IJBSCHS(2009-14-2-8) Biomedical Soft Computing and Human Sciences, Vol.14,No.2,pp.55-60 (2009)
Special Issue: Bio-sensors: Data Acquisition, Processing and Control
[Original article] Copyright©1995 Biomedical Fuzzy Systems Association
(Accepted on 2009.01.01)

Diagnosis of Voice Disorders using Mel Scaled WPT and Functional Link Neural Network

Paulraj M P, Sazali Yaacob, and M. Hariharan

School of Mechatronics Engineering, Universiti Malaysia Perlis, Perlis, Malaysia

The paper was received on March 31, 2008.

Abstract: *Nowadays voice disorders are increasing dramatically due to the modern way of life. Most of the voice disorders cause changes in the voice signal. Acoustic analysis on the speech signal could be a useful tool for diagnosing voice disorders. This paper applies Mel-scaled wavelet packet transform (Mel-scaled WPT) based features to perform accurate diagnosis of voice disorders. A Functional Link Neural Network (FLNN) is developed to test the usefulness of the suggested features. Two simple modifications are newly proposed in the FLNN architecture to improve the classification accuracy. In the first architecture, a hidden layer is newly introduced in a FLNN and trained by Back Propagation (BP) procedure. In the second architecture, the Integral and Derivative controller concepts are introduced to the neurons in the hidden layer and the network is trained by BP procedure. The performance is compared with conventional neural network model. The results prove that the proposed FLNN gives very promising classification accuracy and suggested features can be employed clinically to diagnose the voice disorders.*

Keywords Acoustic Analysis, Voice Disorders, Mel scaled Wavelet Packet Transform (Mel scaled WPT), Functional Link Neural Network(FLNN)

1. Introduction

Voice is a highly multi-variate component of speech and the need to provide for its quantitative description has led to the development of clinical tools. With the rapid development of signal processing technique, vocal or voice signal can be used for the detection of voice disorders. Vocal or voice signal information plays an important role to understand the process of vocal fold pathology formation. In the last years, lot of works have been carried out on the automatic detection and classification of voice pathologies by means of acoustic analysis, parametric and non-parametric feature extraction, automatic pattern recognition or statistical methods [1-4]. The feature extraction using the above methods includes handling of overlapping windows of a speech signal and large number of computations performed during the feature extraction phase. In the recent years, wavelet transform has been used to analyze all kinds of problems in signal and image processing. Speech processing is one of these areas. Speech is a highly non-stationary signal; Fourier transform is not a very useful tool for analyzing non-stationary signals as the time domain information are lost while performing the frequency transformation. When looking at a Fourier transform of a signal, it is impossible to tell when a particular event took place. Wavelet transform approach is a good tool for the analysis of non-stationary signals, as it is useful in localizing a transform approach is a good tool for the analysis symptom both in time and frequency scale. Hence, wavelet analysis has the potential for the identification of voice disorders. In [5], authors have presented a procedure to identify pathological disorders of larynx using wavelet analysis. P.S Bhat et al. [6] have proposed a method for the classification and analysis of speech abnormalities based on wavelet analysis and artificial neural network.

S. Datta and his co-workers have developed a new filter structure using Mel-like Admissible Wavelet Packet Structure for speech recognition [7-9]. These filters have the advantage of having frequency bands spacing similar to the Mel scale. Wavelet Packet has the advantage that it can segment the frequency axis and has uniform translation in time. This property of partitioning of the frequency axis is used for the realization of conjugate mirror filter structure similar to that of Mel filter.

The aim of this paper is to apply Mel-scaled wavelet packet filter for extracting the features from the voice or vocal signal. A Functional Link Neural Network is developed to test the efficacy of the feature vector derived by using Mel-scaled Wavelet packet filter. The experimental results indicate that FLNN gives very promising classification accuracy and the proposed features can be used as an additional acoustic indicator for the diagnosis of voice disorders.

02600, Jejawi, Arau,Kangar,
Perlis, Malaysia.
Phone No. 006049798918 Fax No. 006049798142
Email: paul@unimap.edu.my

2. Voice Disorders

A voice is termed "disordered" when the vocal quality of an individual is changed in such a way that it is thought to be abnormal to the listener. Voice disorders arise due to physiological disorder or psychological disorder, accident, misuse of the voice, or surgery affecting the vocal folds. Voice disorders can be classified into two main categories such as Functional voice disorders and Organic voice disorders. Early detection and treatment of laryngeal tumours, depends on factors such as health awareness among the general public and on the experience of speech therapists and ENT clinicians. Impairment of vocal function can have a major impact on the quality of life, severely limiting communication at work and affecting all social aspects of daily life

Disorders that affect the larynx cause changes in the human's voice quality. These disorders have to be diagnosed and treated at an early stage, especially larynx cancer. The voice signal can be a useful tool to diagnose them. Acoustic analysis of speech signal is a non-invasive technique that has been proved to be an effective tool for the diagnosis of voice disorders. This paper presents feature extraction based on Mel scaled WPT and Functional Link Neural Network classifier for the diagnosis of voice disorders, which will help speech specialists to perform accurate diagnosis of voice disorders.

3. Description of Dataset

The speech samples are taken from the commercial database distributed by Kay Elemetrics for the classification experiments [10]. 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.

Table 1. Distribution of Male and Female Speakers

S. No.	Name of Voice Disorders	Total Speakers	Male Speakers	Female Speakers
1.	Paralysis	67	33	34
2.	Vocal Fold Edema	44	11	33
3.	Gastric Reflux	48	24	24
4.	Normal Condition	53	21	32

The database includes a total of 710 speech samples (657 abnormal+53 normal). All the speech samples were recorded in a controlled environment and sampled with 25 kHz or 50 kHz sampling rate and 16 bits of resolution. The database used in this paper, consists of 53 normal and 159 pathological speakers. The speech samples are down sampled to 16 kHz giving an 8 kHz bandwidth signal. The speech samples of patients having Vocal Fold Edema, Gastric Reflux, and vocal fold paralysis voice disorder problem are considered for the classification problem. The distribution of male and female speakers in normal and pathological condition is shown in Table 1.

4. Design of Filter using Wavelet Transform

4.1. Wavelet Transform

The wavelet Transform provides time frequency representation of the signal. It decomposes signal over dilated and translated wavelets. A wavelet is a waveform of effectively limited duration that has an average value of zero. The Wavelet Transform is defined as the convolution of a signal $f(t)$ with a wavelet function $\psi(t)$ shifted in time by a translation parameter and dilated by a scale parameter. The general definition of the wavelet transform is given as:

$$W(a,b)=\int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{a}}\psi^{*}\left(\frac{t-b}{a}\right)dt \qquad (1)$$

where a and b are real and * denote complex conjugate and $\psi(t)$ is the wavelet function. The Wavelet transform uses multi-resolution technique by which different frequencies are analyzed with different resolution. The Discrete Wavelet Transform (DWT) of a sampled sequence $f_n = f(nT)$ with sampling period T is computed as:

$$DWT\{f_n;a,m\}=\sum_{n} f_n \psi^{*}_{a,m}(n) \qquad (2)$$

where m and n are integers. The value of a is equal to 2. The discrete wavelet transform applies to discrete-time signals-both time and time-scale parameters are discrete. The signal representation is not complete if the wavelet decomposition is computed up to a scale a^j. The information corresponding to the scales larger then a^j is required, which is computed by a scaling filter and is given by:

$$SFf[n,a^{j}]=\sum_{m=0}^{N-1} f[m]\,a^{-j/2}\,\phi^{*}\left(\frac{m-n}{a^{j}}\right) \qquad (3)$$

where $\Phi(n)$ is the discrete scaling filter.

4.2. Wavelet Packets

In DWT decomposition procedure, a signal is decomposed into two frequency bands such as lower

frequency band (approximation coefficients) and higher frequency band (detail coefficients). Low frequency band is used for further decomposition. Hence DWT gives a left recursive binary tree structure. In Wavelet Packet (WP) decomposition procedure, both lower and higher frequency bands are decomposed into two sub-bands. Thereby Wavelet packet gives a balanced binary tree structure. In the tree, each subspace is indexed by its depth *i* and the number of subspaces *p*. The two wavelet packet orthogonal bases at a parent node (*i*,*p*) are given by the following forms

$$\psi_{i+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} l[n]\psi_i^p(k-2^i n) \quad (4)$$

where *l[n]* is a low pass(scaling) filter.

$$\psi_{i+1}^{2p+1}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_i^p(k-2^i n) \quad (5)$$

where *h[n]* is the high pass(wavelet) filter. Wavelet packet decomposition helps to partition the high frequency side into smaller bands which cannot be achieved by using general discrete wavelet transform. A complete set of bases is obtained by Wavelet packet decomposition. From these bases, a best basis is to be selected and it is very important for portioning the frequency axis. However, this method suffers from the problem of giving different tree structures for a small shift in the signal. In order to overcome the above problem, admissible wavelet packet decomposition was proposed [7-9]. In this paper, the usefulness of Mel-like admissible wavelet packet tree structure is investigated for the diagnosis of voice disorders. The frequency bands obtained from wavelet packet decomposition closely follows the Mel scale. Table 2 gives the lower cut off frequency, higher cut off frequency and bandwidth of all the 24 filters obtained by using the wavelet packet decomposition [8].

5. Feature Extraction using Mel-scaled WPT based filters

The speech samples are sampled at 16 kHz giving an 8 kHz bandwidth signal. The speech signals are filtered with the 24 Mel-scale wavelet packet filters designed by 'Daubechies' wavelet 4 and the energy in each frequency sub-band are calculated. The calculated energy is divided by the number of wavelet coefficients in the corresponding sub-band thereby giving average energy per wavelet coefficients per sub-band (F_p). It is given by:

$$F_p = E_p / N_p \quad (6)$$

where E_p is the total energy in the sub-band *p* and N_p is the number of wavelet coefficients. The log of energy is then calculated giving a total of 24 coefficients.

Table 2. Filter and their corresponding frequency bands achieved by wavelet packet decomposition [8]

Filter Number	Lower cut off frequency (Hz)	Upper cut off frequency (Hz)	Band Width (Hz)
1	0	125	125
2	125	250	125
3	250	375	125
4	375	500	125
5	500	625	125
6	625	750	125
7	750	875	125
8	875	1000	125
9	1000	1125	125
10	1125	1250	125
11	1250	1375	125
12	1375	1500	125
13	1500	1750	250
14	1750	2000	250
15	2000	2250	250
16	2250	2500	250
17	2500	2750	250
18	2750	3000	250
19	3000	3500	500
20	3500	4000	500
21	4000	5000	1000
22	5000	6000	1000
23	6000	7000	1000
24	7000	8000	1000

Discrete Cosine Transform (DCT) is applied on these 24 coefficients and the first 12 DCT coefficients are taken as features for training the neural network. The DCT is applied to reduce the redundant information and to model the variation of the logarithmic energy. The complete process of feature extraction is shown in Figure 1.

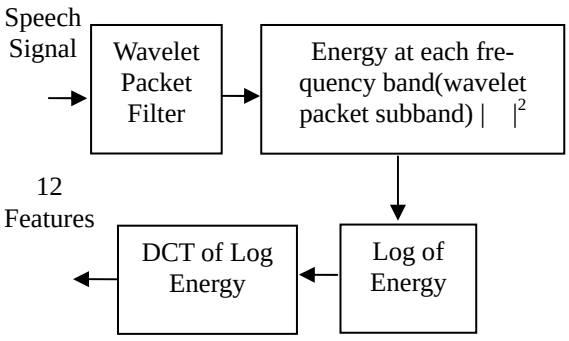

Figure 1. Feature Extraction

6. Neural Network Architecture

Artificial Neural Network (ANN) provides alternative form of computing that attempts to mimic the functionality of the brain [11]. Feedforward neural networks have been the subject of intensive research efforts in recent years because of their interesting learning and generalization properties. The back propagation method is a learning procedure for multilayered feedforward neural networks. In this paper, A FLNN based classifier is developed to test the efficacy of the proposed features and its performance is compared with the conventional three layer neural network.

6.1. Conventional Three Layer Network

A three layer neural network with 12 input neurons, 10 hidden neurons and one output neuron is considered. The hidden neurons have a bias value of 1.0 and the output and hidden neurons are activated by binary sigmoidal activation function. 148 samples are randomly chosen from the database and used as training patterns and tested with 212 samples. Five trials were performed. For each trial, the network is trained with 20 different sets of initial weight samples and randomized between -0.5 and 0.5. The sum squared error tolerance is fixed as 0.1 and testing tolerance is fixed as 0.1. The learning rate and momentum factor are chosen as 0.1 and 0.9 respectively.

6.2. Functional Link Neural Network

The FLNN proposed by Pao [12] can be used for function approximation and pattern classification with faster convergence and lesser computational complexity than a MLP network. In a FLNN, the need of hidden layer is removed. In contrast to linear weighting of the input pattern produced by the linear links of a MLP, the functional link acts on an element or the entire pattern itself by generating a set of linearly independent functions. Researchers have widely used the FLNN concept for noise cancellation, prediction, system identification and control [13-15]. In this paper, FLNN model is developed for the classification of normal and pathological voices. Initially, the conventional FLNN is trained using the suggested features for 15 times and it did not converge for the tolerance level of 0.1. In order to overcome this problem, two simple modifications are proposed in the FLNN architecture. The introduction of hidden layer and the extension of functional link concept enhance the performance of the conventional FLNN.

6.2.1. Modification I

In the first architecture, a hidden layer is newly introduced in a FLNN and trained by Back Propagation (BP) procedure. Without loss of generality and simplicity a FLNN having *'m'* number of neurons in the hidden layer and *'p'* number of neurons in the output layer is considered. The network has *(2n-1)* number of neurons in the input layer. The first *'n'* neurons in the input layer receive input signals $x_1, x_2, x_3, \ldots x_n$ and the remaining *(n-1)* neurons receive the functional composition of the input signal namely $x_1x_2, x_2x_3, x_3x_4, \ldots, x_{n-1}x_n$ [12].

6.2.2. Modification II

This architecture consists of a three layer network with input, hidden and output layers. In this architecture, the Integral and Derivative controller concepts are introduced to the neurons in the hidden layer. For derivative feedback, the derivative constant k_d is multiplied with derivative of the output of the hidden neuron (difference between the present output of the hidden neuron and preceding output of the hidden neuron). For integral input, the integral constant k_i is multiplied with the cumulative of past outputs of the hidden neuron. These two parameters for the $(k+1)^{th}$ epoch is,

$$D_{k+1} = k_d * (z_j(k) - z_j(k-1)) \quad (7)$$

$$I_{k+1} = k_i * \sum_{k=1}^{k} z_j(k) \quad (8)$$

where $z_j(k)$ is the output signals for the hidden units at the k^{th} epoch. The constants k_d and k_i are chosen such that the training time is minimum. The hidden and output neurons are activated by binary sigmoidal activation function as given in Equation (9),

$$f(x) = [1/(1+\exp(-x/q))] \quad (9)$$

The input vector at sample time is defined as:

$$x = [x_1\ x_2\ x_3\ \ldots\ x_n] \quad (10)$$

The functional link input neurons of *(n+1)* to *(2n– 1)* units are generated as:

$$x_{n+i} = x_i\, x_{i+1},\ x_{i+1}x_{i+2}, x_{i+2}x_{i+3}, \ldots, x_{n-1}\, x_n \quad \text{for } i = 1,2,3,\ldots,n-1 \quad (11)$$

The net input to the j^{th} hidden neuron z_j is:

$$z_{in\_j} = \sum_{i=1}^{2n-1} u_{ij}\, x_i \quad (12)$$

and $$z_j = f(z_{in\_j}) \quad \text{for } j = 1,2,3, \ldots, m \quad (13)$$

In the modification I, a three layer FLNN with 12 input neurons and 23 functional link neurons in the input layer, 10 neurons in the hidden layer and 1 neuron in the output layer is considered. The 12 DCT coefficients ($x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+12}$) along with the derived 23 functional composition signals (13 additional input features derived from the 12 original input features) are used as the input feature to the FLNN.

In the modification II, a three layer FLNN with 12 input neurons and 23 functional link neurons in the input layer, 10 hidden neurons in the hidden layer and 1 neuron in the output layer is considered. Further, the

Integral and Derivative outputs from the hidden neurons are fedback to itself. The hidden neurons have a bias value of 1.0 and the output and hidden neurons are activated by binary sigmoidal activation function. 148 samples are randomly chosen from the database and used as training patterns and tested with 212 samples. Five trials were performed. For each trial, the network is trained with 20 different sets of initial weight samples between -0.5 and 0.5. The sum squared error tolerance is fixed as 0.2 and testing tolerance is fixed as 0.1. The learning rate and momentum factor are chosen as 0.1 and 0.9 respectively. The constants k_d and k_i are chosen as 0.1 and 0.005 respectively.

7. Results and Discussion

ENT clinicians and speech therapists use subjective techniques or invasive methods such as evaluation of voice quality by the specialist doctor's direct inspection and the observation of vocal folds by endoscopy techniques. These techniques provide inconvenience to the patients and depend on expertise of medical doctors. Many research works have been carried out for early detection of voice disorders by means of acoustic analysis, parametric and non-parametric feature extraction, and automatic pattern recognition. This paper proposes Mel-scaled WPT filter for extracting the features from the voice or vocal signal and FLNN based classifier for the classification of pathological voices. The conventional three layer neural network model and the FLNN model are trained by using conventional back propagation algorithm and the results are tabulated in Table 3 and Table 4. For the evaluation of the accuracy of the neural network based classifier, 70% of the speech samples were used for the training. 30% of the remaining samples together with trained samples were used for the testing process.

Table 3. Neural Network Training Results- Mean Epoch

Trail No.	CBP	FLNN-I	FLNN - II
1	6435	2943	1351
2	7319	3181	1315
3	6857	3164	1239
4	6286	3646	1364
5	6738	3001	1413

Table 4. Neural Network Training Results- Mean Classification Rate (%)

Trail No.	CBP	FLNN-I	FLNN - II
1	94.17	96.59	96.56
2	94.60	96.66	96.56
3	94.58	96.67	96.49
4	94.39	96.53	96.67
5	94.69	96.53	96.65

From the Table 3, it can be observed that the mean epoch for the conventional BP training is high compared to proposed FLNN architectures. The mean epoch for training the FLNN- II architecture is low compared to FLNN-I architecture and the conventional BP. From Table 4, it can be observed that the mean classification rates (%) for the proposed FLNN architectures are higher when compared to the conventional BP. The mean classification rate (%) of FLNN-II architecture is comparable with FLNN-I but better than the conventional BP. The results prove that the proposed acoustic features and FLNN based classifier help speech specialists and ENT clinicians to perform early diagnosis of voice disorders.

8. Conclusions

This paper suggests additional acoustical indicators for the automatic diagnosis of voice disorders. Mel-scaled wavelet packet filter is designed to derive these acoustical features. A FLNN based classifier is developed to test the performance of the proposed features and two modifications are proposed in its architecture. The advantage of using Mel-scaled WPT is the elimination of using overlapping windows, which reduces the number of computations performed during feature extraction phase. The results prove that the suggested features can be employed clinically to diagnose voice disorders.

Acknowledgement

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

[1] Carlos H.E., Mercedes F.R., Santiago A.N., Pedro G.V., Godino.Llorente J.I. (2000) "Diagnosis of vocal and voice disorders by the speech signal", Proc. of IEEE-INNS-ENNS International Joint Conference on Neural Networks,vol.4, pp, 253-258.

[2] Boyanov, B., Stefan, H. (1997) "Acoustic analysis of pathological voices: A voice analysis system for the screening of laryngeal diseases", Proc. of IEEE International Conference on Engineering in Medicine and Biology Society, vol.16, pp, 74-82.

[3] Dimitar, D.(1993) "Acoustic model and evaluation of pathological voice production", Third International Conference on Speech Communication and Technology, EUROSPEECH'93, pp, 1969-72.

[4] Godino Llorente, J.I., Gomez, V. P., Blanco,V. M. (2006) "Dimensionality reduction of a pathological voice quality assessment system

based on Gaussian mixture models and short-term cepstral parameters, IEEE Transactions on Biomedical Engineering, vol.53, pp,1943-1953.

[5] Nayak, J., Bhat, P.S. (2003) "Identification of voice disorders using speech samples", Conference on Convergent Technologies for Asia-Pacific Region, TENCON(2003), vol.3, pp, 951-953.

[6] Nayak, J., Bhat, P. S.,Acharya, R., Aithal, U.V. (2005) "Classification and analysis of speech abnormalities", ITBM-RBM, vol.26, pp, 319-327.

[7] Farooq, O., Datta, S. (2004) "Wavelet based robust sub-band features for phoneme recognition", Proc. of the IEE Vis. Image Signal Processing, vol. 151, No. 3, pp, 187-193.

[8] Farooq, O., Datta, S. (2001) "Mel filter like admissible wavelet packet structure for speech recognition", Proc. of IEEE Signal Processing Letters, vol. 8(7), pp, 196-198.

[9] Long, C. J., Datta, S. (1996) "Wavelet based feature extraction for phoneme recognition", Proc. 4th International Conference on Spoken Language, vol. 1, pp, 264-267.

[10] Massachusetts Eye and Ear Infirmary.(1994) "Voice disorders database", version 1.03(CDROM), Kay Elemetrics Corporation, Lincoln Park, NJ, USA.

[11] Sivanandam, S.N., Paulraj, M. (2003) "Introduction to artificial neural networks", Vikhas Publication, India.

[12] Pao, Y. H. (1989) "Adaptive Pattern Recognition and Neural Networks Reading", MA: Addison-Wesley.

[13] Dash, P. K., Satpathy, H. P., Liew, A. C and Rahman, S. (1997) "A real-time short-term load forecasting system using functional link network", IEEE Transactions on Power Systems, Vol. 12, No. 2, pp,675-680.

[14] Patra, J.C. and Pal, R.N. (1995) "Functional link artificial neural network-based adaptive channel equalization of nonlinear channels with QAM signal", IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp, 2081-2086.

[15] Patra, J. C., and Alex C. K. (2002) "Nonlinear dynamic system identification using Chebyshev functional link artificial neural networks", IEEE Transactions on Systems, Man, and Cybernetics, vol. 32(4), pp, 505-511.

Paulraj M P is currently an Associate Professor at Universiti Malaysia Perlis, Malaysia. His research interests are in the area of artificial intelligence, Fuzzy systems, Speech processing, acoustic engineering and biomedical signal processing applications. He has published more than 150 papers in referred journals and conferences. He has authored a book titled Introduction to Artificial Neural Networks. He is a member of the Institution of Engineers, India and MISTE, India.

Sazali Yaacob is currently working as a professor at Universiti Malaysia Perlis, Malaysia. He has published more than 150 papers in Journals and Conference Proceedings. His research interests are in Artificial Intelligence applications in the fields of acoustics, vision and robotics. He received Charted Engineer status by the Engineering Council, United Kingdom in 2005 and also a member of the IET (UK).

M. Hariharan has obtained BE in Electrical and Electronics Engineering from Bharathiar University, India and Master degree from Anna University, India. He is currently doing Ph.D in Mechatronic Engineering at Universiti Malaysia Perlis, Malaysia. His research interests include speech signal processing, wavelet transform, image processing and artificial neural network. He is a graduate student member of IEEE, USA.