Scientia Iranica D (2013) 20(6), 2051–2060

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
www.scientiairanica.com

Remote diagnosis of unilateral vocal fold paralysis using matching pursuit based features extracted from telephony speech signal

Y. Shekofteh and F. Almasganj*

Department of Biomedical Engineering, Amirkabir University of Technology, Tehran, 424, Iran.

Received 29 July 2012; accepted 8 June 2013

KEYWORDS
Pathological speech signal;
Unilateral vocal fold paralysis;
Feature extraction;
Matching pursuit;
Support vector machine.

Abstract. Unilateral Vocal Fold Paralysis (UVFP) is a type of neurogenic laryngeal disorder in which vocal folds of patients do not have their normal behaviors, leading to abnormal talking voices. In this paper, a new noninvasive method for processing telephony speech signals is proposed to remotely diagnose the voice of the patients with UVFP disease. The proposed feature extraction method benefits from an adaptive decomposition method, the Matching Pursuit (MP) algorithm, to decompose the involved signals to some predefined atoms. Then, the attributes of the obtained atoms assigned to the speech signal converts to a final feature vector, so called MSDMP. Simulation results indicate the usefulness of the proposed feature vector with respect to a commonly used wavelet based features (EWPD). The MSDMP feature vector has improved the classification rate by 4.98%, as compared to the EWPD feature vector.

1. Introduction

Development of noninvasive methods in the diagnosis of different diseases can lead to spreading and improvement of prevention and care programs. The speech signal is an easily accessible signal that clearly represents the characteristics of larynx and vocal folds. From this point of view, the speech signal could be an efficient subject for noninvasive approaches to diagnose the patients who have larynx problems. By applying proper feature extraction and classification methods to the patients' speech signal, diagnosis of vocal fold diseases such as Unilateral Paralysis, edema, nodules and polyp could be realized. Development of the vocal folds screening systems is an interesting engineering field, as they do not need considerable hardware requirements, e.g. pronouncing some certain words by a suspected case is adequate to be fed to such a recognition system. In order to use such systems, patients must undergo a simple self-training process to properly pronounce some certain words. The processing results of the voice of suspected subjects (with pathological sounds), analyzed by a properly developed system, could be used as a prognosis of laryngeal diseases to assist the physicians.

Employing proper feature extraction methods is a key element in signal classification problems. In the field of this work, there are a large number of researches in which different types of features are extracted related to the operation of vocal folds. Acoustic parameters like shimmer (amplitude perturbation), jitter (pitch perturbation), ratio of the harmonics energy to the noise energy, Normalized First Harmonic Energy (NFHE) and Turbulent Noise Index (TNI) are some of the introduced primary features [1-4]. Besides, some techniques were introduced to estimate the pitch frequency of pathological voices, based on

*. *Corresponding author. Tel.: +98 21 64542372; Fax: +98 21 66495655*
E-mail addresses: y_shekofteh@aut.ac.ir (Y. Shekofteh), almas@aut.ac.ir (F. Almasganj)

autocorrelation function and cepstral coefficients [5]. Hansen et al. utilized the non-linear characterization of the speech production system via the differential "Teager energy" operator and the energy separation algorithms to capture irregularity characteristic of the vocal folds [6].

In the field of the spectral-based feature extraction approaches, Hartl et al. [7] designed a spectrum-based measurement method by elimination of inter-subject variability. They used features such as relative energy levels of the first harmonic, first formant and third formant, spectral slope in the low-frequency zone and relative level of energy above 6 kHz. Based on these features, they observed a relative increase of energy level in the mid-frequency and high-frequency ranges and a decrease in the low-frequency spectral slope of the voices of the patients with Unilateral Vocal Fold Paralysis (UVFP), in contrast to normal voices.

Gelzinis et al. [8] compared the effectiveness of different feature vectors to categorize an input voice signal into three classes (normal and two pathological classes, diffuse and nodular) or two classes (normal and pathological). The obtained results showed that the pitch and amplitude perturbation are two characteristics that are the best among the other ones. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) were the classifiers employed in their work.

Moreover, there are researchers that are searching for nonlinear dynamic characteristic of the voice signals [9-11]. Zhang benefited from the Correlation Dimension (CD) feature, to compare sustained voices generated by normal against patient subjects with UVFP defect [10]. In this field, Vaziri et al. [11] conducted some experiments on the use of phase space-based features such as CD [12], the Largest Lyapunov Exponent (LLE) [13], approximate entropy (ApEn) [14], Fractal Dimension (FD) [15] and Ziv–Lempel complexity (ZL) [16]. They compared these features with some perturbation-based features like jitter and shimmer.

By studying the local characteristic of the speech signal, Behroozmand and Almasganj introduced a modified version of wavelet-based feature extraction method [17]. In their work, a Wavelet Packet Decomposition (WPD) was first applied to the speech signal to divide it into many sub bands. The energy and entropy of each sub band was then evaluated to form the elements of an initial feature vector. Finally, via a genetic algorithm, a properly selected feature set was finalized out of the initial feature vector. In a similar way, Khadivi et al. designed a system to distinguish three of the most common types of vocal diseases; unilateral paralysis, polyp and nodules from each other [18]. They utilized WPD-based feature set and compared three different feature selection methods: Davies-Bouldin criteria, genetic algorithm and k-nearest neighbors in the considered task. Also, in [19,20], the ability of entropy and energy features, obtained from the coefficients of an optimum wavelet packet tree, was proposed along with Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) as feature dimension reduction methods and individual, forward, backward, and branch-and-bound methods were examined as feature selection methods.

In [21], a joint time-frequency approach was proposed using an adaptive time-frequency transform algorithm to decompose the speech signals. Then, several features such as octave max, octave mean, energy ratio, length ratio and frequency ratio were extracted from the decomposition parameters, and analyzed using statistical pattern classification techniques. In addition, Ghoraani and Krishnan proposed a methodology to extract the needed features using adaptive Time-Frequency Distribution (TFD) and Nonnegative Matrix Factorization (NMF) [22]. This led to meaningful features by which one could successfully measure the rate of the abnormality of the speech signal.

On the other hand, a system for remotely detecting vocal fold pathologies, using the telephony speech, was introduced by Moran et al. [23]. This system benefited from a linear classifier fed by measurements of pitch perturbation, amplitude perturbation and harmonic-to-noise ratio. They showed that the amplitude perturbation features are robust features for the telephony speech diagnosis. They subcategorized the pathologic recordings into four groups comprising of normal, neuromuscular pathologic, physical pathologic and mixed pathologic.

In this paper, we follow an approach that its overall framework and test set is nearly similar to [17], with two main differences: In this work, a different wavelet-based feature extraction approach is followed; second, the experiments are conducted over a known pathological speech corpus, while its speech samples are passed through a telephone line simulator. We follow a Matching Pursuit (MP)-based method to extract the useful features. Matching Pursuit algorithm is an adaptive signal decomposition method by which the time-frequency characteristics of a signal could be extracted, properly. This method provides the main basis vectors of an input signal, by assigning a number of atoms, with their corresponding attributes, to every signal. The attributes deviations of an assigned atoms group, for a suspected speech signal, perform high potential features for diagnosing the signal. In this work, the proposed feature vector is used to distinguish the voices of patients with Unilateral Vocal Fold Paralysis (UVFP) from normal voices.

The rest of this paper is organized as follows: In Section 2, UVFP disease is introduced in details. Section 3 describes the database used in this work. In

Section 4, the Matching Pursuit method is introduced. In the following section, our proposed feature vector (MSDMP) and a commonly used feature vector based on wavelet are presented. The classification procedure and the experimental results are provided in Sections 6 and 7, respectively. In the final sections, we have discussion and conclusion of this paper.

2. Unilateral vocal fold paralysis

Larynx is the source of the sound generated by the human vocal tract system. It produces periodic sounds through a rhythmic opening and closing of the vocal folds. The vocal folds, also known as vocal cords, are composed of twin infoldings of mucous membrane stretched horizontally across the larynx. One of the most common neurological diseases of the vocal folds is unilateral paralysis disease. A dysfunction in the vogues or recurrent nerves innervating the larynx could lead to a laryngeal disease, so-called Unilateral Vocal Fold Paralysis (UVFP) [8]. In details, UVFP commonly occurs for three major reasons: nerve injury during the surgeries, pressure on the nerve from an adjacent tumor and viral infection. Together, these three reasons account for more than 85% of the cases of paralyzed vocal folds. Of course, there are other less common causes such as stroke and other neurologic-based diseases [24]. In Figure 1, two video stroboscopic images are shown. They are the vocal folds of two different cases, a normal person and a patient case with UVFP disease. In both cases, the opening phase of vocal folds is shown during phonation of a vowel sound.

We could see that normal vocal folds in opening phase are completely isolated from each other. But, this is not the case for the ill vocal folds. In addition, vocal folds of a normal person, in the closing phase, are completely jointed; but the glottis fissure of the patients with either unilateral or bilateral paralysis of the vocal folds remains continuously open, which leads to the glottal air leakage. As a result, not only the incomplete closure of vocal folds, but also the incomplete opening of vocal folds results in airflow turbulence and chaotic behavior of the voiced portions of the speech signals of the patients [25]. Such voices are referred to as breathy or creaky. Because of the chaotic behavior of the UVFP patients' voices, caused by the airflow turbulence in glottis, this effect could be tracked by studying the form and harmonics of the related speech signals.

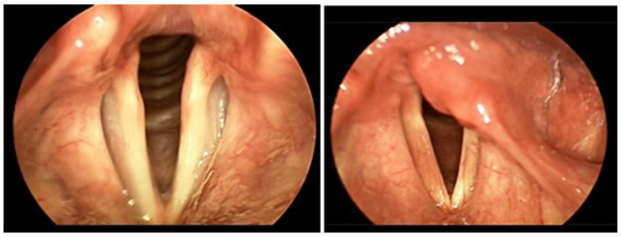

Figure 1. Video stroboscopic image of the vocal cords at the opening phase during phonation of a vowel sound: (a) Normal vocal folds; and (b) unilaterally paralyzed vocal folds [24].

3. Database

The acoustic samples used in this paper are selected from the disordered voice database of Kay Elemetrics Corporation [26]. This database is developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Lab. To conduct our experiments, we employ 53 normal acoustic samples and 67 unilateral vocal fold paralysis samples (34 women and 33 men) from this database. The selected samples are identical to ones used in [17]. The acoustic samples include vowel /a/ during long phonation of a word that contains vowel /a/. The voice signals are recorded on a DAT recorder at a sampling rate of 44.1 kHz with 16-bit resolution in a controlled condition.

In the recent years, a big focus has been made towards the design of remote systems [23]. This is due to the rapid technological improvements and the accessibility of communicational devices. As mentioned earlier, in this work, we are interested to apply and evaluate our approaches over the telephony speech signals; so, we have to convert the aforementioned data set to its telephony version, using the following block diagram depicted in Figure 2.

Toward this conversion, the sampling rate of the acoustic signals must be set to 8 kHz. Applying such distortion to speech signals is necessary to simulate the telephone transmission line. The final processed signal is approximately selected in the middle of each signal (phonation of vowel /a/) with the length of 350 milliseconds.

4. The overall structure of the proposed feature extraction approach

Matching Pursuit (MP) algorithm is introduced by Mallat and Zhang [27]. In this method, adaptive decomposition of a signal is done based on the locally

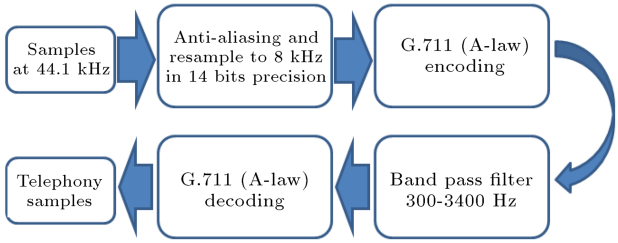

Figure 2. Block diagram of the voice conversion (44100 Hz to 8000Hz) method.

temporal characteristics of the signal; so, this method would be an appropriate time-frequency decomposition method applicable to non-stationary signals. The MP method uses a set of predefined basis functions, so-called dictionary; a signal could be optimally represented by a limited set of these functions [28]. Each of the basis functions is a so-called atom. A group of properly selected atoms, which constitutes the basis functions of a considered subspace of signals, creates a dictionary; by employing such a dictionary, a signal could be represented by a limited number of atoms. If a dictionary is properly designed, the MP method will provide a small set of basis functions that could reproduce a good approximate of the signal.

A variety of applications such as video coding and music note detection successfully exploited the MP method [29,30]. The MP method is also used for signal classification [31]. Umapathy et al. showed that the MP method could act as an adaptive time-frequency transform algorithm [32]. In a similar way, Chu et al. proposed an MP-based feature extraction method in the context of processing environmental sounds [33]. They followed the task of analyzing the environmental sound to understand the scene or context surrounding an audio sensor.

To understand the MP algorithm, suppose a dictionary, D, is a collection of some proper atoms [34], $h_\gamma(t)$ given by:

$$D = \{h_\gamma(t) : \gamma \in \Gamma\}, \tag{1}$$

where Γ is the index collection of the atoms belonging to the dictionary. The linear approximation of a signal, $x(t)$, using a limited number of available atoms, m, can be written as:

$$x(t) = \sum_{i=1}^{m} A_{\gamma_i} h_{\gamma_i}(t) + R^{(m)}(t), \tag{2}$$

where $R^{(m)}(t)$ is the residual signal (error) of the applied approximation. Notice that the residual signal is a signal with the same length as the original signal $x(t)$. In the decomposition process, the coefficient, A_{γ_i}, is the approximated weight of the atom, $h_{\gamma_i}(t)$, with the index of γ_i. In the MP method, for a desired number of atoms (for $i = 1, \cdots, m$), iteratively finding the best index, γ_i, and the corresponding weight, A_{γ_i}, is the main problem.

To implement the MP algorithm, first, the inner products of the signal with all atoms, included in the dictionary D, must be evaluated. The atom corresponding to the largest inner product is assigned as the winner atom, with the assigned index γ_1. The corresponding inner product magnitude is the approximated weight of the atom, A_{γ_1}, as given by:

$$A_{\gamma_1} = \langle x(t), h_{\gamma_1}(t) \rangle. \tag{3}$$

The first residual signal $R^{(1)}(t)$ is then defined by:

$$R^{(1)}(t) = x(t) - A_{\gamma_1} h_{\gamma_1}(t). \tag{4}$$

The first atom chosen by the MP algorithm has the best correlation with the signal structure. The subsequential atoms are then determined by a similar manner and iteratively applying the same procedure on the residual signal obtained in each iteration. By m times applying to the approach, the signal will be decomposed to m atoms with the evaluated weights and a final residual signal as $R^{(m)}(t)$.

4.1. Designing the dictionary

The performance of the MP method is highly related to its dictionary. Designing a proper dictionary makes the MP method a quick and efficient tool for signal decomposition purposes. Comparison among different sets of dictionaries such as Fourier, Haar and Gabor [28] shows that the set of Gabor functions is the best choice. Vetterli and Kovacevic showed that the Gabor-based time-frequency representation is optimal in minimizing the joint two-dimensional uncertainty in the combined spatial-frequency space [35]. Based upon these investigations, we decided to employ the Gabor dictionary in this research. Gabor functions are sine-modulated Gaussian functions determined by a set of four different parameters $\gamma = (s, u, w, \theta)$; they represent the scale, time-shift, frequency and phase of an atom, respectively. The Gabor atoms could simultaneously provide very good time and frequency localizations. The general definition of a discrete Gabor atom is given by:

$$g_{s,u,w,\theta}(n) = \frac{K_{s,u,w,\theta}}{\sqrt{s}} e^{-\pi^{(n-u)^2}/s^2} \times \cos[2\pi w(n-u) + \theta]. \tag{5}$$

In this work, to prepare the needed dictionary, the parameters are selected similar to [28] as follows:

- s: The time scale parameter, which corresponds to the atom width in time;

  $s = 2^p, \qquad (1 \le p \le 8),$

- u: The discrete time shift parameter;

  $u = 64 \times (j-1), \qquad (1 \le j \le 4),$

- w: The nonlinear normalized frequency parameter;

  $w = Ki^{2.6}, \qquad (1 \le i \le 35, \quad K = 0.5 \times 35^{-2.6}),$

- θ: The phase parameter: We used fixed-parameter;

  $\theta = 0.$

According to the selected parameters, the number of Gabor atoms accommodated in the constructed dictionary is equal to:

$$L = 8 \times 4 \times 35 \times 1.$$

The processed signal (with the length of 350 mSec) is segmented to 21 frames ($k = 21$) with half-frame shift. For each frame, the length of the window applied to signal (equal to the length of each atom) is selected as $N = 256$ samples, which covers a period of about 32 mSec, according to the selected telephony sampling rate. In Figures 3 and 4, a frame of normal signal and a frame of UVFP signal (disordered) are shown, respectively. In each of the figures, besides the original signals, the first three selected atoms for decomposition of the signals are plotted. In addition, the corresponding values of the scales (s), frequencies (w) and inner products (as its correlation) are shown in the figures.

Comparing Figures 3 and 4, the normal voice shows a nearly periodic pattern; whereas, the disordered signal (UVFP voice) shows a nearly irregular pattern. Based on the selected atoms for decomposition of the signals, we can see that the normal voice atoms have relatively lower frequencies (w) and higher scaling parameter (s) than those of the disordered signal. These differences could be efficiently employed as a feature vector to discriminate normal signals from disordered ones.

4.2. Computational cost of the MP method

For an input signal, it is first divided to k overlapped frames of the lengths of N sample. The MP algorithm

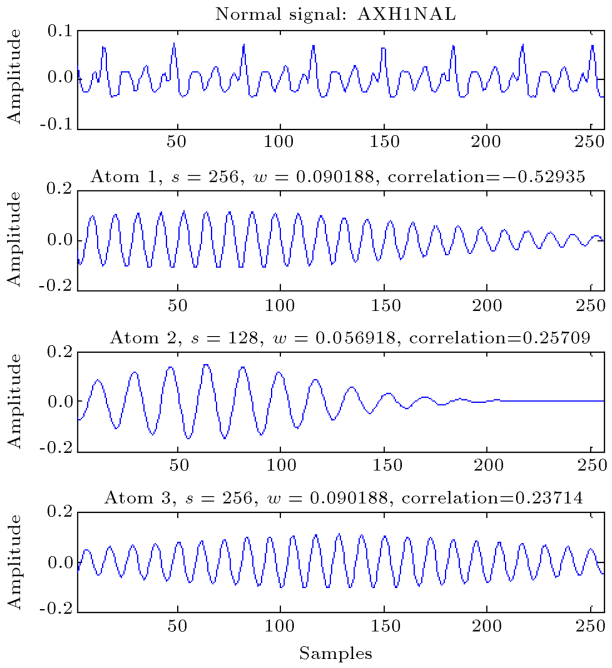

Figure 3. Speech waveforms of the vowel /a/ for a normal voice (top) and its first three main atoms.

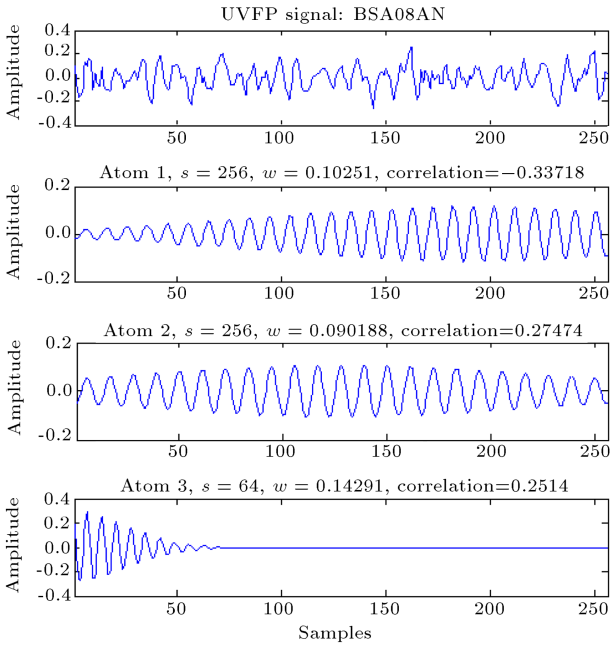

Figure 4. Speech waveforms of the vowel /a/ for a UVFP signal (top) and its first three main atoms.

is then applied individually to each of the frames. For each frame, the MP algorithm requires L inner products for all of the L atoms accommodated in the dictionary. So, the overall computational cost of the MP method would be $O(kmLN)$, where m is the number of atoms selected for the decomposition of each frame. During the decomposition process, the value of correlation and the selected atoms with their corresponding specifications such as scale (s) and frequency (w) parameters will be saved for the next processing to extract the proposed feature vector.

5. Feature extraction approaches

In this study, the performance of the proposed Feature Extraction (FE) method is compared to that of an older successfully used FE method [17]. In this section, we first introduce EWPD algorithm as the commonly-used FE method; we then compare the proposed MSDMP approach with the former method, by conducting some experiments arranged individually via these two FE methods.

5.1. EWPD feature extraction method

The Wavelet Transform (WT) is an effective tool for representing various behaviors of signals such as repeated patterns and discontinuities. This transform is especially a powerful tool for analyzing non-stationary signals. The wavelet-based decomposition approaches have been used extensively in the signal feature extraction tasks [17]. In the traditional Discrete Wavelet Transform (DWT), an input signal splits into two sub bands, detail and approximation; then in the

second level of decomposition, the approximation band splits into new detail and approximation bands, and for ith level of decomposition, the approximation band of $(i-1)$th level splits into new detail and approximation bands. An expansion of DWT is the Wavelet Packet Decomposition (WPD) that presents more details for a signal [36,37]. In each level of WPD, decomposition method would be applied to both approximation and detail sub bands.

To complete the feature extraction process, energy and entropy measures could be evaluated for the decomposed signals. The energy of each decomposed sub band is simply the sum of its squared coefficients. The main part of the energy of a voiced speech signal is mostly found in its approximation sub band [38,39]. On the other hand, entropy measure indicates the amount of the information stored in a signal. Different evaluations for entropy are introduced, such as Shannon, log energy, sure and threshold [40]. Behroozmand showed that in the task of pathological voice assessments, the Shanon entropy of WPD coefficients are more effective than the energy features [17]. So, in this study, Shannon entropy is employed to extract the feature vector fed to the classifier.

In this paper, WPD is utilized while using "Daubechies" mother wavelet and its decomposition over 3 levels. Then, Shannon entropy of each decomposed WPD subband is evaluated by:

$$E(i,j) = -\sum_{k=1}^{L(i,j)} G_{i,j}^2(k)\log(G_{i,j}^2(k)),$$

$$0 \le i \le 3, \qquad 0 \le j \le (2^i - 1), \tag{6}$$

where $G_{i,j}$ is the values of WPD coefficients at level i and subband j, and $L(i,j)$ is the number of WPD coefficients at level i and subband j. As mentioned before, in this work, three levels of WPD decomposition and entropy computation of each subband would be used of which the block diagrams are shown in Figure 5.

Adding the entropy of the original signal, $E(0,0)$, to the other 14 entropy coefficients, the final 15-dimensional feature vector would be obtained as represented by:

$$\text{EWPD}_{15\times 1} = [E(0,0), E(1,0), E(1,1), \cdots, E(3,7)]^T. \tag{7}$$

In this case, the whole processed signal with the length of 2688 time-samples (350 mSec) is utilized to compute EWPD feature vector.

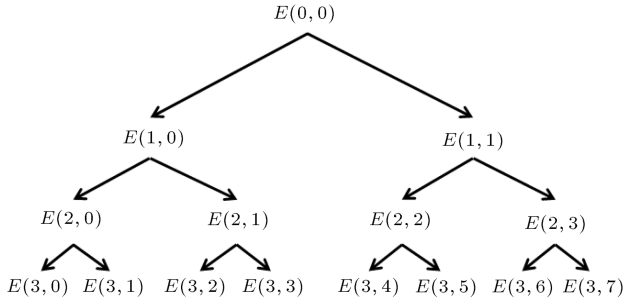

Figure 5. Entropy computation of the Wavelet Packet Decomposition (WPD) over 3 levels.

5.2. *Proposed MSDMP feature extraction method*

As mentioned previously, the proposed FE method is based on the Matching Pursuit method. In [28], MP-based features are utilized for the classification of environmental sounds. The main motivation toward using the MP-based features is based on this assumption that the most important information of a signal is located in its main atoms that are highly correlated to it. The MP method selects main atoms in a sequential order by eliminating the largest residual energy. So, the most useful atoms of the input signal can be obtained after a few iterations.

In order to extract the proposed MP-based features, each frame of pathological voice are decomposed by applying MP decomposition method, using the dictionary D with Gabor functions as its atoms. The first m atoms assigned to each frame are then selected, and their main parameters (nonlinear normalized frequency parameter, w, and time scale parameter, s) are gathered. Then, the Mean (M) and the Standard Deviation (SD) of these parameters (w and s) are calculated over all of the frames. In this manner, for a frame of signal, a 4-dimensional feature vector can be obtained as given by:

$$\text{MSDMP}_{4\times 1} = [\text{M}_w, \text{SD}_w, \text{M}_s, \text{SD}_s]^T,$$

$$\text{M}_w = 1/m\left(\sum_{i=1}^{m} w_i\right),$$

$$\text{SD}_w = \sqrt{1/(m-1)\sum_{i=1}^{m}(w_i - M_w)^2},$$

$$\text{M}_s = 1/m\left(\sum_{i=1}^{m} s_i\right),$$

$$\text{SD}_s = \sqrt{1/(m-1)\sum_{i=1}^{m}(s_i - \text{M}_s)^2}. \tag{8}$$

It should be mentioned that the extracted MSDMP features are highly robust against variations such as environmental noises. One reason may be due to the omission of atom orders, while evaluating the mean and standard deviation parameters.

6. Classification procedure

Supporting Vector Machine (SVM) is an important technique in the field of classification and pattern recognition. This method could be efficiently employed in different classification and regression tasks [41]. The SVM is basically a binary classifier that has a supervised learning phase in which the decision boundary is chosen in a manner to maximize the margine between data samples belonging to different classes. In this paper, in order to discriminate the pathological voices (UVFP) from the normal ones, the SVM classifier is employed. We select the "linear" function as the kernel of the SVM classifier, similar to the work introduced in [17]. The regularization constant of the SVM classifier is also set to 1.

To evaluate the performances of the implemented discrimination tasks, a 10-fold cross validation scheme is used. In this method, the samples of each class are randomly divided into 10 subsets. In each step of the experiment, a single subset is retained as the test set, and the 9 remained ones are used as the training data. This process will be repeated 10 times, for different subsets as the test data, and the results are averaged over the different folds.

In this paper, several measures are used such as sensitivity, specificity and accuracy for the evaluation of classification performance with different feature extraction methods [42]. These measures are determined as follows:

- Sensitivity = 100*TP/(TP + FN);
- Specificity = 100*TN/(TN + FP);
- Accuracy = 100*(TP + TN)/(TP + FP + TN + FN);

with the following definitions:

- True Positive (TP): The pathological cases which are classified as pathological samples;
- True Negative (TN): The normal cases which are classified as normal samples;
- False Positive (FP): The normal cases which are classified as pathological samples;
- False Negative (FN): The pathological cases which are classified as normal samples.

7. Experimental results

This section describes the experiments conducted to verify the effectiveness of the proposed feature extraction method for the Telephony-based diagnosis of Unilateral Vocal Fold Paralysis (UVFP) disease.

7.1. Designing the EWPD feature vector

In this study, the Daubechies (db) wavelet is selected to extract the feature vector for a baseline system. Order of the Daubechies wavelet is equal to wavelet vanishing moments. A high number of vanishing moments allows bettering the compressing regular parts of the pathological voice. On the other hand, the high orders of wavelet function are sensitive to noise, so, low-order invariants must be utilized. In this paper, to determine the best order of the Daubechies wavelet, some experiments are conducted using the measure of recognition accuracy. The accuracy results obtained by the EWPD feature vector are presented in Figure 6.

Based on the obtained results of Figures 6, db9 (Daubechie's wavelet function with order of 9) gives the best classification accuracy compared to all the other ones. Therefore, using this wavelet function in the baseline system, the traditional entropy based WPD feature vector (EWPD) results in the overall recognition accuracy of 86.07%.

7.2. Evaluation of MSDMP feature vector

In Figure 7, the recognition accuracy for the MSDMP features is plotted as a function of first main atoms (m) obtained from 10-fold cross validation. In this experiment, the recognition accuracy is calculated for various number of the first main atoms ($m = 1, 2, \cdots, 10$).

Results of Figure 7 indicate that using the first four atoms ($m = 4$) in the proposed MSDMP method, the highest classification performance is gained by 91.05%. According to Figure 7, using the value of $m > 2$, the classification performance of the MSDMP feature vector is greater than that of the EWPD feature vector. In addition, in Table 1, the classification performance of the EWPD and the MSDMP features are presented in terms of accuracy, sensitivity and specificity.

The obtained results of Table 1 demonstrate that the proposed MSDMP features performs more acceptably than the traditional EWPD feature extrac-

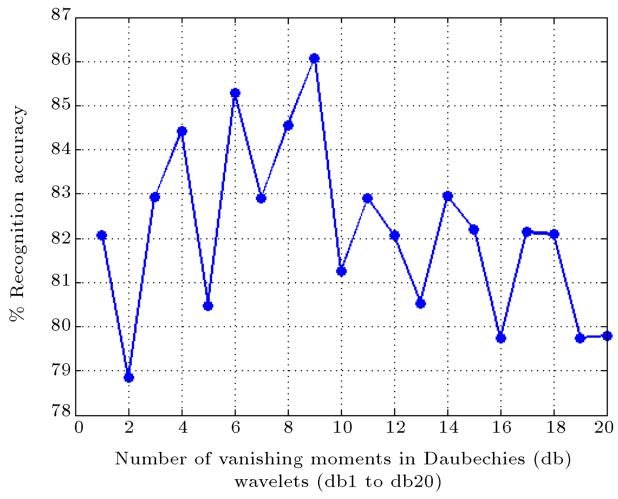

Figure 6. Recognition accuracy of the EWPD feature vector, using different Daubechie's wavelet functions with varying order (vanishing moments).

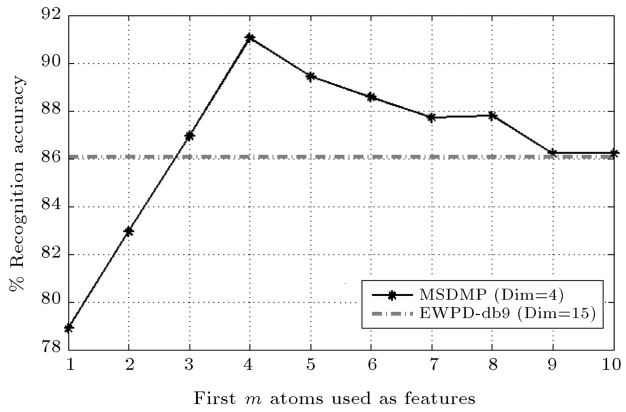

Figure 7. Recognition accuracy of the MSDMP feature vector, using various numbers of the first main atoms ($m = 1, \cdots, 10$).

Table 1. Comparing the performance of EWPD and MSDMP in terms of accuracy, sensitivity and specificity.

Feature extraction method	Accuracy	Sensitivity	Specificity
MSDMP ($m = 4$)	91.05%	91.08%	91.67%
EWPD (db9)	86.07%	89.15%	82.50%

tion method in terms of the accuracy, specificity and sensitivity.

7.3. Computational times

The computational times required to calculate the needed feature vectors, via different approaches, are shown in Table 2. In this table, the computational time of the proposed MSDMP method is presented for different values of m. All of the classification methods are implemented on a Pentium 4 processor with 2G RAM, and simulated using Matlab 7.11.0 (The Mathworks Web-Site [http://www.mathworks.com]).

As we can see in Table 2, implementing the proposed MSDMP feature extraction method, with different choices of the parameter m, needs lower computational time than the EWPD method. Also, the MSDMP method with 4-dimentional feature vector ($m = 4$) requires lower computational time and memory than the 15-dimensional EWPD feature vector in training and test stage of SVM classifier. Therefore, the proposed method is capable of analyzing pathological voice with faster computational time and less memory than the conventional method.

Table 2. Comparison of the computational times for different feature vectors.

Feature extraction method	Elapsed time (mSec)
EWPD	401.5
MSDMP ($m = 2$)	32.2
MSDMP ($m = 4$)	50.9
MSDMP ($m = 10$)	88.7

8. Discussion

Implementing the classification task, via the traditional WPD-based feature vector (EWPD), resulted in the overall recognition accuracy of 86.07%. For the MSDMP features, the recognition accuracy was evaluated for various values of m ($m = 1, 2, \cdots, 10$). It is observed that using the first four atoms ($m = 4$) of the proposed MSDMP method leads to the highest classification rate as 91.05%. Based on the results of Figure 7 which shows the recognition rates obtained via the MSDMP features as a function of m, the classification performance obtained via the proposed MSDMP features is greater than the EWPD method for $m > 2$. In addition, based on the results scheduled in Table 2, it is worthwhile to mention that the MSDMP method with $m = 4$ needs one-eighth of the computational time needed for the EWPD method. This also verifies the superiority of the MSDMP method over the EWPD method. The overall results of this paper show that the MSDMP features produce an effective way to extract discriminant features of the speech signals, using MP-based time-frequency localized representation.

9. Conclusion

In this paper, a feature extraction method is presented for remote diagnosis of UVFP disease from telephony-based pathological speech signals. This method is based on the Matching Pursuit (MP) algorithm. The MP method adaptively extracts the main atoms of a signal that include useful information. Therefore, the proposed feature extraction method (MSDMP) makes use of the time–frequency characteristic of the input signal. Simulation results show that the performance of the MP-based feature vector is better than that of the conventional WPD-based feature vector (EWPD) to get not only high classification metrics, but also in the low computational time and dimension of final feature vector.

Acknowledgments

The authors would like to thank Mr. Ahmadreza Rezaei for some comments that enhanced the quality of this paper.

References

1. Hadjitodorov, S. and Mitev, P. "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening", *Med. Eng. Phys.*, **24**(6) pp. 419-429 (2002).

2. Wallen, E.J. and Hansen, J.H.L. "A screening test for speech pathology assessment using objective quality measures", *Proceedings of the International Conference Spoken Language*, Philadelphia, USA, pp. 776-779 (1996).
3. Hadjitodorov, S., Boyanov, B. and Teston, B. "Laryngeal pathology detection by means of class-specific neural maps", *IEEE Trans Inf. Technol. Biomed.*, **4**(1), pp. 68-73 (2000).
4. Awan, S.N. and Frenkel, M.L. "Improvements in estimating the harmonics-to-noise ratio of the voice", *J. Voice*, **8**(3), pp. 255-262 (1994).
5. Mitev, P. and Hadjitodorov, S. "Fundamental frequency estimation of voice of patients with laryngeal disorders", *Inf. Sci.*, **156**(1-2), pp. 3-19 (2003).
6. Hansen, J.H.L., Gavidia-Ceballos, L. and Kaiser, J.F. "A non-linear operator based speech feature analysis method with application to vocal fold pathology assessment", *IEEE Trans. Biomed. Eng.*, **45**(3), pp. 300-313 (1998).
7. Hartl, D.M., Hans, S., Vaissiére, J., Riquet, M. and Brasnu, D.F. "Objective voice quality analysis before and after onset of unilateral vocal fold paralysis", *J. Voice*, **15**(3), pp. 351-361 (1999).
8. Gelzinisa, A., Verikasa, A. and Bacauskienea, M. "Automated speech analysis applied to laryngeal disease categorization", *Comput. Methods Programs Biomed.*, **91**(1), pp. 36-47 (2008).
9. Shekofteh, Y. and Almasganj, F. "Feature extraction based on speech attractors in the reconstructed phase space for automatic speech recognition systems", *ETRI J.*, **35**(1), pp. 100-108 (2013).
10. Zhang, Y., Jiang, J.J., Biazzo, L. and Jorgensen, M. "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis", *J. Voice*, **19**(4), pp. 519-528 (2005).
11. Vaziri, G., Almasganj, F. and Behroozmand, R. "Pathological assessment of patients' speech signals using nonlinear dynamical analysis", *Comput. Biol. Med.*, **40**(1), pp. 54-63 (2010).
12. Kantz, H. and Schreiber, T., *Nonlinear Time Series Analysis*, Cambridge University, Cambridge Press. (1997).
13. Petry, A. and Barone, D. "Preliminary experiments in speaker verification using time-dependent largest Lyapunov exponents", *Comput. Speech Lang.*, **17**(4), pp. 403-413 (2003).
14. Johnson, M.L., Straume, M. and Lampl, M. "The use of regularity as estimated by approximate entropy to distinguish saltatory growth", *Ann. Hum. Biol.*, **28**(5), pp. 491-504 (2001).
15. Ezeiza, A., Ipia, K.L., Hernndez, C. and Barroso, N. "Enhancing the feature extraction process for automatic speech recognition with fractal dimensions", *Cogn. Comput.* (2013). DOI: 10.1007/s12559-012-9165-0 (In Press).
16. Doganaksoy, A. and Gologlu, F. "On Lempel-Ziv complexity of sequences", *Lecture Notes in Computer Science*, **4086**, Springer Berlin/Heidelberg (2006).
17. Behroozmand, R. and Almasganj, F. "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis", *Comput. Biol. Med.*, **37**(4), pp. 474-485 (2007).
18. Khadivi-Herisa, H., SeyedAghazadeh, B. and Nikkhah-Bahrami, M. "Optimal feature selection for the assessment of vocal fold disorders", *Comput. Biol. Med.*, **39**(10), pp. 860-868 (2009).
19. Arjmandi, M.K. and Pooyan, M. "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine", *Biomed. Signal Process. Control*, **7**(1), pp. 3-19 (2012).
20. Arjmandi, M.K., Pooyan, M., Mikaili, M., Vali, M. and Moqarezadeh, A. "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods", *J. Voice*, **25**(6), pp. e275-e289 (2011).
21. Umapathy, K., Krishnan, S., Parsa, V. and Jamieson, D.G. "Discrimination of pathological voices using a time-frequency approach", *IEEE Trans. Biomed. Eng.*, **52**(3), pp. 421-430 (2005).
22. Ghoraani, B. and Krishnan, S. "A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification", *EURASIP. J. Adv. Signal. Process.*, Article ID 928974 (2009).
23. Moran, R.J., Reilly, R.B., Chazal, P. and Lacy, P.D. "Telephony-based voice pathology assessment using automated speech analysis", *IEEE Trans. Biomed. Eng.*, **53**(3), pp. 468-477 (2006).
24. Sulica, L. and Blitzer, A., *Vocal Fold Paralysis*, Heidelberg, Springer (2006). http://voicemedicine.com/unilateral.htm
25. Dekrom, G. "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments", *J. Speech Hear. Res.*, **38**(4), pp. 794-811 (1995).
26. DVD, Disordered Voice Database (CD-ROM), Version 1.03, Massachusetts Eye and Ear Infirmary, Kay Elemetrics Corporation, Boston, MA, Voice and Speech Lab. (1994).
27. Mallat, S. and Zhang, Z. "Matching pursuits with time-frequency dictionaries", *IEEE Trans. Signal Process.*, **41**(12), pp. 3397-3415 (1993).
28. Chu, S., Narayanan, S. and Kuo, J. "Environmental sound recognition with time-frequency audio features", *IEEE Trans. Audio, Speech Lang. Process*, **17**(6), pp. 1142-1158 (2009).
29. Neff, R. and Zakhor, A. "Very low bit rate video coding based on matching pursuits", *IEEE Trans. Circuits Syst. Video Technol.*, **7**(1), pp. 158-171 (1997).

30. Gribonval, R. and Bacry, E. "Harmonic decomposition of audio signals with matching pursuit", *IEEE Trans. Signal Process*, **51**(1), pp. 101-111 (2003).
31. Ebenezer, S.P., Papandreou-Suppappola, A. and Suppappola, S.B. "Classification of acoustic emissions using modified matching pursuit", *EURASIP J. App. Signal Process*, pp. 347-357 (2004).
32. Umapathy, K., Krishnan, S. and Jimaa, S. "Multigroup classification of audio signals using time-frequency parameters", *IEEE Trans. Multimedia*, **7**(2), pp. 308-315 (2005).
33. Chu, S., Narayanan, S. and Kuo, C.C.J. "Environmental sound recognition using mp-based features", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, pp. 1-4 (2008).
34. Chen, S.S., Donoho, D.L. and Saunders, M.A. "Atomic decomposition by basis pursuit", *SIAM J. Sci. Comput.*, **20**(1), pp. 33-61 (1998).
35. Vetterli, M. and Kovacevic, J., *Wavelets and Subband Coding*, Englewood Cliffs, NJ, Prentice-Hall (1995).
36. Ekici, S., Yildirim, S. and Poyraz, M. "Energy and entropy-based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition", *Expert Syst. Appl.*, **34**(4), pp. 2937-2944 (2008).
37. Learned, R.E. and Willsky, A.S. "A wavelet packet approach to transient signal classification", *Appl. Comput. Harmonic Anal.*, **2**(1), pp. 265-278 (1995).
38. Pham, T.V. and Kubin, G. "DWT-based classification of acoustic-phonetic classes and phonetic units", *Proceedings of IEEE International Conference on ICSLP*, South Korea, pp. 985-988 (2004).
39. Morchen, F. "Time series feature extraction for data mining using DWT and DFT", Technical Report, Department of Mathematics and Computer Science, University of Marburg, Germany, 33, pp. 1-31 (2003).
40. Coifman, R.R. and Wickerhauser, M.V. "Entropy-based algorithms for best basis selection", *IEEE Trans. Inform. Theory*, **38**(2), pp. 713-18 (1992).
41. Vapinik, V., *Statistical Learning Theory*, Wiley, New York (1998).
42. Duda, R.D., Hart, P.E. and Stroke, D.G., *Pattern Recognition*, 2nd Edn., John Wiley, New York (2001).

Biographies

Yasser Shekofteh received his BS in Biomedical Engineering and Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2005 and 2006, respectively. He received his MS in Biomedical Engineering from Amirkabir University of Technology in 2008. He is currently a PhD candidate in the Biomedical Engineering Department at Amirkabir University of Technology. His research interests include signal processing, speech recognition and keyword spotting.

Farshad Almasganj received his MS in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 1987 and his PhD in Biomedical Engineering from Tarbiat Modares University, Tehran, Iran, in 1998. He is currently an associate professor in the Biomedical Engineering Department of Amirkabir University of Technology. His research interests include automatic detection of voice disorders, speech recognition, prosody and language modeling for ASR systems.