] Stefan Hadjitodorov and Petar Mitev. A computer system for acoustic analysis of pathological voices and laryngeal diseases screening. *Medical Engineering & Physics*, 24(6):419–429, July 2002.

[7] Zvi Kons, Aharon Satt, Ron Hoory, Virgilijus Uloza, Evaldas Vaiciukynas, Adas Gelzinis, and Marija Bacauskiene. On Feature Extraction for Voice Pathology Detection from Speech Signals. pages 3–6.

[8] Ian McLoughlin. *Applied Speech and Audio Processing with MATLAB Examples*. Cambridge University Press, 2009.

[9] Jagannath Nirmal, Pramod Kachare, Suprava Patnaik, and Mukesh Zaveri. Cepstrum liftering based voice conversion using RBF and GMM. In *International Conference on Communication and Signal Processing, ICCSP 2013 - Proceedings*, pages 570–575, 2013.

[10] Lc C Quilty, Km M Godfrey, and S H Kennedy. Harm avoidance as a mediator of treatment response to antidepressant treatment of patients with major depression. *Psychotherapy and*, 8:1–7, 2010.

[11] L. Rabiner and R. Schafer. *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice Hall, 1978.

[12] R.T. Ritchings, M. McGillion, and C.J. Moore. Pathological voice quality assessment using artificial neural networks. *Medical Engineering & Physics*, 24(7-8):561–564, September 2002.

[13] V Sellam and J Jagadeesan. Classification of Normal and Pathological Voice Using SVM and RBFNN. *Journal of Signal and Information Processing*, 05(01):1–7, 2014.

[14] Virgilijus Uloza, Antanas Verikas, Marija Bacauskiene, Adas Gelzinis, Ruta Pribuisiene, Marius Kaseta, and Viktoras Saferis. Categorizing normal and pathological voices: Automated and perceptual categorization. *Journal of Voice*, 25(6):700–708, 2011.

[15] Xiang Wang, Jianping Zhang, and Yonghong Yan. Discrimination between pathological and normal voices using GMM-SVM approach. *Journal of Voice*, 25(1):38–43, 2011.