Vocal Fold Disorder Detection by applying LBP Operator on Dysphonic Speech Signal

Ghulam Muhammad, Zulfiqar Ali, Mansour Alsulaiman, Khalid Almutib
1Department of Computer Engineering, 2Department of Software Engineering
King Saud University
P.O. Box 51178, Riyadh 11543
SAUDI ARABIA
{ghulam, zuali, msuliman, muteb}@ksu.edu.sa
3Department of Electrical and Electronic Engineering
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750, Tronoh, Perak
MALAYSIA

Abstract: - The automatic system for voice pathology assessment is one of the active areas for researchers in the recent years due to its benefits to the clinicians and presence of a significant number of dysphonic patients around the globe. In this paper, a voice disorder detection system is developed to differentiate between a normal and pathological voice signal. The system is implemented by applying the local binary pattern (LBP) operator on Mel-weighted spectrum of a signal. The LBP is considered as one of the sophisticated techniques for the image processing. The technique also provided very good results for voice pathology detection during this study. The English voice disorder database MEEI is used to evaluate the performance of the developed system. The results of the LBP operator based system are compared with MFCC and found to be better than MFCC.

Key-Words: - LBP operator, MFCC, Vocal fold disorders, Sustained vowel, MEEI database, disorder detection system.

1 Introduction
Voice pathologies affect the vocal folds, and these disorders produce irregular vibrations in the vocal folds due to the malfunctioning of the voice box. Vocal fold pathologies exhibit variations in a vibratory cycle of the vocal folds due to their incomplete closure. The voice disorder also changes the shape of the vocal tract and produces irregularities in spectral properties. In the last decade, much research has been done on the automatic detection and classification of vocal fold diseases, and these tasks continue to require further investigation due to the lack of standard diagnosing approaches/equipment for voice disorders.

Automatic voice pathology detection can be accomplished by various types of features, which can be obtained by the long-term and short-term signal analysis. The long-term parameter can be derived by acoustic analysis [1, 2] of speech. The short-term parameters are further divided into two groups: parametric features and non-parametric features [3]. The parametric features represent the resonant structure of the human vocal cord and can be obtained by linear prediction coefficients (LPC) [4], and LPC-based cepstrum (LPCC) [5]. The non-parametric features mimic the behavior of the human auditory system and can be derived from the FFT based Mel-frequency cepstral analysis (MFCC) [6, 7]. Various classification techniques of pattern recognition such as HMM (Hidden Markov Model) [8, 9], GMM (Gaussian Mixture Model) [10], Vector Quantization (VQ) [11], Support Vector Machine (SVM) [12], Multilayer Perceptron (MLP) [13], Neural Networks (NN) [14], k-means Nearest Neighbors (KNNs) [15], Linear Discriminant Analysis (LDA) [16], and Learning VQ (LVQ) [17] are used to detect and/or classify voice disorders.

The vocal tract properties can be modeled using the all-pole model with the help of LPC features. These features represent the main vocal tract resonance properties in the acoustic spectrum. LPC highlights these formant structures for a speaker to
make a differentiation between them [18]. LPC coefficients and LPC based cepstral coefficients (LPCC) have been extracted from pathological and normal voices to develop voice disorder systems [19, 20, 21, 22]. LPC and LPCC have provided correct acceptance rates of 73% each, and the efficiencies were 85% and 80%, respectively, when the vocal folds edema was detected from other pathologies or normal voices [19]. Five pathologies, i.e., edema, nodules, polyps, cyst and paralysis are used in this research. The sustained vowel /a/ samples taken from the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Laboratory database [23]. The number of samples considered in [19] for pathological patients and normal persons is 67 and 53, respectively. MFCC features are also extracted to compare with LPC and LPCC, and the efficiency obtained for MFCC is 52%. The false acceptance rate of MFCC is 74%, which is 70% higher than LPC and 61% higher than LPCC, showing that MFCC is unable to differentiate between edema and other pathologies. In the second case, all pathologies are included in class one and normal speakers are grouped into class two, the performance of the MFCC is better than LPC and LPCC regarding efficiency and false acceptance rates. Therefore, it may be concluded that MFCC cannot discriminate between pathologies as well as it can discriminate between pathological and normal voices. It is also easier for a dysphonic expert to determine if a person’s voice is normal or not than to determine the type of disorder only by hearing him/her. Similarly, non-parametric features perform well in detection but not in classification. Parametric features, such as LPC and LPCC, resemble the resonant structure of the human vocal tract. Hence, LPC may perform well in the classification of pathological voices. In [20], LPC and MFCC features are inputted to KNN and SVM for the classification of three classes. The sustained vowel /a/ was recorded to develop the database at the Department of Medicine, Lithuania, and voice samples are labelled as healthy, nodular and diffuse. The best recognition rate with LPC is 67.31%, and the rate with MFCC was 73.08%.

Many voice pathology detection systems have been proposed using MFCC [20, 24, 25, 26, 27, 28], and MFCC has shown better performance in disease detection compared to LPC. In [27], various experiments are performed on a voice pathology detection system using the MEEI database. MDVP parameters and MFCC features are extracted from all sustained vowel /a/ voice samples and input into various modelling techniques. In [22], MFCC are extracted from the sustained vowel /a/ of dysphonic and normal voices from the database at the ENT department of Busan National University Hospital, Korea, and used with HMM, GMM, SVM and ANN for disease detection. The voice pathologies recorded for this database are cyst, edema, laryngitis, nodule, palsy, polyp, and glottis cancer. The highest pathology detection rate is 95.2% obtained with GMM.

The LPB features widely use in various image processing applications such as face recognition, gender detection, and texture classification, etc. [29, 30]. An approach to find cepstral speech feature by LBP operator is presented by Mansour et al. in [31]. The features were used for speech recognition, and named as local binary pattern cepstral coefficients (LBPC). The obtained results were good, and a recognition rate of 98.48% was achieved.

In this paper, we use LBP operator in a different way to extract the speech features to capture the voice characteristics of normal and pathological speech. The obtained results are good and comparable to existing techniques.

The rest of the paper is organized as follows: section 2 describes the speech corpus, existing and proposed speech feature, and classification technique. Section 3 provides experimental setup and obtained results, and finally; section 4 draws some conclusion.

2 Material and Method
This section provides discussion regarding the proposed speech feature obtained by applying LBP operator. The classification technique and existing speech features are also described.

2.1 Speech Corpus
MEEI database contained speech sample of normal person and dysphonic patients of various types. The database includes sustained vowel /a/ and running speech ‘Rainbow passage’ recorded by more than 700 patients and 53 normal speakers. The samples are recorded at 25 KHz and 50 K Hz sampling frequency with 16 bit resolution. A total number of speech samples is approximately 1400, recorded in control environment. The vowel contained only phonation part without any onset and offset
information. The vowel and text are recorded without any repetition. For normal persons, the sample duration is 2 to 3 sec, while for patients; the duration is 0.4 to 1.4 sec.

2.2 Mel-frequency Cepstral Coefficients
MFCC simulates human auditory mechanism and performs reasonably well under robust conditions. Fig.1 shows a block diagram of the MFCC calculation. First, digitized wave data is divided into overlapping frames, where the frame length is 16 milliseconds. This division is needed to analyze the speech in small pseudo-stationary segments.

The resultant frame is multiplied by a Hamming window to minimize the effect of spectral leakage. The Hamming window has almost zero values towards the both ends ensuring the continuity of the signal in successive frames. The Hamming window is given in Eq. 1

\[ s_n' = 0.54 - 0.46 \cos \left( \frac{2\pi(n-1)}{N-1} \right) s_n \]  

Fourier transformation (FT) is applied to the windowed signal to convert the time-domain signal into a frequency-domain signal (spectrum). Triangular band-pass filters (BPFs) are applied to divide the spectrum into certain frequency bands. The center frequencies of the BPFs are spaced on a Mel-scale, and the bandwidths correspond to well-known auditory perception phenomena called critical bandwidth. The relation between Mel-scale and linear scale (Hz) is almost linear up to around 800 Hz, and logarithmic beyond that (see Eq. 2).

\[ m = 2595 \log_{10} \left( 1 + \frac{f}{700} \right), \quad m = 1,2,\ldots, P \]  

In Eq. 2, \( m \) corresponds to Mel's index (\( P \) filters are used) and \( f \) refers to frequency in Hz. Therefore, by applying (\( P = 24 \)) BPFs, \( N \) points of the spectrum are converted to only 24 values. Log is applied to the 24 outputs to make the convolution components additive and to adjust the dynamic range in the spectrum. In this way, source excitation signal and vocal tract filter response become additive. The log outputs are then passed through discrete cosine transform (DCT) to de-correlate the components and reduce the dimension. The output of DCT is called MFCC. Typically, 0-th coefficient is ignored and 1\(^{st}\) to 12\(^{th}\) coefficients are retained to represent 12 MFCC.

2.3 Voice Features by applying Local Binary Pattern Operator (LBPVF)
The LBP voice features, abbreviated as LBPVF, extracted by applying LBP operator on the spectrum of speech signal. The block diagram of LBPVF extractor is depicted in Fig. 2. To extract the LBPVF, a signal divided into frames of 20msec with an overlapping of 10msec. Each frame is passed through the hamming window to taper its ends to zero. This will avoid the spectral leakage and to ensure the continuity in consecutive frames. The frequency domain representation is obtained by applying FT, and passed through 24 BPFs to divide the spectrum into certain frequencies. The last component of LBPVF extractor is the use of LBP operator.

The LBP operator applied on each value of the logarithmic Mel-weighted spectrum of a speech sample to get the features. The spectrum is 3-dimensional image, where, horizontal axis
represents time, vertical axis represents frequency, and gray scale values represent intensity of the speech. Use of the LBP operator on a 3x3 neighborhood is presented in Fig. 3. Each value of the spectrum will be replaced by a decimal value within the range of 0 to 255. The LBP operator compares the central value with all of its eight neighbors. The value of a neighbor will be replaced by 1 if it is greater or equal to the central value, otherwise, by 0. The operator then move from left of the central value in anti-clockwise direction to form eight-bit binary number. After converting to decimal number, it will replace the central value [30].

In a 3x3 window, horizontal values correspond to change in time, vertical values correspond to change in frequency, and diagonal values representing a change in both, time and frequency. Hence, the generated LBPVF can be looked as a summarization of all these changes.

2.4 Gaussian Mixture Model
GMM [32] is a state of the art modeling technique that copes more with the space of the features, rather than the time sequence of their appearance. Healthy and pathological persons are modeled by GMMs that represent, in a weighted manner, the occurrence of the feature vectors. The well-known method to model the speaker GMM is the Expectation-Maximization algorithm, where model parameters (Mean, variance and mixture coefficients) are adapted and tuned to converge to a model giving a maximum log-likelihood value.

The GMM model is given by the weighted sum of individual Gaussians

\[
p(X | \lambda) = \sum_{i=1}^{M} w_i g(X | \mu_i, \Sigma_i)
\]

where \(X\) is a D-dimensional continuous-valued data vector (i.e. measurements or features), \(w_i\) are the mixture weights, and, are the component Gaussian densities. Each component density is a D-dimensional Gaussian function of the form,

\[
g(X | \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)\right)
\]

with mean vector \(\mu_i\) and covariance matrix \(\Sigma_i\). The mixture weights satisfy the constraint \(\sum_{i=1}^{M} w_i = 1\). The model of the GMM is denoted as \(\lambda = (w_1, \mu_1, \Sigma_1, \ldots, w_M, \mu_M, \Sigma_M)\), \(i = 1, 2, 3, \ldots, M\).

3 Experiments and Results
The voice disorder detection system is developed by considering speech samples of sustained vowels /a/ available in MEEI database. The total number of voice disorder patients in MEEI database is 53 including male and female subjects. To make a balance between normal and pathological subjects, the same number of pathological subjects is considered from the database. Therefore, the total number of subjects to perform experiments becomes 106, and they are divided into two parts. The first part is 60% of the database and is used for the training of the developed disorder detection system.
The second part is 40% and is used for testing to evaluate the performance of the system. The performance of the automatic systems is evaluated by following parameters:

- True negative (TN): The system detects normal voice as normal voice.
- True positive (TP): The system detects disordered voice as disordered voice.
- False negative (FN): The system detects disordered voice as normal voice.
- False positive (FP): The system detects normal voice as disordered voice.
- Sensitivity (SE): The likelihood that the system detects disordered voice when the input is a disordered voice [33].
  \[ SE = \frac{TP}{TP + FN} \times 100 \]
- Specificity (SP): The likelihood that the system detects normal voice when the input is a normal voice [33].
  \[ SP = \frac{TN}{TN + FP} \times 100 \]
- Accuracy (Acc. %): The ratio between correctly detected files and the total number of files.

### 3.1 Results with LBPVF
The results of the developed voice disorder detection system with LBPVF by using different number of Gaussian mixtures are provided in Table 1.

<table>
<thead>
<tr>
<th>No. of Mixtures in GMM</th>
<th>Performance Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>4</td>
<td>80%</td>
<td>70%</td>
</tr>
<tr>
<td>8</td>
<td>60%</td>
<td>80%</td>
</tr>
<tr>
<td>16</td>
<td>70%</td>
<td>95%</td>
</tr>
<tr>
<td>32</td>
<td>80%</td>
<td>100%</td>
</tr>
<tr>
<td>50</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>70</td>
<td>70%</td>
<td>95%</td>
</tr>
<tr>
<td>90</td>
<td>95%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Sensitivity of 95% is achieved with 90 Gaussian mixtures represent that 95% of the disordered samples are recognized correctly. At 32 GMMs, the system obtained specificity of 100%; this means all normal samples are detected truly. The highest obtained accuracy is 92.50%. A comparison between different performance parameters is depicted in the Fig. 4.

![Fig. 4. A comparison between performance parameters for LBPVF](image)

### 3.2 Results with MFCC
The experiments for MFCC are conducted by using the same experimental setup. The results of performance parameters with MFCC are provided in the Table 2.

The maximum achieved sensitivity with MFCC is 95% with 8 Gaussian mixtures, and maximum specificity is 90% with 16 mixtures. The accuracy of the system built with MFCC is 90%. A comparison of performance parameters for MFCC is presented in Fig. 5.

<table>
<thead>
<tr>
<th>No. of Mixtures in GMM</th>
<th>Performance Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>4</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>8</td>
<td>95%</td>
<td>80%</td>
</tr>
<tr>
<td>16</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td>32</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>50</td>
<td>90%</td>
<td>50%</td>
</tr>
<tr>
<td>70</td>
<td>90%</td>
<td>75%</td>
</tr>
<tr>
<td>90</td>
<td>90%</td>
<td>70%</td>
</tr>
</tbody>
</table>
4 Conclusion

The specificity of 90% of both LBPVF and MFCC shows that both types of features recognized the disordered samples with same precision. On the other hand, the specificity with LBPVF is 10% better than that of MFCC. This shows that LBPVF detected the normal subject better than MFCC. The accuracy of the detection system with LBPVF is also better than MFCC by 2.50%. Overall, the LBPVF based voice disorder detection system performed well as compared to MFCC.

The LBP based features captures variation along the time axis when left and right neighbors are compared to central value, while, upper and lower neighbors provide variation along frequency axis. Similarly, neighbors along diagonal capture changes along time-frequency axis. This may be the reason that the LBP based features exhibit performance better than MFCC.

Acknowledgment

This work is supported by the National Plan for Science and Technology in King Saud University under grant number 12-MED2474-02. The authors are grateful for this support.

References:


