Cepstral Analysis of Speech for the Vocal Fold Pathology Detection

Jennifer C Saldanha
Asst. Professor,
Dept. of E&C
SJEC, Mangalore

Ananthakrishna T
Asst. Professor,
Dept. of E&C,
MIT, Manipal

Rohan Pinto
Asst. Professor,
Dept. of E&C
SJEC, Mangalore

ABSTRACT

It is possible to identify voice disorders using certain features of speech signals. A complementary technique could be acoustic analysis of the speech signal, which is shown to be a potentially useful tool to detect voice diseases[2]. The focus of this study is to compare the performances of mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) features in the detection of vocal fold pathology and also bring out scale to measure severity of the disease. The speech processing algorithm proposed estimates features necessary to formulate a stochastic model to characterize healthy and pathology conditions from speech recordings. Two different set of features such as MFCC and LPCC are extracted from acoustic analysis of voiced speech of normal and pathological subjects. A linear discriminant analysis (LDA) classifier is designed and the classification results have been reported.

Keywords

Mel frequency cepstral coefficients, Linear predictive cepstral coefficients, Linear discriminant analysis**.**

1. INTRODUCTION

The vocal fold pathology mainly affects on the normal vibration pattern of the glottis and this in turn brings changes in the voice quality [1]. The analysis methods found in the literature are mainly based on the periodicity of vocal fold vibration and the turbulence in the glottal flow resulting from malfunctioning of the vocal folds. Researchers have used vocal noise level in the voiced speech as one of the parameter for the analysis of normal and pathological voices [2]. Time-domain based acoustical parameters are also found in the literature to evaluate pathologic voices which include pitch, jitter, shimmer etc [1]. In this study cepstral coefficients of voiced speech are used as parameters to classify normal voice from pathologic voice. The use of cepstrum in the assessment of pathological voices is supported by two arguments: on the one hand, cepstrum analysis is appropriate for estimating the noise level of the voice signal. On the other hand, for the case of sustained vowels, the variability of the glottal waveform can also be easily detected from cepstral parameters. Cepstral analysis de-convolves the speech sample in to source and system components, compresses the range of the magnitude spectrum, and reduces correlation between coefficients. Two of the most common are mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC), which are used in this study for speech feature extraction. A linear discriminant analysis (LDA) classifier is used separately on these sets of features and also on the combination of these features to test their efficacy as a tool for the detection of laryngeal pathology. As the same classifier is used on the three feature sets independently, three different sets of classification results were obtained [2]. As a pre-classified database of voices is used in this study, this allows us to make a comparison between the efficiencies of the three sets of features, apart from their individual efficiencies. Next section deals with the methodology used in this application.

2. METHODOLOGY

This application has two phases, namely training and testing. In the training phase speech sample from different normal and pathological subjects is preprocessed to convert it in to a form suitable for extraction of features. Feature extraction is a process of converting a sequence of speech samples into a set of observation vectors which represent events in a probabilistic space over which classification is performed. This process is also called as speech signal parameterization. The required features are extracted and stored as separate normal and pathologic templates. This acts as a model for classification. Figure 1 shows the block diagram of the application. In the testing phase, test sample is preprocessed, converted to a parametric form and compared with all other stored templates.

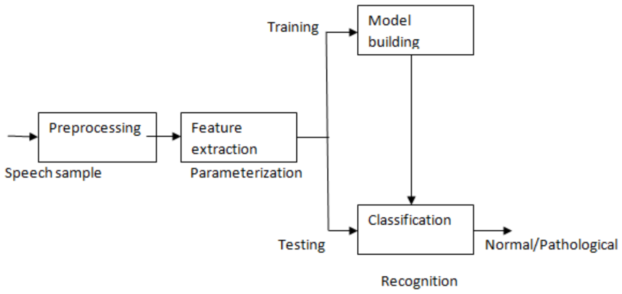

Figure 1. Block diagram of the application

The next step is to select the most accurately matching template and classify the sample. For the evaluation of the operating capacity of the recognition system 50% of the voice samples have been used for the training phase of the classifier, and the 50% of the remaining samples have been used for the testing phase [1]. The next section deals with computation techniques and measurements used in this study.

3. COMPUTATIONAL PROCEDURE AND MEASUREMENTS

3.1 Preprocessing

The voiced speech data sample (sustained phonation of the vowel \a\) is down-sampled at 8 KHz. And divided into 20msec equal frames with 10msec overlap using hamming window [3]. Figure 2 and Figure 3 shows the windowed speech frame of both normal and pathological samples. The source for the voiced speech is often modeled as quasi-periodic glottal pulses[2]. Hence in Figure 2 normal speech sample shows the periodic pattern and in Figure 3 pathological speech sample shows no periodicity due to turbulence of airflow through the glottis and pitch perturbations [2].

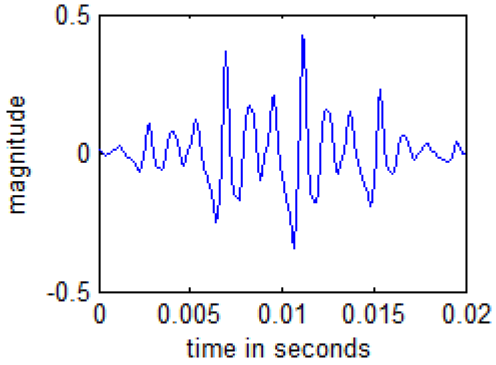

Figure 2. Windowed speech frame for normal voice sample

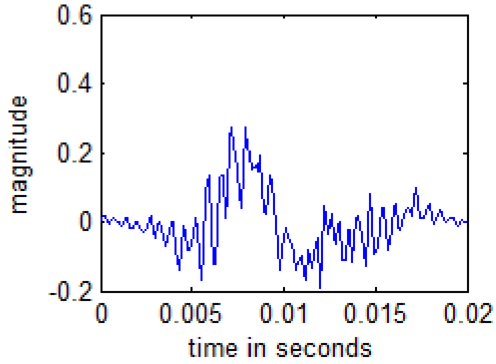

Figure 3. Windowed speech frame for pathological voice sample

3.2 MFCC estimation

The term “Mel” is some kind of measurements of perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scale (Mels) is approximately linear below 1 KHz and logarithmic at higher frequency [4]. Figure 4 shows the block diagram for the computation of MFCC.

To obtain MFCC, first perform discrete Fourier transform (DFT) on each frame of speech signal. Power spectrum of each speech frame is then weighted by a series of filter frequency response whose center frequencies and bandwidths roughly match those of auditory critical band filters. These filters follow mel-scale whereby band edges and center frequencies of the filters are linear for low frequency (<1000Hz) and logarithmically increase with increasing frequency [4].
Thus these filters are called as mel–scale filters and collectively a mel-scale filter bank. Therefore we can use the following formula (1) to compute the mels for a given frequency *f* in Hz [5]:

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

The mel-scale filter bank implementation used in this study includes 24 triangular filters, non-uniformly spaced along the frequency axis [6], as shown in Figure 5. The next step in determining the mel cepstrum is to compute the energy in the each mel-filter. The real cepstrum associated with this energy is called as mel cepstrum and computed using discrete cosine transform (DCT) of log energies. The coefficients of mel cepstrum are called as MFCC. The 1st MFCC (zero-order MFCC), indicates average log energy in each frame and is usually dropped [7]. The remaining first 12 MFCC coefficients are considered from each speech frame. The first and second order derivatives of MFCC are also estimated in order to improve the classification accuracy resulting in a feature vector of 36 coefficients representing each frame [6].

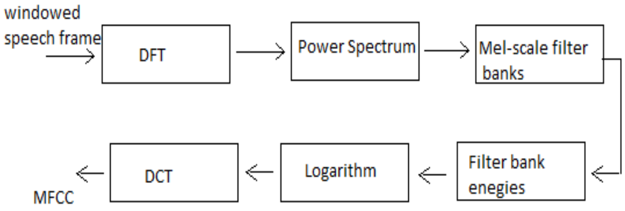

Figure 4. Block diagram of MFCC estimation

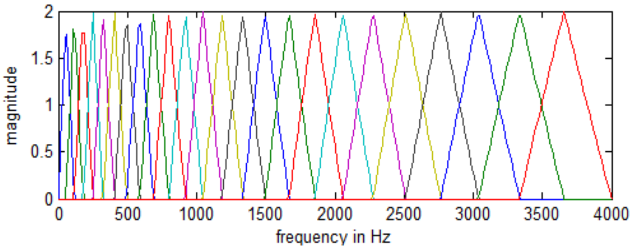

Figure 5. Mel-filter bank containing 24 triangular overlapping filters.

3.3 LPCC estimation

The linear prediction method provides a robust accurate method for estimating the parameters of time varying system representing vocal tract. It is used to separate vocal tract components and excitation components in time domain, hence easy to implement, and achieves data compression. Figure 6 shows the block diagram of the computation of LPCC.

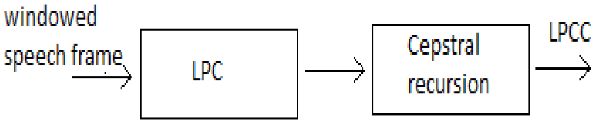

Figure 6. Block diagram of LPCC estimation

It is used for the prediction of current sample as a linear combination of past samples from the basis of linear prediction analysis. By minimizing the sum of squared error between actual speech samples and the linear predicted ones, a unique set of prediction coefficients are determined. Vocal tract transfer function is given as [8],

$$H(z) = \frac{G}{1-\sum_{k=1}^{p} a_k z^{-k}} \qquad (2)$$

where a_k are the LPC coefficients, *G* is the gain term in the LPC model and *p* is the order of the prediction filter.

For the computation of LPCC, auto-correlation sequence is calculated from windowed speech segment. LPC coefficients are calculated by means of Levinson-Durbin algorithm from auto-correlation sequence. LPCC coefficients c_k can be derived directly from the LPC coefficient set a_k by means of the next recursion formula (3),

$$c_k = -a_k + \frac{1}{k}\sum_{m=1}^{k-1}[-(k-m)a_m c_{(k-m)}];\ \ 1 \le k \le p \quad (3)$$

LPCC feature vector containing 12 coefficients per frame is extracted for the population of normal and pathological samples.

Figure 7 and Figure 8 shows the variations in MFCC and LPCC for normal and pathologic speech frame. Significant variation between normal and pathological voice sample is observed in case of MFCC than LPCC.

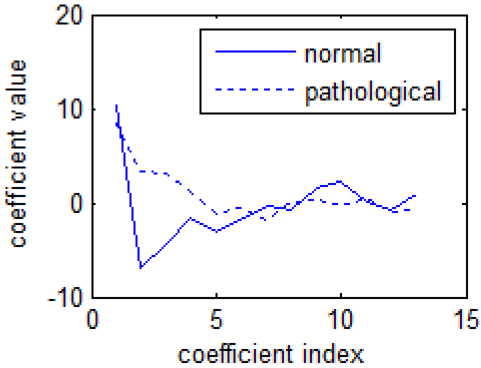

Figure 7. Plot of MFCC for normal and pathological voice sample

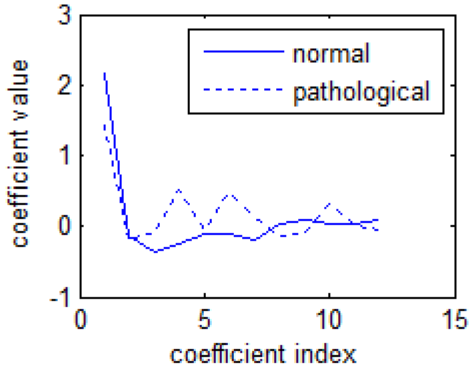

Figure 8. Plot of LPCC for normal and pathological voice sample

3.4 LDA Classifier

Objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. In LDA, between class scatter and within class scatter are used to formulate criteria for class seperability [9]. The solution obtained by maximizing this criterion gives the direction of the projection of the data down to one dimension, and is estimated as [10],

$$w = Sw^{-1}\ (\mu_1 - \mu_2) \quad (4)$$

where μ_1 denotes mean of normal class, μ_2 denotes mean of pathological class, S_w denotes within class scatter.

Mean of each class is defined as,

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{j\,i} \qquad j = 1, 2 \quad (5)$$

where N_j denotes number of samples in the class ‘j’. Within class scatter is the expected covariance of each class and is defined as,

$$S_w = \sum_{j=1}^{2} S_j \quad (6)$$

where

$$S_j = \sum_{i=1}^{N_j}(x_{j\,i} - \mu_j)(x_{j\,i} - \mu_j)^T \quad j = 1, 2 \quad (7)$$

Projections or transformations are obtained from (4) which represents 1-D invariant subspace of the vector space in which the transformation is applied. Once the transformation matrix is obtained data sets are transformed to the new vector space. This completes the training phase of the classifier. Assuming a set of M-dimensional samples $\{x^{(1)}, x^{(2)} \dots\dots\dots\dots\dots x^{(N)}\}$, N_1 of which belong to the class ω_1, N_2 to class ω_2 , we obtain a scalar y by projecting the samples x onto a line,

$$y = w^T x \quad (8)$$

Test vectors are also transformed according to (8) to the new vector space. The decision threshold used for the classification in the transformed space is the mean of normal class and pathological class mean, in the transformed space, and is given as ,

$$\mu_0 = \frac{\widetilde{\mu_1} + \widetilde{\mu_2}}{2} \quad (9)$$

where $\widetilde{\mu_1}$, $\widetilde{\mu_2}$ are the mean of normal and pathological classes in the transformed space and are estimated as,

$$\widetilde{\mu_j} = w^T \mu_j \qquad j = 1, 2 \quad (10)$$

Based on above threshold the given test sample, ‘t’ is classified as belonging to normal class if t ≥ μ_0 or to pathology class otherwise. Severity level of vocal fold pathology is estimated by finding the distance of the test vector from pathological class mean. Figure 9, Figure 10 and Figure 11 shows the LDA plot based on MFCC, LPCC and combination of MFCC and LPCC feature vectors into 1-D (dimensional) space respectively. One can observe a clear discrimination between normal and pathological voice samples in case of MFCC as compared to LPCC and combination of features, obtained by maximizing the ratio of between class scatter to within class scatter in LDA.

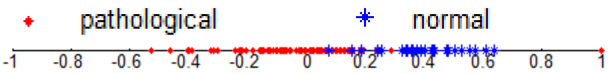

Figure 9. LDA 1-D plot for MFCC

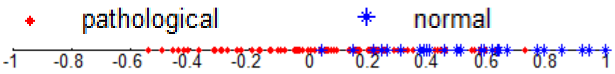

Figure 10. LDA 1-D plot for LPCC

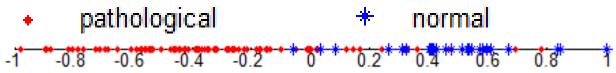

Figure 11. LDA 1-D plot for MFCC & LPCC

3.5 Voice samples

The voice samples for this study are taken from such a database distributed by Kay Elemetrics Corporation [11]. This database of acoustic records originally developed by Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab. Speech database consists of sustained phonation of the vowel /a/ sampled at sampling frequencies of 25 KHz or 50 KHz with 16-bit resolution.

The normal voice records are about 3 seconds long, whereas pathologic voice records are about 1 second long. The recordings were made in a controlled environment. In this study, 50 normal and 300 pathological voice samples pertaining to different types such as adductor, paralyses, leukoplakia, vocal nodule have been used. These include both male and female voice samples of different age group.

4. PERFORMANCE EVALUATION AND RESULTS

Performance of the classifier is evaluated as follows [1],

(1) True positive (TP): The classifier detected pathology when pathology was present.

(2) True negative (TN): The classifier detected normal when normal voice was present.

(3) False positive (FP): The classifier detected pathology when normal voice was present (false acceptance).

(4) False negative (FN): The classifier detected normal when pathology was present (false rejection).

(5) Sensitivity (SE): Likelihood that pathology will be detected given that it is present.

(6) Specificity (SP): Likelihood that the absence of pathology will be detected given that it is absent.

(7) Accuracy: The accuracy with which the classifier is able to classify the given sample to the correct group.

Sensitivity (%) = 100[TP/ (TP+FN)]

Specificity (%) = 100[TN / (TN+FP)]

Accuracy (%) = 100 [(TN+TP)/ (TN+TP+FN+FP)]

Simulation is done using MATLAB and the results of classification are depicted in Table 1. These results were calculated based on the number of samples used for testing.

Table 1: Results

Features	Sensitivity (%)	Specificity (%)	Accuracy (%)
MFCC	94	88	93.14
LPCC	82	76	81.14
MFCC &LPCC	89.33	92	89.71

5. DISCUSSIONS

Due to the fact that sounds were recorded in a controlled environment, no other preprocessing than the windowing is used. MFCC provides better results as compared to LPCC because, it takes human perception sensitivity with respect to frequencies into consideration. The bandwidth and the center frequency of the mel-filter banks roughly match the critical bandwidths of auditory neurons. MFCC has the advantage of energy compaction and dimensionality reduction, as most of the information lies in the first cepstral coefficients. MFCC allows modeling of the effects induced by the presence of pathology over the excitation (vocal folds) and the system (vocal tract). LPCC models the vocal tract response, and most of the voice pathologies affects the vocal folds (affects such as change in mass, elasticity and tension), hence less significant in identifying pathological voice from the normal ones. It is observed that the combination of features also give less accurate results as compared to MFCC alone. This clearly shows that, combination of features decreases the between class variance. Though LDA algorithm is found to be optimal when class distributions are Gaussian, it suffers from a small sample size problem when dealing with high dimensional data, resulting in a within class scatter matrix to be nearly singular. This problem can be eliminated by first reducing the dimensionality using principal component analysis and then using LDA on resulting data.

6. CONCLUSION

In this study two different set of features such as mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) were extracted from acoustic analysis of voiced speech of normal and pathological subjects. The LDA classifier is designed and implemented for pathologic voice and the performances of different set of features were compared. The MFCC have shown better classification results compared to other feature combinations. A scale based on MFCC could be used to evaluate the level of severity of the disease.

7. REFERENCES

[1] Anantha Krishna, K. Shama, and U. C. Niranjan, “k Means nearest neighbor classifier for voice pathology,” *Proceedings of IEEE India Annual Conference (INDICON’04)*, pp. 232–234, IIT Kharagpur, India, December 2004 .

[2] Kumara Shama, Ananthakrishna, and Niranjan U.Cholayya, “Study of Harmonic- to - Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology”, *EURASIP Journal on Advances in Signal Processing*, Volume 2007.

[3] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice-Hall Inc., 2002.

[4] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *Journal of the Acoustical Society of America*,vol.8, pp.185–190, Jan.1937.

[5] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, Md. Saifur Rahman, “Speaker Identification Using Mel Frequency Cepstal Coefficients”, ICECE, Dec 2004, Dhaka, Bangladesh.

[6] Remzi Serdar Kurcan, “Isolated word recognition from in-ear microphone data using hidden markov models(HMM)”, A thesis presented to the faculty of Naval Postgraduate School for the degree of Master of Science in electrical engineering and Master of Science in system engineering**.,**March 2006.

[7] H. P. Combrinck and E. C. Botha, “On The Mel-Scaled Cepstrum”, *Proceedings of the Seventh Annual South African Workshop on* Pattern Recognition, University of Pretoria, South Africa,1996.

[8] Juan I.Godino-Llorente, Santigo Agulera-Navarro,Pedro Gomez-Vilda,”LPC, LPCC and MFCC parameterization applied to the Detection of voice impairments”, Univeridad Politecnica de Madrid, Spain.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification (Second Edition), Wiley Interscience, New York, USA, ISBN: 0- 471-05669-3, 2000.

[10] Ricardo-Guzierrez-Osuna, “Introduction to pattern analysis” , Texas A&M University.

[11] Kay Elemetrics Corporation, “Disordered voice database.” 1994.