Diagnosis of Vocal Fold Pathology using Time-Domain Features and Systole Activated Neural Network

Paulraj M P  Sazali Yaacob  M.Hariharan
School of Mechatronic Engineering,
Universiti Malaysia Perlis,
Perlis, Malaysia.

Abstract- Due to the nature of job, unhealthy social habits and voice abuse, the people are subjected to the risk of voice problems. It is well known that most of vocal fold pathologies cause changes in the acoustic voice signal. Therefore, the voice signal can be a useful tool to diagnose them. Acoustic voice analysis can be used to characterize the pathological voices. This paper presents the detection of vocal fold pathology with the aid of the speech signal recorded from the patients. Time-domain features are proposed and extracted to detect the vocal fold pathology. The main advantages of this method are less computation time, possibility of real-time system development and it requires no transformation techniques (frequency transformation or time-frequency transformation). In order to test the effectiveness and reliability of the proposed time-domain features, a simple neural network model with systole activation function is proposed and trained by conventional back propagation (BP) algorithm. The classification accuracy of the proposed systole activated neural network is comparable with the results of neural network model with sigmoidal activation function. The simulation results show that the proposed systole activated neural network reduces the time taken for training the neural network.

Keywords – Voice disorders, Time-domain features, Artificial Neural Network, systole Activation Function.

I. INTRODUCTION

The vocal fold pathology has to be diagnosed in the early stage. To detect the vocal fold pathology, ENT clinicians and speech therapists use subjective techniques or invasive methods such as the direct inspection of the vocal folds and the observation of the vocal folds by endoscopic instruments [1]. These techniques are expensive, risky, time consuming, discomfort to the patients and require costly resources, such as special light sources, endoscopic instruments and specialized video-camera equipment. In order to circumvent the above problems, non-invasive methods have been developed to help the ENT clinicians and speech therapists for early detection of vocal fold pathology. In the bibliography, there are many algorithms have been found for the automatic detection of vocal fold pathology by means of long-time signal analysis [1-5]. In recent years, more modern approaches have been invented which use short-time speech analysis or Electroglottograph (EGG) signals [6,7]. The short-time acoustical features extracted from the EGG signal can be examined to depict the aspects of normal or abnormal vocal fold vibration motion. Various classifiers have been proposed such as multi-layer perceptron [8,9], learning-vector quantization [10], Hidden Markov models [11], linear discriminant analysis [12], Gaussian mixture models [1], and k- nearest neighbourhood classifier[13,14].

The proper diagnosis of vocal fold pathology is essential. This paper presents the detection of vocal fold pathology with the aid of the speech signal recorded from the patients. Time-domain features are proposed and extracted to detect the vocal fold pathology. In order to test the effectiveness and reliability of the proposed time-domain features, a simple neural network model with systole activation function is proposed and trained by conventional back propagation algorithm. The classification accuracy of the proposed systole activated neural network is comparable and reduces the neural network training time compared to neural network model with binary sigmoidal activation function.

II. SPEECH DATA

The speech data are taken from the commercial database distributed by Kay Elemetrics for the classification experiments [15]. The database contains approximately 1400 voice samples developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs. The acoustic samples are the sustained phonation of the vowel /ah/(1-3s) long from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders in different stages. All the speech samples were collected in a controlled environment and sampled
with 25 kHz or 50 kHz sampling rate and 16 bits of resolution. In order to test the effectiveness of the method and features, a total of 173 voices samples (173 abnormal+53 normal) are used according to Parsa and Jamieson [16] and downsampled to 16 kHz for our analysis. A database is created to assure that all the files have a diagnosis, and gender and age characteristics are uniformly distributed between the two classes as shown in table I.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Margin (years)</th>
<th>Mean (years)</th>
<th>Standard Deviation (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>26-58</td>
<td>22-52</td>
<td>8.49</td>
</tr>
<tr>
<td>Female</td>
<td>38.8</td>
<td>34.2</td>
<td>7.87</td>
</tr>
</tbody>
</table>

TABLE I
GENDER AND AGE DISTRIBUTION OF THE RECORDINGS IN THE CHOSEN SUBSET OF MEEI DATABASE, ADAPTED FROM [16]

III. TIME-DOMAIN FEATURES

Feature extraction from the speech signal plays very important role in the area of automatic detection of vocal fold pathology. Many feature extraction algorithms have been developed in the last three decades based on acoustic analysis of speech signals [1-7]. A great amount of acoustic parameters have been proposed and its effectiveness has been proven by experimental researches. This paper proposes a simple feature extraction method that extracts features from the time-domain energy of speech signal in order to detect the vocal fold pathology. All the features are calculated from short-time frames extracted from the speech signals. The short-time frame length is selected as 64 ms (1024 samples per frame), with an overlap of 50% between adjacent frames.

Fig. 1(a) Time-domain energy plot of normal speech signal

Fig. 1(b). Time-domain energy plot of pathological speech signal

Figure 1(a) and (b) shows the time-domain energy plot of normal and pathological speech signal. These energy plots clearly indicate the difference between the normal and pathological speech signal.

Fig. 2 Illustration the energy peaks of a speech signal in one short-time windows

Fig. 2 shows the energy peaks of a speech signal. Consider a speech signal is divided into N number of short-time frames. Consider the ith frame, from the ith short-time frame the following features are extracted.

1. Three maximum energy levels are computed.
   \( f_{i1}, f_{i2}, f_{i3} \)
2. Three maximum energy levels multiplied with its locations \( n_1, n_2, n_3 \) are found.
   \( f_{i4}, f_{i5}, f_{i6} \)
3. Three straight lines are drawn between three maximum energy peaks. The areas \( A_1, A_2, \text{ and } A_3 \) under the straight line between the energy peaks are computed.
   \( f_{i7}, f_{i8}, f_{i9} \)
4. The total energies of each short frame are calculated.
   \( f_{i10} \)
5. Energy difference between first and second maximum, similarly first and third maximum, second and third maximum are calculated.
   \( f_{i11}, f_{i12}, f_{i13} \)
6. Energy difference of first and second maximum between adjacent frames, similarly first and third maximum between adjacent frames, second and third maximum between adjacent frames are calculated.
   \( f_{i14}, f_{i15}, f_{i16} \)
After extracting the above features from the $i^{th}$ frame, the feature set can be represented as

\[ F = [f_{i1}, f_{i2}, f_{i3}, f_{i4}, f_{i5}, f_{i6}, f_{i7}, f_{i8}, f_{i9}, f_{i10}, f_{i11}, f_{i12}, f_{i13}, f_{i14}, f_{i15}, f_{i16}] \]

\[ i = 1, 2, \ldots, N \]

Finally, from the feature matrix the variation of each features are calculated using the following equation 1.

\[ V_j = \frac{1}{N-j} \sum_{i=1}^{N} |F_{i,j} - F_{i+1,j}| \]

\[ j = 1, 2, \ldots, 16 \]

\[ i = 1, 2, \ldots, N \]

where \( V_j \) represents the variation of each features.

Table II shows difference of mean and standard deviation of the proposed features of normal and pathological speech signal.

### TABLE II

<table>
<thead>
<tr>
<th>Features</th>
<th>Mean Normal</th>
<th>Pathological Normal</th>
<th>Mean Pathological</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>3.59</td>
<td>6.58</td>
<td>1.36</td>
</tr>
<tr>
<td>F2</td>
<td>3.80</td>
<td>7.79</td>
<td>1.47</td>
</tr>
<tr>
<td>F3</td>
<td>3.72</td>
<td>7.76</td>
<td>1.51</td>
</tr>
<tr>
<td>F4</td>
<td>3.70</td>
<td>7.70</td>
<td>1.49</td>
</tr>
<tr>
<td>F5</td>
<td>75.76</td>
<td>77.99</td>
<td>7.67</td>
</tr>
<tr>
<td>F6</td>
<td>70.55</td>
<td>73.36</td>
<td>6.71</td>
</tr>
<tr>
<td>F7</td>
<td>68.42</td>
<td>72.60</td>
<td>7.73</td>
</tr>
<tr>
<td>F8</td>
<td>83.25</td>
<td>105.25</td>
<td>9.77</td>
</tr>
<tr>
<td>F9</td>
<td>79.11</td>
<td>93.31</td>
<td>9.75</td>
</tr>
<tr>
<td>F10</td>
<td>81.67</td>
<td>98.71</td>
<td>11.10</td>
</tr>
<tr>
<td>F11</td>
<td>74.79</td>
<td>9.80</td>
<td>19.69</td>
</tr>
<tr>
<td>F12</td>
<td>83.79</td>
<td>10.70</td>
<td>22.56</td>
</tr>
<tr>
<td>F13</td>
<td>55.68</td>
<td>7.03</td>
<td>14.89</td>
</tr>
<tr>
<td>F14</td>
<td>4877.74</td>
<td>289.33</td>
<td>16910.29</td>
</tr>
<tr>
<td>F15</td>
<td>1641.99</td>
<td>319.49</td>
<td>20102.22</td>
</tr>
<tr>
<td>F16</td>
<td>-3370.52</td>
<td>650.43</td>
<td>16060.69</td>
</tr>
</tbody>
</table>

From the table II, it can be observed that the proposed features clearly differentiate the normal and pathological signal and shows that the features can be used to diagnose vocal fold pathology clinically. The extracted features are used as input of the proposed neural network classifier for classifying speech signal as normal or pathological.

### IV. NEURAL NETWORK CLASSIFIER

Artificial Neural Network (ANN) provides an alternative form of computing that attempts to mimic the functionality of the brain [17]. One of the most used learning methods in ANN is back propagation. The Back propagation method is a learning procedure for multilayered feedforward networks. A three layer neural network model with 16 input neurons, 20 hidden neurons and one output neuron is developed. 50%, 60% and 70% of the samples are randomly chosen from the database and used as training patterns and tested with 100% of the samples. The initial weights for the neural network are randomized between -0.5 and 0.5 and normalized [18]. A trial weight set consist of 10 sets of randomized weight samples are considered. The sum squared error tolerance is fixed as 0.05 and testing tolerance is fixed as 0.05. The learning rate and momentum factor are chosen as 0.08 and 0.6 respectively. The hidden and output neurons have a bias value of 1.0 and are activated by binary sigmoidal activation function of the form

\[ f(x) = \frac{1}{1 + \exp(-x/q)} \]

Similarly, the input and hidden neurons are activated by systole activation function of the form:

\[ f(x) = k_1 x e^{-k_2 x^2} \]

Through experimentation, the values of \( k_1 \) and \( k_2 \) are chosen resulting in better classification rate. The two neural network models are developed and trained by conventional back propagation algorithm with two activation functions namely, binary activation function and proposed systole activation function. The maximum number of epoch is fixed as 20000. The mean epoch for neural network training and mean classification rate (CR) in percentage (%) are tabulated in Table III and IV.

From the Table III and IV, it can be observed that the mean number of epoch for training using conventional back propagation algorithm with binary activation function are high and compared to BP with systole activation function. The classification rates of the neural network trained with systole activation function are comparable with binary activation. The results show that the proposed neural network with systole activation function provides comparable classification rate and reduces the time taken for the neural network training.
A simple time-domain feature extraction method to detect the vocal fold pathology is proposed with the aid of the speech signal recorded from the patients. In order to test the effectiveness and reliability of the proposed time-domain features, a simple neural network model with systole activation function is proposed and compared with conventional BP algorithm with binary activation function. The classification accuracy of the proposed systole activated neural network is comparable with the results of neural network model with sigmoidal activation function. The simulation results show that the proposed systole activated neural network reduces the time taken for training the neural network and the proposed method can be used as a valuable tool for researchers and speech pathologists to detect the vocal fold pathology.

ACKNOWLEDGMENT

This work is supported by the grant: FRGS-9003-00043 from the Ministry of Higher Education of Malaysia. The authors wish to thank our Vice Chancellor Prof. Dato’ Dr. Kamarudin Hussin, for his valuable support during the research work.

REFERENCES


