A Novel Method for Feature Extraction in Vocal Fold Pathology Diagnosis

Vahid Majidnezhad and Igor Kheidorov

Department of Computer Engineering, Shabestar Branch,
Islamic Azad University, Shabestar, Iran
vahidmn@yahoo.com, ikheidorov@sakrament.com

Abstract. Acoustic analysis is a proper method in vocal fold pathology diagnosis so that it can complement and in some cases replace the other invasive, based on direct vocal fold observation, methods. There are different approaches for vocal fold pathology diagnosis. These algorithms usually have two stages which are Feature Extraction and Classification. While the second stage implies a choice of a variety of machine learning methods, the first stage plays a critical role in performance of the classification system. In this paper, three types of features which are Energy and Entropy resulting from the Wavelet Packet Tree and Mel-Frequency-Cepstral-Coefficients (MFCCs), and also their combination are investigated. Finally a new type of feature vector, based on Energy and Mel-Frequency-Cepstral-Coefficients, is proposed. Support vector machine is used as a classifier for evaluating the performance of our proposed method. The results show the priority of the proposed method in comparison with other methods.

Keywords: Vocal fold pathology diagnosis, Wavelet Packet Decomposition, Mel-Frequency-Cepstral-Coefficients (MFCCs), Energy, Entropy, Support Vector Machine (SVM).

1 Introduction

Vocal signal information often plays an important role for specialists to understand the process of vocal fold pathology formation. In some cases vocal signal analysis can be the only way to analyze the state of vocal folds. Nowadays diverse medical techniques exist for direct examination and diagnostics of pathologies. Laryngoscopy, glottography, stroboscopy, electromyography and videokimography are most frequently used by medical specialists. But these methods possess a number of disadvantages. Human vocal tract is hardly-accessible for visual examination during phonation process and that makes it more problematic to identify a pathology. Moreover, these diagnostic means may cause patients much discomfort and distort the actual signal, that may lead to incorrect diagnosis as well [1-4].

Acoustic analysis as a diagnostic method has no drawbacks, peculiar to the above mentioned methods. It possesses a number of advantages. First of all, acoustic analysis is a non-invasive diagnostic technique that allows pathologists to examine

B. Godara and K.S. Nikita (Eds.): MobiHealth 2012, LNICST 61, pp. 96–105, 2013.

many people in short time period with minimal discomfort. It also allows pathologists to reveal the pathologies on early stages of their origin. This method can be of great interest for medical institutions.

In recent years a number of methods were developed for segmentation and classification of speech signals with pathology. The general scheme of vocal fold pathology diagnosis is illustrated in Fig. 1.

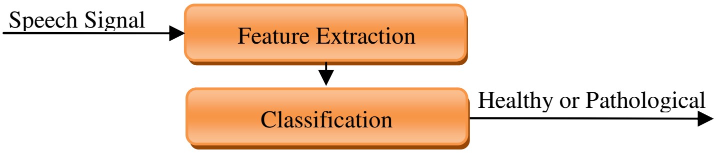

Fig. 1. The general scheme of vocal fold pathology diagnosis

The wavelet transform, as was shown in [5], is a flexible tool for time-frequency analysis of speech signals, especially for short data frames, like separate phonemes. In Fig.2 wavelet transform of a stressed vowel [a:], pronounced by a healthy speaker, is shown. But the situation changes in case of pathological voices. In Fig. 3, Fig. 4 and Fig. 5 wavelet transforms of the same vowel are given, but in these cases it is pronounced by speakers with different voice pathologies. The instability of the formant frequency is obviously seen.

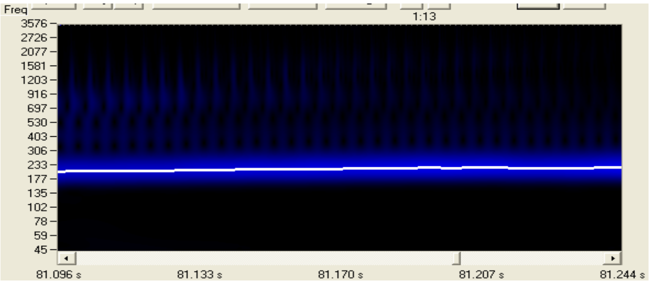

Fig. 2. Wavelet transform of a stressed vowel [a:] pronounced by a healthy speaker

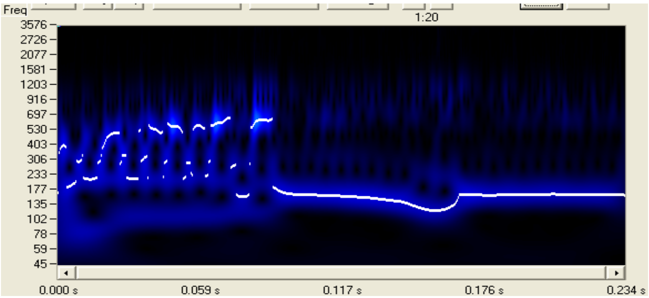

Fig. 3. Wavelet transform of a stressed vowel [a:] pronounced by the speaker with hypertrophic laryngitis

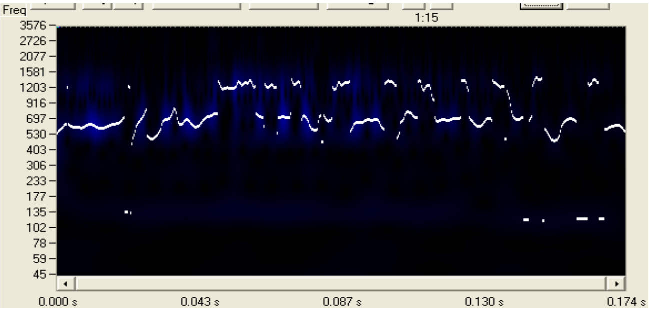

Fig. 4. Wavelet transform of a stressed vowel [a:] pronounced by the speaker with hypertonic dysphonia

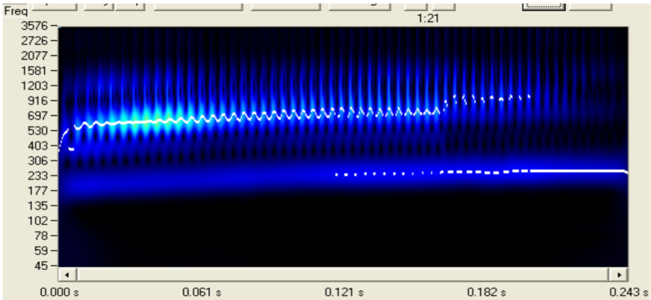

Fig. 5. Wavelet transform of a stressed vowel [a:] pronounced by the speaker with chronic catarrhal laryngitis

Different parameters for feature extraction are used. Traditionally, one deals with such parameters like pitch, jitter, shimmer, amplitude perturbation, pitch perturbation, signal to noise ratio, normalized noise energy [5] and others [6-9]. Feature extraction, using the above mentioned parameters, has shown its efficiency for a number of practical tasks. These parameters are frequently used in systems for automatic vocal fold pathology diagnosis, in speaker identification systems or in multimedia database indexing systems.

Finally, the extracted features are used for speech classification into the healthy and pathological class. Different machine learning methods such as Support Vector Machines [10], Artificial Neural Networks [11], etc. can be used as a classifier.

The presence of a pathology in a vocal tract inevitably leads to voice signal distortion. Depending on pathology severity the distortion may be more or less significant. Among all sounds that are produced by vocal tract, sustained vowels and some sonorant consonants are most easily distorted if a pathology is present.

2 Feature Extraction

In the first stage of the proposed method, as it is shown in Fig. 1, features extraction must be done. For this purpose, first, by the use of cepstral representation of the input signal, 13 Mel-Frequency-Cepstral-Coefficients (MFCC) are extracted. Then the wavelet packet decomposition in 5 levels is applied on the input signal to make the wavelet packet tree. Then, from the nodes of resulting wavelet packet tree, 63 energy features along with 63 Shannon entropy features are extracted. Finally, these three types of features and all possible states of their combination are considered and investigated to find the best state for constructing the proposed feature vector.

2.1 Mel-Frequency-Cepstral-Coefficients (MFCCs)

MFCCs are widely used features to characterize a voice signal and can be estimated by using a parametric approach derived from linear prediction coefficients (LPC), or by the non-parametric discrete fast Fourier transform (FFT), which typically encodes more information than the LPC method. The signal is windowed with a hamming window in the time domain and converted into the frequency domain by FFT, which gives the magnitude of the FFT. Then the FFT data is converted into filter bank outputs and the cosine transform is found to reduce dimensionality. The filter bank is constructed using 13 linearly-spaced filters (133.33Hz between center frequencies,) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency.) Each filter is constructed by combining the amplitude of FFT bin. The Matlab code to calculate the MFCC features was adapted from the Auditory Toolbox (Malcolm Slaney). The MFCCs are used as features in [12] to classify the speech into pathology and healthy class. Reduction of MFCC information has been used by averaging the sample's value of each coefficient.

2.2 Wavelet Packet Decomposition

Recently, wavelet packets (WPs) have been widely used by many researchers to analyze voice and speech signals. There are many out-standing properties of wavelet packets which encourage researchers to employ them in widespread fields. The most important, multi resolution property of WPs is helpful in voice signal synthesis [13-14].

The hierarchical WP transform uses a family of wavelet functions and their associated scaling functions to decompose the original signal into subsequent sub-bands. The decomposition process is recursively applied to both the low and high frequency sub-bands to generate the next level of the hierarchy. WPs can be described by the following collection of basic functions:

$$W_{2n}(2^{p-1}x-1)=\sqrt{2^{1-p}}\sum_{m}h(m-2l)\sqrt{2^{p}}W_{n}(2^{p}x-m) \tag{1}$$

$$W_{2n+1}(2^{p-1}x-1)=\sqrt{2^{1-p}}\sum_{m}g(m-2l)\sqrt{2^{p}}W_{n}(2^{p}x-m) \tag{2}$$

where *p* is scale index, *l* the translation index, *h* the low-pass filter and *g* the high-pass filter with

$$g(k) = (-1)^k h(1-k) \tag{3}$$

the WP coefficients at different scales and positions of a discrete signal can be computed as follows:

$$C_{n,k}^{p} = \sqrt{2^p} \sum_{m=-\infty}^{\infty} f(m) W_n (2^p m - k) \tag{4}$$

$$C_{2n,l}^{p-1} = \sum_{m} h(m-2l) C_{n,m}^{p} \tag{5}$$

$$C_{2n+1,l}^{p-1} = \sum_{m} g(m-2l) C_{n,m}^{p} \tag{6}$$

for a group of wavelet packet coefficients, energy feature in its corresponding sub-band is computed as

$$Energy_n = \frac{1}{N^2} \sum_{k=1}^{n} \left| C_{n,k}^{p} \right|^2 \tag{7}$$

The entropy evaluates the rate of information which is produced by the pathogens factors as a measure of abnormality in pathological speech. Also, the measure of Shannon entropy can be computed using the extracted wavelet-packet coefficients, through the following formula

$$Entropy_n = -\sum_{k=1}^{n} \left| C_{n,k}^{p} \right|^2 \log \left| C_{n,k}^{p} \right|^2 \tag{8}$$

In this study, mother wavelet function of the tenth order Daubechies has been chosen and the signals have been decomposed to five levels. The mother wavelet used in this study is reported to be effective in voice signal analysis [15-16] and is being widely used in many pathological voice analyses [14]. Due to the noise-like effect of irregularities in the vibration pattern of damaged vocal folds, the distribution manner of such variations within the whole frequency range of pathological speech signals is not clearly known. Therefore, it seems reasonable to use WP rather than DWT or CWT to have more detail sub-bands.

3 Classification by Support Vector Machine

For classification, a statistical learning algorithm called support vector machine (SVM) is used. SVMs which were proposed by Vapnik [17], have become an acknowledged classification method in the task of musical genre recognition. Their usage in this task was already justified by works of Li et al. [18] were SVMs outperformed other commonly used classification methods (Gaussian Mixture models, K-Nearest Neighbors classifier, Hidden Markov Models, etc.). We will consider the basic theory of SVM below.

Given a set of training vectors belonging to two separate classes, $(x_1,y_1),\ldots,(x_l,y_l)$, where $x_i \in R^N$ and $y_i \in \{-1,\ldots,1\}$, one wants to find a hyper-plane $wx + b = 0$ to separate the data. In fact, there are many possible hyper-planes, but there is only one that maximizes the margin (the distance between the hyper-plane and the nearest data point of each class). The solution to the optimization problem of SVM is given by the saddle point of the Lagrange functional

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \{ y_i[(w.x_i)+b]-1\} \tag{9}$$

where α_i are the Lagrange multipliers. Classical Lagrangian duality enables the primal problem (15) to be transformed to its dual problem, which is easier to solve. The solution is given by

$$\overline{w} = \sum_{i=1}^{l} \overline{\alpha}_i y_i x_i, \overline{b} = -\frac{1}{2}\overline{w}.[x_r + x_s] \tag{10}$$

where x_r and x_s are any two support vectors with $\overline{\alpha}_i, \overline{\alpha}_s > 0$, $y_r=1$, $y_s=-1$.

To solve the non-separable problem slack variables $\xi_i > 0$ and a penalty function, $f(\xi_i) = \sum_i \xi_i$, where the ξ_i are measures of the misclassification error. The solution is identical to the separable case except for a modification of the Lagrange multipliers as $0 \leq \alpha_i \leq C$, $i=1,..l$. The choice of C is not strict in practice.

The SVM can realize nonlinear discrimination by kernel mapping [17], when the samples in the input space cannot be separated by any linear hyper-plane, but can be linearly separated in the nonlinear mapped feature space. Of course in the proposed method, linear function is used as the kernel function of the SVM.

4 Experiments and Results

In this section, seven experiments have been designed. These experiments are simulated in Matlab 7.11.0. For displaying and comparing the results, four indicators (TP, FN, TN and FP) have been used.

True positive rate (TP), also called sensitivity, is the ratio between pathological files correctly classified and the total number of pathological voices. False negative rate (FN) is the ratio between pathological files wrongly classified and the total number of pathological files. True negative rate (TN), sometimes called specificity, is the ratio between normal files correctly classified and the total number of normal files. False positive rate (FP) is the ratio between normal files wrongly classified and the total number of normal files.

The final accuracy of the system is the ratio between all the hits obtained by the system and the total number of samples. Also for displaying the results, the ROC curve (AUC) has been used which is a very common way in medical decision systems.

4.1 Database Description

The database was created by specialists from the Belarusian Republican Center of Speech, Voice and Hearing Pathologies. 40 pathological speeches and 40 healthy speeches, which are related to sustained vowel “a”, have been selected randomly. All the records are in PCM format, 16 bits, mono, with 16 kHz sampling frequency. A random partition is also created for applying the holdout validation on the dataset. This partition divides the dataset into a training set with 40 records and a test (or holdout) set with 40 records.

4.2 Results

In the first experiment, only the MFCC is used to make the final feature vector. In the second experiment, only the energy resulting from the wavelet packet tree is used to make the final feature vector. In the third experiment, only the entropy resulting from the wavelet packet tree is used to make the final feature vector. In the fourth experiment, the MFCC along with the energy resulting from the wavelet packet tree are used to make the final feature vector. In the fifth experiment, the MFCC along with the entropy resulting from the wavelet packet tree are used to make the final feature vector. In the sixth experiment, the energy along with the entropy resulting from the wavelet packet tree are used to make the final feature vector. In the seventh experiment, the MFCC along with the energy and entropy resulting from the wavelet packet tree are used to make the final feature vector. Finally, in all experiments, for each speech signal the samples according its feature vector are extracted and fed to the SVM classifier. The classification results are shown in table 1. Also the comparison diagrams, including ROC curves and accuracy charts, are illustrated in Fig. 6 and Fig. 7.

Table 1. The results of experiments

Feature Extraction Method	**TP**	**TN**	**FP**	**FN**	**Accuracy**
MFCC	85%	100%	0%	15%	92.5%
Energy	85%	100%	0%	15%	92.5%
Entropy	85%	95%	5%	15%	90%
MFCC + Energy	95%	100%	0%	5%	97.5%
MFCC + Entropy	85%	100%	0%	15%	92.5%
Energy + Entropy	95%	90%	10%	5%	92.5%
MFCC + Energy + Entropy	90%	100%	0%	10%	95%

With considering the results, it is clear that the feature vector base on the combination of MFCC and energy of wavelet packet tree nodes has more potential and accuracy for using in the classification in comparison with other types. So, it is suggested to use the combination of MFCC and energy of wavelet packet tree nodes as the final feature vector.

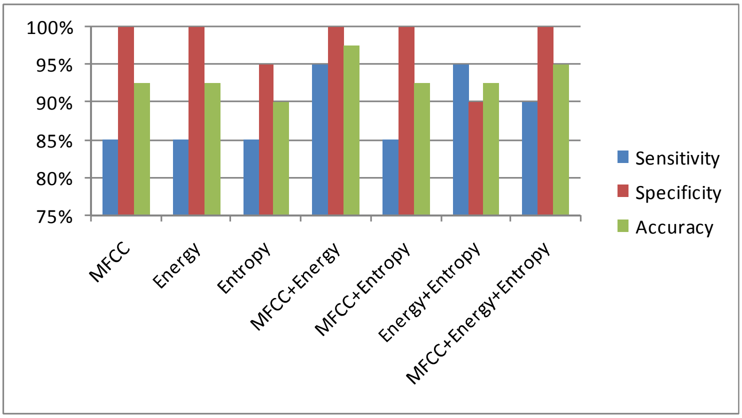

Fig. 6. The comparative results of different feature extraction methods

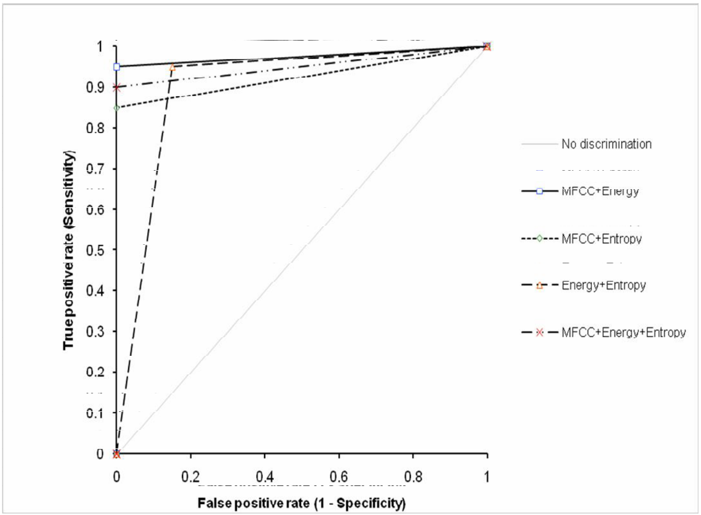

Fig. 7. The ROC curves of different feature extraction methods

4.3 Discussion

In pervious works, different parameters are used as the feature vector. For example, in [9] amplitude perturbation, pitch perturbation, etc. are used as the features and an average classification rate of 85.8% is reported. In [7] jitter, peak of autocorrelation, etc. are used as the features and an accuracy of 54.79% is reported for clustering of voice types.

But in this article, utilizing of the MFCC and the energy of the wavelet packet tree nodes as the features are proposed. The classification results in Fig. 6 show the accuracy of 97.5% and the sensitivity of 95% and the specificity of 100% for the proposed method.

Moreover, this fact can be seen in Fig. 7 that the AUC obtained with other types of features are decreased, while with the proposed feature vector (MFCC + Energy) the expected performance of system increases. This fact shows better performance of the proposed feature vector in comparison with the other types.

5 Conclusion

In this article, it is shown that features based on wavelet transformation have potential for detection of vocal fold pathology. So, three types of features which are energy and entropy resulting from the wavelet packet tree and the Mel-Frequency-Cepstral-Coefficients (MFCCs) and also their combination are investigated.

Also, a novel feature vector base on the combination of MFCC and energy of the wavelet packet tree nodes is proposed. By the means of SVM classifier, seven experiments are designed to investigate the efficiency of the proposed feature vector. The results of experiments show better performance of the proposed feature vector in comparison with other types.

Although it may be possible to try to build a complete multi-class classification system with a hierarchy of support vector machines so that detection of different type of pathological speech will be possible. For further research, it is suggested to work on the more sophisticated feature extraction phase.

Acknowledgments. This work was supported by the speech laboratory of NASB in Belarus. The authors wish to thank the Belarusian Republican Center of Speech, Voice and Hearing Pathologies by its support in the speech database.

References

1. Alonso, J.B., Leon, J.D., Alonso, I., Ferrer, M.A.: Automatic Detection of Pathologies in the Voice by HOS Based Parameters. EURASIP Journal on Applied Signal Processing 2001(4), 275–284 (2001)
2. Ceballos, L.G., Hansen, J., Kaiser, J.: A Non-Linear Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Trans. Biomedical Engineering 45(3), 300–313 (2005)
3. Ceballos, L.G., Hansen, J., Kaiser, J.: Vocal Fold Pathology Assessment Using AM Autocorrelation Analysis of the Teager Energy Operator. In: Proc. of the ICSLP 1996, pp. 757–760 (1996)
4. Adnene, C., Lamia, B.: Analysis of Pathological Voices by Speech Processing. In: 2003 Proc. of the Signal Processing and Its Applications, vol. 1(1), pp. 365–367 (2003)

5. Manfredi, C.: Adaptive Noise Energy Estimation in Pathological Speech Signals. IEEE Trans. Biomedical Engineering 47(11), 1538–1543 (2000)
6. Llorente, J.I.G., Vilda, P.G.: Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors. IEEE Trans. Biomedical Engineering 51(2), 380–384 (2004)
7. Rosa, M.D.O., Pereira, J.C., Grellet, M.: Adaptive Estimation of Residue Signal for Voice Pathology Diagnosis. IEEE Trans. Biomedical Engineering 47(1), 96–104 (2000)
8. Mallat, S.G.: A Theory for Multi-resolution Signal Decomposition: the Wavelet Representation. IEEE Trans. Pattern Analysis and Machine Intelligence 11(7), 674–693 (1989)
9. Wallen, E.J., Hansen, J.H.: A Screening Test for Speech Pathology Assessment Using Objective Quality Measures. In: Proc. of the ICSLP 1996, pp. 776–779 (1996)
10. Chen, W., Peng, C., Zhu, X., Wan, B., Wei, D.: SVM-based identification of pathological voices. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, pp. 3786–3789 (2007)
11. Ritchings, R.T., McGillion, M.A., Moore, C.J.: Pathological voice quality assessment using artificial neural networks. Medical Engineering & Physics 24(8), 561–564 (2002)
12. Lee, J.-Y., Jeong, S., Hahn, M.: Classification of pathological and normal voice based on linear Discriminant analysis. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4432, pp. 382–390. Springer, Heidelberg (2007)
13. Herisa, H.K., Aghazadeh, B.S., Bahrami, M.N.: Optimal feature selection for the assessment of vocal fold disorders. Computers in Biology and Medicine 39(10), 860–868 (2009)
14. Fonseca, E.S., Guido, R.C., Scalassarsa, P.R., Maciel, C.D., Pereira, J.C.: Wavelet time frequency analysis and least squares support vector machines for identification of voice disorders. Computers in Biology and Medicine 37(4), 571–578 (2007)
15. Guido, R.C., Pereira, J.C., Fonseca, E.S., Sanchez, F.L., Vierira, L.S.: Trying different wavelets on the search for voice disorders sorting. In: Proceedings of the 37th IEEE International Southeastern Symposium on System Theory, pp. 495–499 (2005)
16. Umapathy, K., Krishnan, S.: Feature analysis of pathological speech signals using local discriminant bases technique. Medical and Biological Engineering and Computing 43(4), 457–464 (2005)
17. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
18. Li, T., Oginara, M., Li, Q.: A comparative study on content based music genre classification. In: Proc. of the 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 282–289 (2003)