Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 928974, 11 pages
doi:10.1155/2009/928974

Research Article

A Joint Time-Frequency and Matrix Decomposition Feature Extraction Methodology for Pathological Voice Classification

Behnaz Ghoraani and Sridhar Krishnan

Signal Analysis Research Lab, Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada M5B 2K3

Correspondence should be addressed to Sridhar Krishnan, krishnan@ee.ryerson.ca

Received 1 November 2008; Revised 28 April 2009; Accepted 21 July 2009

Recommended by Juan I. Godino-Llorente

The number of people affected by speech problems is increasing as the modern world places increasing demands on the human voice via mobile telephones, voice recognition software, and interpersonal verbal communications. In this paper, we propose a novel methodology for automatic pattern classification of pathological voices. The main contribution of this paper is extraction of meaningful and unique features using Adaptive time-frequency distribution (TFD) and nonnegative matrix factorization (NMF). We construct Adaptive TFD as an effective signal analysis domain to dynamically track the nonstationarity in the speech and utilize NMF as a matrix decomposition (MD) technique to quantify the constructed TFD. The proposed method extracts meaningful and unique features from the joint TFD of the speech, and automatically identifies and measures the abnormality of the signal. Depending on the abnormality measure of each signal, we classify the signal into normal or pathological. The proposed method is applied on the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database which consists of 161 pathological and 51 normal speakers, and an overall classification accuracy of 98.6% was achieved.

1. Introduction

Dysphonia or pathological voice refers to speech problems resulting from damage to or malformation of the speech organs. Dysphonia is more common in people who use their voice professionally, for example, teachers, lawyers, salespeople, actors, and singers [1, 2], and it dramatically effects these professional groups's lives both financially and psychosocially [2]. In the past 20 years, a significant attention has been paid to the science of voice pathology diagnostic and monitoring. The purpose of this work is to help patients with pathological problems for monitoring their progress over the course of voice therapy. Currently, patients are required to routinely visit a specialist to follow up their progress. Moreover, the traditional ways to diagnose voice pathology are subjective, and depending on the experience of the specialist, different evaluations can be resulted. Developing an automated technique saves time for both the patients and the specialist and can improve the accuracy of the assessments.

Our purpose of developing an automatic pathological voice classification is training a classification system which enables us to automatically categorize any input voice as either normal or pathological. The same as any other signal classification methods, before applying any classifier, we are required to reduce the dimension of the data by extracting some discriminative and representative features from the signal. Once the signal features are extracted, if the extracted features are well defined, even simple classification methods will be good enough for classification of the data. There have been some attempts in literature to extract the most proper features. Temporal features, such as, amplitude perturbation and pitch perturbation [3, 4] have been used for pathological speech classification; however, the temporal features alone are not enough for pathological voice analysis. Spectral and cepstral domains have also been used for pathological voice feature extraction; for example, mean fundamental frequency and standard deviation of the frequency [4], energy spectrum of a speech signal [5], mel-frequency cepstral coefficients (MFCCs) [6], and linear prediction cepstral

coefficients (LPCCs) [7] have been used as pathological voice features. Gelzinis et al. [8] and Sáenz-Lechón et al. [9] provide a comprehensive review of the current pathological feature extraction methods and their outcomes. We mention only few of the techniques which reported a high accuracy; for example, Parsa and Jamieson in [10] achieves 96.5% classification using four fundamental frequency dependent features and two independent features based on the linear prediction (LP) modeling of vowel samples. In [7], Godino-Llorente et al. feed MFCC coefficients of the vowel */ah/* from both normal and pathological speakers into a neural-network classifier, and achieve 96% classification rate. In [11], Umapathy et al. present a new feature extraction methodology. In this paper, the authors propose a segment free approach to extract features such as octave max and mean, energy ratio and length, and frequency ratio from the speech signals. This method was applied on continuous speech samples, and it resulted in 93.4% classification accuracy.

In this paper, we study feature extraction for pathological voice classification and propose a novel set of meaningful features which are interpretable in terms of spectral and temporal characteristics of the normal and pathological signals. In Section 2, we explain the proposed methodology. Section 3 provides an overview of the desired characteristics of the selected signal analysis domain and chooses a signal representation which satisfies the criteria. Section 4 describes nonnegative matrix factorization (NMF) as a part-based matrix decomposition (MD). In Section 5, we propose a novel temporal and spectral feature set and apply a simple classifier to train the pattern classifier. Results are given in Section 6, and conclusion is described in Section 7.

2. Methodology

In this paper, we propose a novel approach for automatic pathological voice feature extraction and classification. The majority of the current methods apply a short time spectrum analysis to the signal frames, and extract the spectral and temporal features from each frame. In other words, these methods assume the stationarity of the pathological speech over 10–30 milliseconds intervals and represent each frame with one feature vector; however, to our knowledge, the stationarity of the pathological speech over 10–30 milliseconds has not been confirmed yet, and as a matter of fact, our observation from the TFD of abnormal speech evident that there are more transients in the abnormal signals, and the formants in pathological speech are more spread and are less structured. Another shortcoming of the current approaches is that they require to segment the signal into short intervals. Using an appropriate signal segmentation has always been a controversial topic in windowed TF approaches. Since the real world signals have nonstationary dynamics, segmentation at nonstationarity parts of the signal could loose the useful information of the signal. To overcome these limitations, we propose a novel approach to extract the TF features from the speech in a way that it captures the dynamic changes of the pathological speech.

Figure 1 is a schematic of the proposed pathological speech classification approach. As shown in this figure, a joint TF representation of the pathological and normal signals is estimated. It has been shown that TF analysis is effective for revealing non-stationary aspects of signals such as trends, discontinuities, and repeated patterns where other signal processing approaches fail or are not as effective. However, most of the TF analyses have been utilized for visualization purpose, and quantification and parametrization of TFD for feature extraction and automatic classification have not been explicitly studied so far. In this paper, we explore TF feature extraction for pathological signal classification. As we mention in Section 3, not every TF signal analysis is suitable for our purpose. In Section 3, we explain the criteria for a suitable TFD and propose Adaptive TFD as a method which successfully captures the temporal and spectral localization of the signals components.

Once the signal is transformed to the TF plane, we interpret the TFD as a matrix $V_{M\times N}$ and apply a matrix decomposition (MD) technique to the TF matrix as given below

$$V_{M\times N} = W_{M\times r}H_{r\times N} = \sum_{i=1}^{r} w_i h_i \tag{1}$$

where N is the length of the signal, M is the frequency resolution of the constructed TFD, and r is the order of MD. Applying an MD on the TF matrix V, we derive the TF matrices W and H, which are defined as follows:

$$W_{M\times r} = [w_1 w_2 \cdots w_r],$$
$$H_{r\times N} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_r \end{bmatrix}. \tag{2}$$

In (1), MD reduces the TF matrix (V) to the base and coefficient vectors ($\{w_i\}_{i=1,\ldots,r}$ and $\{h_i\}_{i=1,\ldots,r}$, resp.) in a way that the former represents the bases components in the TF signal structure, and the latter indicates the location of the corresponding base vectors in time. The estimated base and coefficient vectors are used in Section 5 to extract novel joint time and frequency features. Despite the window-based feature extraction approaches, the proposed method does not take any assumption about the stationarity of the signal, and MD automatically decides at which interval the signal is stationarity. In this paper, we choose nonnegative matrix factorization (NMF) as the MD technique. NMF and the optimization method are explained in Section 4.

Finally, the extracted features are used to train a classifier. The classification and the evaluation are explained in Section 5.3.

3. Signal Representation Domain

The TFD, $V(t, f)$, that could extract meaningful features should preserve joint temporal and spectral localization of

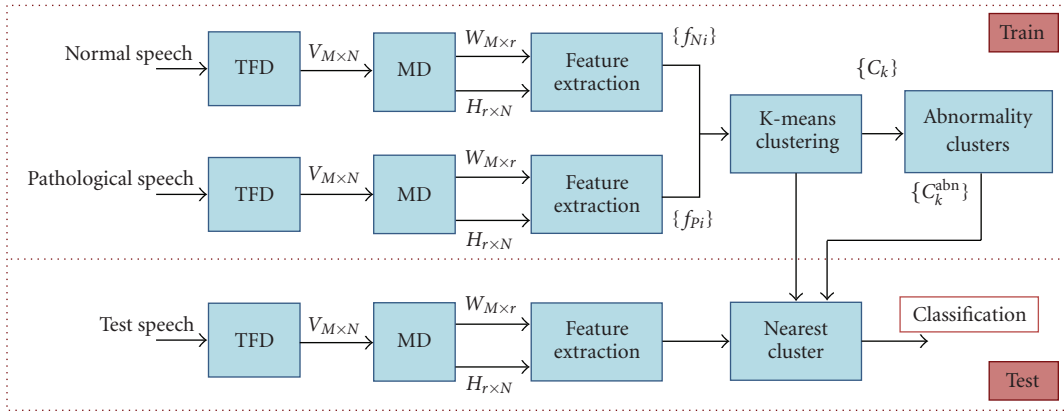


Figure 1: The schematic of the proposed pathological feature extraction and classification methodology.

the signal. As shown in [12], the TFD that preserves the time and frequency localized components has the following properties:

(1) There are nonnegative values.

$$V(t,f) \geq 0. \tag{3}$$

In order to produce meaningful features, the value of the TFD should be positive at each point; otherwise the extracted features may not be interpretable, for example, Wigner-Ville distribution (WVD) always gives the derivative of the phase for the instantaneous frequency which is always positive, but it also gives that the expectation value of the square of the frequency, for a fixed time, can become negative which does not make sense [13]. Moreover, it is very difficult to explain negative probabilities.

(2) There are correct time and frequency marginals.

$$\int_{-\infty}^{+\infty} V(t,f)\,df = |x(t)|^2, \tag{4}$$

$$\int_{-\infty}^{+\infty} V(t,f)\,dt = |X(f)|^2, \tag{5}$$

where $V(t,f)$ is the TFD of signal $x(t)$ with Fourier transform of $X(f)$. The TFD which satisfies the above criteria is called positive TFD [13]. A positive TFD with correct marginals estimates a cross-term free distribution of the true joint TF distribution of the signal. Such a TFD provides a high TF localization of the signal energy, and it is therefore a suitable TF representations for feature extraction from non-stationary signals. In this study, we use a TFD that satisfies the criteria in (5) and (3). This TFD is called Adaptive TFD as it is constructed according to the properties of the signal being analyzed. Adaptive TFD has been used for instantaneous feature extraction from Vibroarthrographic (VAG) signals in knee joint problems to classify the pathological conditions of the articular cartilage [14].

3.1. Adaptive TFD. Adaptive TFD method [14] uses the matching pursuit TFD (MP-TFD) as an initial TFD estimate to construct a positive, high resolution, and cross-term free TFD. As explained in Appendix A, MP-TFD decomposes the signal into Gabor atoms with a wide variety of modulated frequency and phase, time shift and duration, and adds up the Wigner distribution of each component. MP-TFD eliminates the cross-term problem with bilinear TFDs and provides a better representation for multicomponent signals. However, the shortcoming of MP-TFD is that it does not necessarily satisfy the marginal properties.

As described by Krishnan et al. [14], we apply a cross-entropy minimization to the matching pursuit TFD (MP-TFD) denoted by $\hat{V}(t,f)$, as a prior estimate of the true TFD, and construct an optimal estimate of TFD, denoted by $V(t,f)$ in a way that the estimated TFD satisfies the time and frequency marginals, $m_0(t)$ and $m_0(f)$, respectively.

The Adaptive TFD is iteratively estimated from the MP-TFD as given below.

(1) The time marginal is satisfied by multiplying and then dividing the TFD by the desired and the current time marginal:

$$V^{(0)}(t,f) = \hat{V}(t,f)\frac{m_0(t)}{\hat{p}(t)}, \tag{6}$$

where $\hat{p}(t)$ is the time marginal of $\hat{V}(t,f)$. At this stage, $V^{(0)}(t,f)$ has the correct time marginal.

(2) The frequency marginal is satisfied by multiplying and then dividing the TFD by the desired and the current frequency marginal:

$$V^{(1)}(t,f) = V^{(0)}(t,f)\frac{m_0(f)}{p^{(0)}(f)}, \tag{7}$$

where $p^{(0)}(f)$ is the frequency marginal of $V^{(0)}(t,f)$. At this stage $V^{(1)}(t,f)$ satisfies the frequency marginal condition, but the time marginal could be disrupted.

(3) It is shown that repeating the above steps makes the estimated TFD closer to the true TF representation of the signal.

4. Matrix Decomposition

We consider the TFD, $V(t, f)$, as a matrix, $V_{M\times N}$, where N is the number of samples, and M is the frequency resolution of the constructed TFD, for example, given an 81.92 ms frame with sampling frequency of 25 kHz, N is 2048 and the highest possible frequency resolution, M, is 1024, which is half of the frame length. Next, we apply an MD technique to decompose the TF matrix to the components, $W_{M\times r}$ and $H_{r\times N}$, in a way that $V \approx WH$. W and H matrices are called basis and encoding, matrices respectively, and $r < N$ is the number of the decomposition.

Depending on the utilized matrix decomposition technique, the estimated components satisfy different criteria and offer variant properties. The MD techniques that is suitable for TF quantification has to estimate the encoding and base components with a high TF localization. Three well-known MD techniques are Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Nonnegative Matrix Factorization (NMF). PCA finds a set of orthogonal components that minimizes the mean squared error of the reconstructed data. The PCA algorithm decomposes the data into a set of eigenvectors W corresponding to the first r largest eigenvalues of the covariance matrix of the data, and H, the projection of the data on this space. ICA is a statistical technique for decomposing a complex dataset into components that are as independent as possible. If r independent components $w_1 \cdots w_r$ compose r linear mixtures $v_1 \cdots v_n$ as $V = WH$, the goal of ICA is estimating H, while our observation is only the random matrix V. Once the matrix H is estimated, the independent components can be obtained as $W = VH^{-1}$. NMF technique is applied to a nonnegative matrix and constraints the matrix factors W and H to be nonnegative. In a previous study [15], we demonstrated that NMF decomposed factors promise a higher TF representation and localization compared to ICA and PCA factors. In addition, as it was mentioned in Section 3, the negative TF distributions do not result in interpretable features, and they are not suitable for feature extraction. Therefore, in this paper, we use NMF for TF matrix decomposition.

NMF algorithm starts with an initial estimate for W and H and performs an iterative optimization to minimize a given cost function. In [16], Lee and Seung introduce two updating algorithms using the least square error and the Kullback-Leibler (KL) divergence as the cost functions:

Least square error

$$W \longleftarrow W \cdot \frac{VH^T}{WHH^T}, \quad H \longleftarrow H \cdot \frac{W^TV}{W^TWH},$$

KL divergence

$$W \longleftarrow W \cdot \frac{(V/WH)H^T}{1 \cdot H}, \quad H \longleftarrow H \cdot \frac{W^T(V/WH)}{W \cdot 1}. \tag{8}$$

In these equations, $A \cdot B$ and A/B are term by term multiplication and division of the matrices A and B.

Various alternative minimization strategies have been proposed [17]. In this work, we use a projected gradient bound-constrained optimization method which is proposed by Lin [18]. The optimization method is performed on function $f = V - WH$ and is consisted of three steps.

(1) Updating the Matrix. W In this stage, the optimization of $f_H(W)$ is solved with respect to W, where $f_H(W)$ is the function $f = V - WH$, in which matrix H is assumed to be constant. In every iteration, matrix W is updated as

$$W^{t+1} = \max\{(W^t - \alpha^t \nabla f_H(W^t)), 0\}, \tag{9}$$

where t is the iteration order, $\nabla f_H(W)$ is the projected gradient of the function f, while H is constant, and α^t is the step size to update the matrix. The step size is found as $\alpha^t = \beta^{K_t}$. Where $\beta^1, \beta^2, \beta^3, \ldots$ are the possible step sizes, and K_t is the first nonnegative integer for which

$$f(W^{t+1}) - f(W^t) \le \sigma\langle \nabla f_H(W^t), W^{t+1} - W^t\rangle, \tag{10}$$

where the operator $\langle \cdot, \cdot \rangle$ is the inner product between two matrices as defined

$$\langle A, B\rangle = \sum_i \sum_j a_{ij} b_{ij}. \tag{11}$$

In [18], values of σ and β are suggested to be 0.01 and 0.1, respectively. Once the step size, α^t, is found, the stationarity condition of function $f_H(W)$ at the updated matrix is checked as

$$\left|\left|\nabla^P f_H(W^{t+1})\right|\right| \le \epsilon ||\nabla f_H(W^1)||, \tag{12}$$

where $\|\nabla f_H(W^1)\|$ is the the projected gradient of the function $f_H(W)$ at first iteration ($t = 1$), ϵ is a very small tolerance, and $\nabla^P f_H(W)$ is the projected gradient defined as

$$\nabla^P f_H(W) = \begin{cases} \nabla f_H(W), & w_{mr} > 0, \\ \min(0, \nabla f_H(W)), & w_{mr} = 0. \end{cases} \tag{13}$$

If the stationary condition is met, the procedure stops, if not, the optimization is repeated until the point W^{t+1} becomes a stationary point of f_H.

(2) Updating the Matrix. H: This stage solves the optimization problem respect to H assuming W is constant. A similar procedure to what we did in stage 1 is repeated in here. The only difference is that in the previous stage, H is constant, but here W is constant.

(3) The Convergence Test. Once the above sub-optimum problems are solved, we check for the stationarity of the W and H solutions together:

$$\begin{aligned} &||\nabla f_H(W^t)|| + ||\nabla f_W(H^t)|| \\ &\quad \le \epsilon(||\nabla f_H(W^1)|| + ||\nabla f_W(H^1)||). \end{aligned} \tag{14}$$

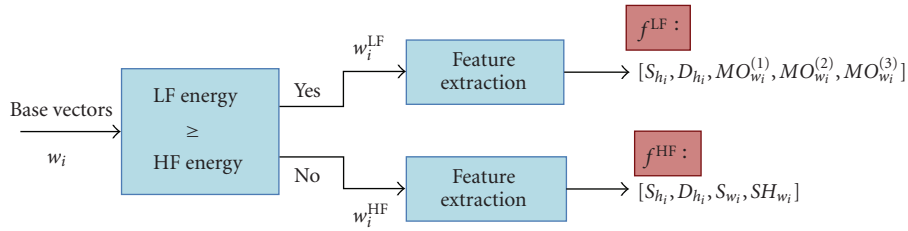


Figure 2: Block diagram of the proposed feature extraction technique.

The optimization is complete if the global convergence rule (14) is satisfied; otherwise, the steps 1 and 2 are iteratively repeated until the optimization is complete.

The gradient-based NMF is computationally competitive and offers better convergence properties than the standard approach, and it is, therefore, used in the present study.

5. Feature Extraction and Classification

In this section, we extract a novel feature set from the decomposed TF base and coefficient vectors (W and H). Our observations evident that the abnormal speech behaves differently for voiced (vowel) and unvoiced (constant) components. Therefore, prior to feature extraction, we divide the base vectors into two groups: (a) *Low Frequency* (LF): the bases with dominant energy in the frequencies lower than 4 kHz, and (b) *High Frequency* (HF): the bases with major energy concentration in the higher frequencies.

Next, as depicted in Figure 2, we extract four features from each LF base and five features from each HF base while only two of these two feature sets are the same. In order to derive the discriminative features of normal and abnormal signals, we investigate the TFD difference of the two groups. To do so, we choose one normal and one pathological speech and construct the Adaptive TFD of each 80 ms frame of the signals. The sum of the TF matrices for each speech is shown in Figure 3. We observed two major differences between the pathological and the normal speech: (1) the pathological signal has more transient components compared to the normal signal, and (2) the pathological voice presents weaker formants compared to the normal signal.

Base on the above observations, we extract the following features from the coefficient and base vectors.

5.1. Coefficient Vectors. It is observed that the pathological voice can be characterized by its noisy structure. The more transients and discontinuities are present in the signal, the more abnormality is observed in the speech. Two features are proposed to represent this characteristic of the pathological speech.

5.1.1. Sparsity. Sparsity of the coefficient vector distinguishes the nonfrequent transient components of the abnormal signals from the natural frequent components. Several sparseness measures have been proposed in the literature. In this paper, we use the function defined as

$$S_{h_i} = \frac{\sqrt{N} - \left(\sum_{n=1}^{N} h_i(n)\right) / \sqrt{\sum_{n=1}^{N} h_i^2}}{\sqrt{N} - 1}. \tag{15}$$

The above function is unity if and only if h_i contains a single nonzero component and is zero if and only if all the components are equal. The sparsity measure in (15) has been used for applications such as NMF matrix decomposition with more part-based properties [19]; however, it has never been used for feature extraction application.

The next proposed feature differentiates the discontinuity characteristics of the pathological speech from the normal signal.

5.1.2. Sum of Derivative. We have

$$D_{h_i} = \sum_{n=1}^{N-1} h_i'(n)^2, \tag{16}$$

where

$$h_i'(n) = h_i(n+1) - h_i(n), \quad n = 1, \ldots, N-1. \tag{17}$$

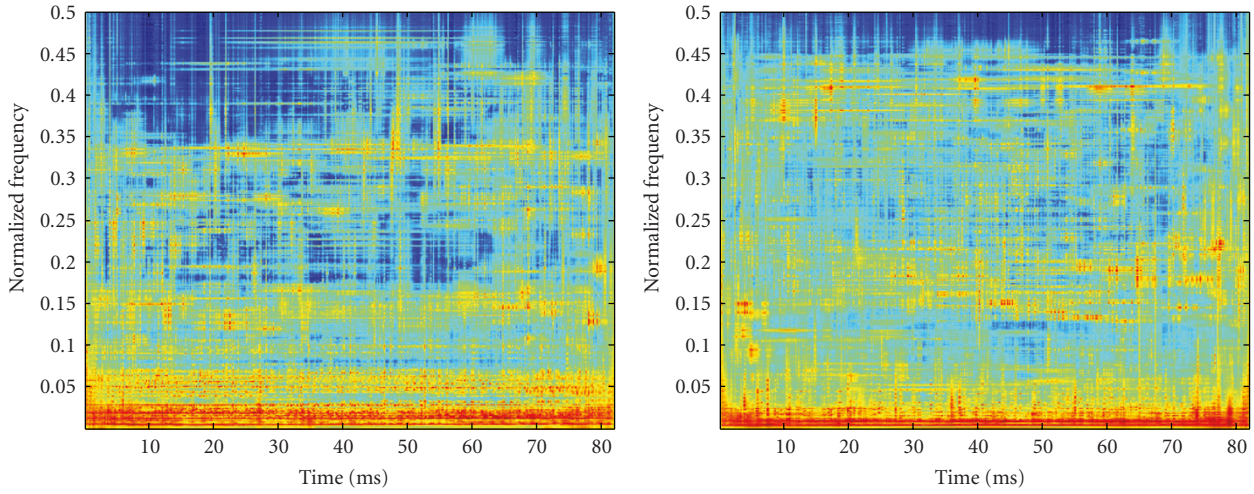
D_{h_i} captures the discontinuities and abrupt changes, which are typical in pathological voice samples.

5.2. Base Vectors. The base vectors represent the frequency components present in the signal. The dynamics of the voice abnormality varies between HF and LF-bases groups. Hence, we extracted different frequency features for each group.

5.2.1. Moments. Our observation showed that in the pathological speech, the HF bases tend to have bases with energy concentration at higher frequencies compared to the normal signals. To discriminate this abnormality property, we extract the first three moments of the base vectors as the features:

$$MO^{(o)} w_i = \sum_{m=1}^{M} f^o w_i(m), \quad o = 1, 2, 3 \tag{18}$$

where MO^1, MO^2, and MO^3 are the three moments, and M is the frequency resolution. The moment features are extracted from HF bases; the higher are the frequency energies, the larger will the feature values be. Although these features are useful for distinguishing the abnormalities of



(a) TF distribution of a normal voice with a male speaker

(b) TF distribution of a pathological voice with a male speaker

Figure 3: TFD of a normal (a) and an abnormal signal (b) is constructed using adaptive TFD with Gabor atoms, 100 MP iterations and 5 MCE iterations. As evident in theses figures, the pathological signal has more transient components specially at high frequencies. In addition, the TF of the pathological signal presents weak formants, while the normal signal has more periodicity in low frequencies, and introduces stronger formants.

the HF components, there are not useful for representing the abnormalities of the LF bases. The reason is that the major frequency changes in the LF components is dominated by the difference in pitch frequency of speech from one speaker to another speaker, and it does not provide any discrimination between normality or abnormality of the speech. Two features are proposed for LF bases.

5.2.2. Sparsity. As is known in literature, it is expected to observe periodic structures in the low frequency components of the normal speech. Therefore, when a large amount of scattered energy is observed in the low frequency components, we conclude that a level of abnormality is present in the signal. To measure this property, we propose the sparsity of the base vectors $\{w_i\}_{i=1,\ldots,M}$ as given below:

$$S_{w_i} = \frac{\sqrt{M} - \left(\sum_{m=1}^{M} w_i(m)\right)/\sqrt{\sum_{m=1}^{M} w_i^2}}{\sqrt{M} - 1}. \tag{19}$$

For normal signals we expect to have higher sparsity features, while pathological speech signals have lower sparsity values.

5.2.3. Sharpness. S_{w_i} measures the spread of the components in low frequencies. In addition, we need another feature to provide an information on the energy distribution in frequency. Comparing the LF bases of the normal and the pathological signals, we notice that normal signals have strong formants; however, the pathological signals have weak and less structured formants.

For each base vector, first we calculate the Fourier transform as given

$$W_i(\nu) = \left| \sum_{f=1}^{M} e^{-j(2\pi m\nu/M)} w_i(m) \right|. \tag{20}$$

where M is length of the base vector, and $W_i(\nu)$ is the Fourier transform of the base vector w_i. Next, we perform a second Fourier transform on the base vector, and obtain $W_i(\kappa)$ as follows:

$$W_i(\kappa) = \left| \sum_{\nu=1}^{M/2} e^{-j(2\pi\nu\kappa/(M/2))} W_i(\nu) \right|. \tag{21}$$

Finally, we sum up all the values of $|W(\kappa)|$ for κ more than m_0, where m_0 is a small number:

$$SH_{w_i} = \sum_{\kappa=m_0}^{M/4} |W_i(\kappa)|. \tag{22}$$

In Appendix B, we demonstrate that SH_{w_i} is a large value for bases representing strong formants, such as in normal speech, but is a small value for distorted formants, such as in pathological speech.

5.3. Classification. As it is shown in Figure 1, once the features are extracted, we feed them into a pattern classifier, which consists of a training and a testing stage.

5.3.1. Training Stage. Various classifiers were used for pathological voice classification [8], such as, the linear discriminant analysis, hidden Markov models, and neural networks. In the proposed technique, we use K-means clustering as a simple classifier.

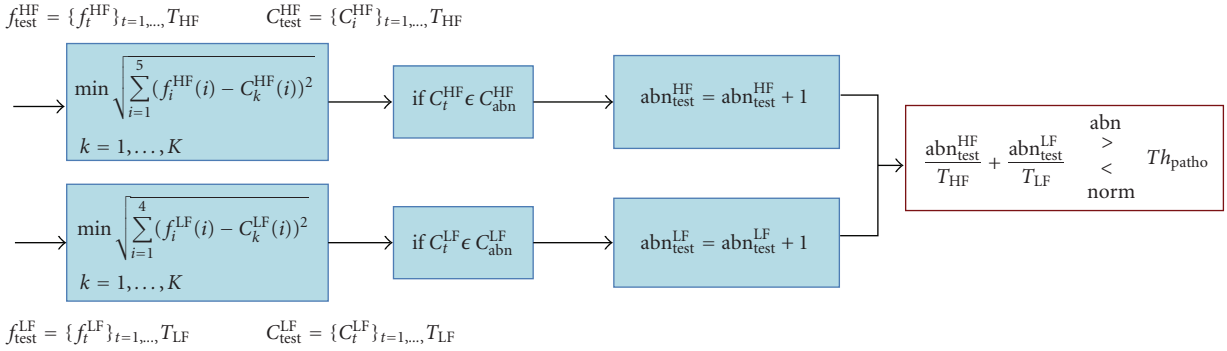


Figure 4: The block diagram of the test stage.

K-means clustering is one of the simplest unsupervised learning algorithms. The method starts with an initial random centroids, and it iteratively classifies a given data set into a certain number of clusters (K) by minimizing the squared Euclidean distance of the samples in each cluster to the centroid of that cluster. For each cluster, the centroid is the mean of the points in that cluster C_i.

Since separate features are extracted for LF and HF components, we have to train a separate classifier for each group: C^{LF} and C^{HF} for LF and HF components, respectively. Once the clusters are estimated, we count the number of abnormality feature vectors in each cluster, and the cluster with a majority of abnormal points is labeled as abnormal clusters; otherwise, the cluster is labeled as normal

$$C_k \in \begin{cases} \text{Abnormality}, & \text{if } \sum f_{\text{abn}}^{C_k} > \alpha \sum f_n^{C_k}, \\ \text{Normality}, & \text{if } \sum f_{\text{abn}}^{C_k} < \alpha \sum f_n^{C_k}, \end{cases} \tag{23}$$

where $\sum f_{\text{abn}}^{C_k}$ and $\sum f_n^{C_k}$ are the total number of abnormality and normality features in the cluster C_k, respectively. We found the value of α equal to 1.2 to be a proper choice for this threshold.

In (23), we choose the classes that represent the abnormality in the speech. The equation distinguishes a cluster as abnormal if the number of the features estimated from the pathological voice is more than features derived from the normal speech. The abnormality clusters are denoted as $C_{\text{abn}}^{\text{LF}}$ and $C_{\text{abn}}^{\text{HF}}$ for LF and HF groups, respectively.

5.3.2. Testing Stage. In this stage, we test the trained classifier. For a voice sample, we find the nearest cluster to each of its feature vectors using Euclidean distance criterion. If the number of the feature vectors that belong to the abnormality clusters is dominant, the voice sample is classified as a pathological voice; otherwise, it is classified as a normal speech.

Figure 4 demonstrates the testing stage. $f_{\text{Test}}^{\text{LF}}$ and $f_{\text{Test}}^{\text{HF}}$ feature vectors are derived from the base and coefficient vectors in LF and HF groups, respectively. For each feature vector, we find the closest cluster, C_{k_0}, as given in

$$\begin{aligned} f_t^{\text{LF}} \in C_{k_0}^{\text{LF}} \quad \text{if } k_0 = \min_{k=1,\ldots,K} \sqrt{\sum_{i=1}^{4}\left(f_t^{\text{LF}}(i) - C_k^{\text{LF}}(i)\right)^2}, \\ t = 1,\ldots,T_{\text{LF}}, \\ f_t^{\text{HF}} \in C_{k_0}^{\text{HF}} \quad \text{if } k_0 = \min_{k=1,\ldots,K} \sqrt{\sum_{i=1}^{5}\left(f_t^{\text{HF}}(i) - C_k^{\text{HF}}(i)\right)^2}, \\ t = 1,\ldots,T_{\text{HF}}, \end{aligned} \tag{24}$$

where f_t^{LF} and f_t^{HF} are the input feature vectors, and T_{HF} and T_{LF} are the total numbers of test feature vectors for HF and LF components, respectively.

Next, the number of all the features that belong to abnormal and normal clusters is calculated

$$\begin{aligned} &\text{if } C_{k_0}^{\text{LF}} \in C_{\text{abn}}^{\text{LF}} \Longrightarrow \text{abn}_{\text{test}}^{\text{LF}} = \text{abn}_{\text{test}}^{\text{LF}} + 1, \\ &\text{if } C_{k_0}^{\text{HF}} \in C_{\text{abn}}^{\text{HF}} \Longrightarrow \text{abn}_{\text{test}}^{\text{HF}} = \text{abn}_{\text{test}}^{\text{HF}} + 1, \end{aligned} \tag{25}$$

where $\text{abn}_{\text{test}}^{\text{LF}}$ and $\text{abn}_{\text{test}}^{\text{HF}}$ are the numbers of all the feature vectors of LF and HF groups that belong to an abnormal cluster. The signal is classified as normal if

$$L_{\text{abnormality}} < Th_{\text{patho}}, \tag{26}$$

where Th_{patho} is the abnormality threshold, and $L_{\text{abnormality}}$ is the number of the abnormality features in the voice sample:

$$L_{\text{abnormality}} = \left(\frac{\text{abn}_{\text{test}}^{\text{LF}}}{T_{\text{LF}}} + \frac{\text{abn}_{\text{test}}^{\text{HF}}}{T_{\text{HF}}}\right). \tag{27}$$

If the criterion in (26) is not satisfied, the signal is classified as a pathological speech.

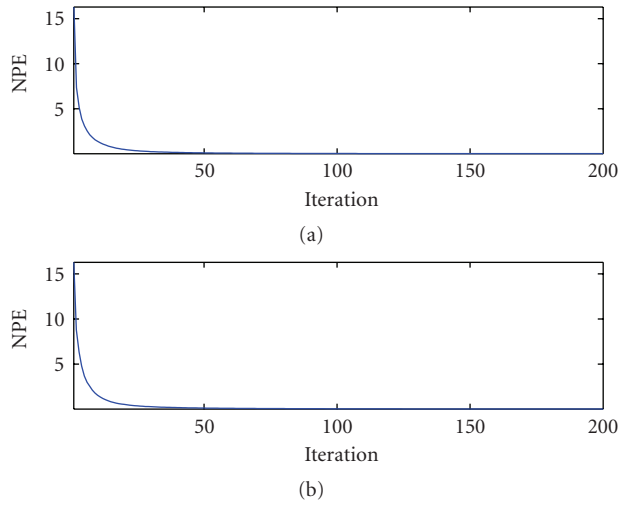


Figure 5: The normalized projected energy (NPE) at each iteration is plotted for one normal (a) and one pathological signal (b). As it can be observed in this figure, most of the coherent structure of the signal is projected before 100 iterations, and the remaining energy is negligible.

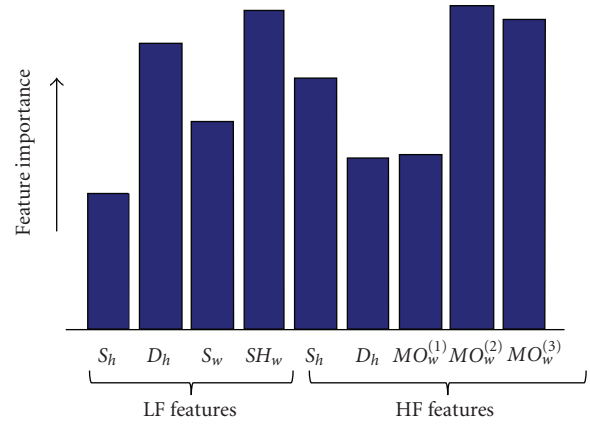


Figure 6: The relative height of each feature represents the relative importance of the feature compared to the other features.

6. Results

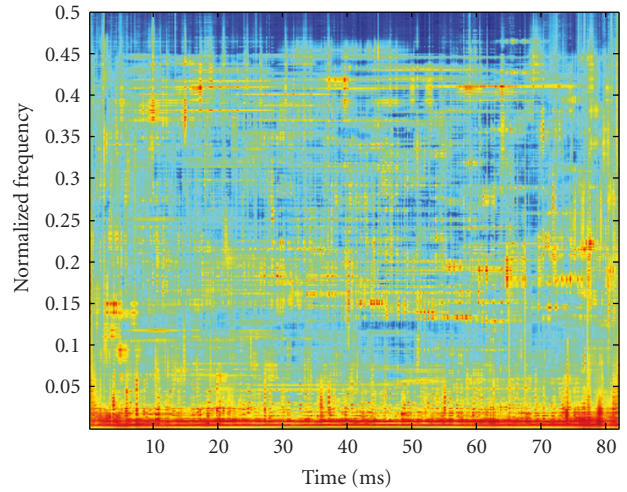
The proposed methodology was applied to the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database, distributed by Kay Elemetrics Corporation [20]. The database consists of 51 normal and 161 pathological speakers whose disorders spanned a variety of organic, neurological, traumatic, and psychogenic factors. The speech signal is sampled at 25 kHz and quantized at a resolution of 16 bits/sample. In this paper, 25 abnormal and 25 normal signals were used to train the classifier.

MP-TFD with Gabor atoms is estimated for each 80 ms of the signal. Gabor atoms provide optimal TF resolution in the TF plane and have been commonly used in MP-TFD. To acquire the required iterations (I) in the MP decomposition, we calculate the energy of the projected signal at each iteration, $\langle R^i x, g_{\gamma_i}\rangle$ in (A.2). Figure 5 illustrates the mean of the projected energy per iteration for one normal and one pathological signal. As evident in this figure, most of the coherent structure of the signal is projected before 100 iterations. Therefore, in this paper, MP-TFD is constructed using the first 100 iterations and the remaining energy is ignored. As explained in Section 3.1, the Adaptive TFD is constructed by performing MCE iterations to the estimated MP-TFD. It can be shown that after 5 iterations, the constructed TFD satisfies the marginal criteria in (5).

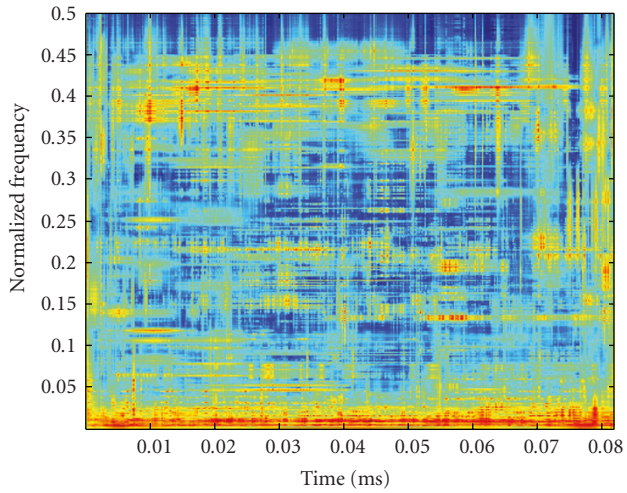
Next, we apply NMF-MD with base number of $r = 15$ to each TF matrix and estimate the base and coefficient matrices, W and H, respectively. Each base vector is categorized into either LF or HF group a base vector is grouped as LF component if its energy is concentrated more in the frequency range of 4 kHz or less; otherwise, it is grouped as HF component. We extract 4 features (S_h, D_h, S_w, SH_w) from each LF base vector w and its coefficient vector h, and 5 features $(S_h, D_h, MO_w^{(1)}, MO_w^{(2)}, MO_w^{(3)})$ from each HF base vector and its coefficient vector. In order to obtain the role of each feature in the classification accuracy, we calculate the P-value of each feature using the Student's t-test. The feature with the smallest P-value plays the most important role in the classification accuracy. Figure 6 demonstrates the relative importance of each 9 features. As shown in this figure, D_h and SH_w from LF features, and S_h, $MO_w^{(2)}$ and $MO_w^{(3)}$ from HF features play the most significant role in the classification accuracy.

Finally, we apply the K-means clustering to the logarithm of the derived feature vectors, and define the abnormality clusters. Figures 7 illustrates the application of the proposed methodology for a pathological voice sample which is shown in Figure 7(a). As explained in Section 5.3, the test procedure determines the feature vectors that belong to the abnormality clusters. We use the base and coefficient matrices, W_{abn} and H_{abn}, corresponding to the abnormality feature vectors to reconstruct the abnormality TF matrix, V_{abn}, as $V_{\text{abn}} = W_{\text{abn}} H_{\text{abn}}$. Figure 7(b) depicts the reconstructed TF matrix. As it is expected, the proposed method successfully distinguishes transients, high frequency components, and week formants as abnormality.

In the test stage, the trained classifier is used to calculate the measure of abnormality ($L_{\text{abnormality}}$ in (27)) for each voice sample. Figure 8 shows the abnormality measure for 51 normal and 161 pathological speech signals in MEEI database. As evident in this figure, the pathological samples have higher abnormality measure compared to the normal samples. Each signal is classified as normal if its abnormality measure is smaller than a threshold (Th_{patho} in (26)); otherwise it is classified as pathological. In order to find the abnormality threshold, receiver operating curves (ROCs) of $L_{\text{abnormality}}$ are computed with the area under the curve indicating relative abnormality detection (Figure 9). Based on the ROC, the cut point of 0.59 is chosen as the abnormality threshold ($Th_{\text{patho}} = 0.59$). Table 1 shows the accuracy of the classifier. From the table, it can be observed that out of 51 normal signals, 50 were classified as normal, and only 1 was misclassified as pathological. Also, the table shows that out of 161 pathological signals, 159 were classified



(a) TFD of a pathological speech



(b) TFD of the estimated abnormality

Figure 7: The classifier of Figure 4 is applied to the TF matrix of a pathological speech shown in (a), and the estimated abnormality TF matrix is shown in (b) . As evident in this figure, the abnormality components are mainly transients, high frequency components, and week formants.

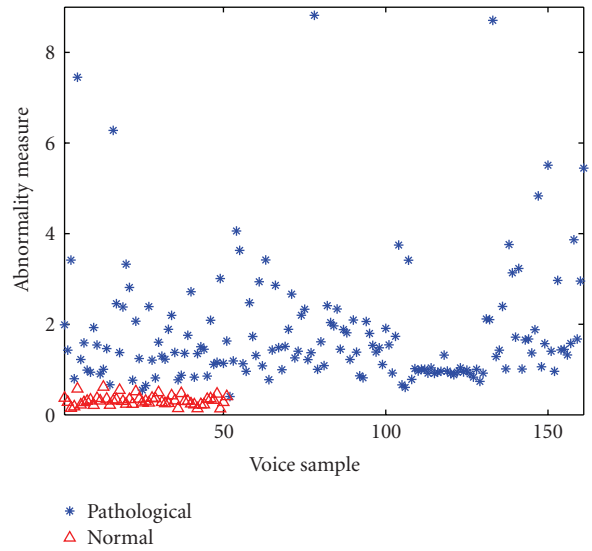


Figure 8: For each voice sample, the number of the feature vectors that belong to an abnormality cluster is calculated, and the abnormality measure is calculated as the ratio of the total number of the abnormal feature vectors to the total number of feature vectors in the voice sample.

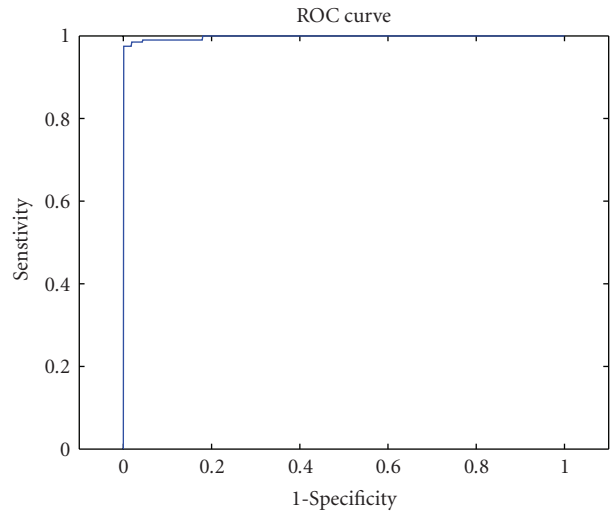


Figure 9: Receiver operating curve for the pathological voice classification is plotted. In this analysis, pathological speech is considered negative, and normal is considered positive. The area under the ROC is 0.999, and the maximum sensitivity for pathological speech detection while preserving 100% specificity is 98.1%.

as pathological and only 2 were misclassified as normal. The total classification accuracy is 98.6%. As it can be concluded from the result, the extracted features successfully discriminate the abnormality region in the speech.

In Figure 9 and Table 1, we utilized MD with decomposition order (r) of 15. We repeated the proposed method using different decomposition orders. Our experiment showed that the decomposition order of 5 and higher is suitable for our application. Table 2 shows the P-values of three decomposition orders obtained with the Student's t-test.

As explained in Section 2, our proposed feature extraction methodology performs a longer term modeling compared to the current methods. The pathological speech classification is conventionally performed on 10–30 ms of signal. At sampling frequency of 8 kHz, the number of sample is 80–240 samples per segment. In this paper, we use 80 ms of speech at sampling frequency of 25 kHz. As a result, we are working with 2048 samples/frame which is 10 times the conventional length. The results shown in this section demonstrate that the proposed methodology successfully discriminates the pathological characteristics of the speech. In addition to the high accuracy rate, the advantage of our proposed methodology can be concluded in 3 points. (1) By performing MP on the speech signal, we project the most coherent structure of the signal. The

Table 1: Classification result.

Classes	Normal	Abnormal	Total
Normal	50	1	51
Pathological	2	159	161
Normal	98.0%	2.0%	100%
Pathological	1.2%	98.8%	100%

Table 2: P-value of the classifiers obtained with three different decomposition orders.

Decomposition order (r)	5	10	15
P-value	3×10^{-10}	1×10^{-11}	1×10^{-13}

remaining part represents the random noise presented in the signal. Hence, we perform an automatically denoising on the signal which allows the technique to be practical in the low SNR speech signals. (2) In this method, we reconstruct the TF matrix of the abnormality part of the signal, and we estimate the amount of abnormality in the speech signal. The reconstructed TF matrix and the abnormality measure have potential to be used as a patients' progress measure over the course of voice therapy. (3) In this work, we use a very simple classifier rather than a complex classifier, such as hidden Markov models or neural networks.

7. Conclusion

TF analysis are effective for revealing non-stationary aspects of signals such as trends, discontinuities, and repeated patterns where other signal processing approaches fail or are not as effective; however, most of the TF analysis are restricted to visualization of TFDs and do not focus on quantification or parametrization that are essential for feature analysis and pattern classification.

In this paper, we presented a joint TF and MD feature extraction approach for pathological voice classification. The proposed methodology extracts meaningful speech features that are difficult to be captured by other means. TF features are extracted from a positive TFD that satisfies the marginal conditions and can be considered as a true joint distribution of time and frequency. The utilized TFD is a segment free TF approach, and it provides a high-resolution and cross-term free TFD.

The TF matrix was decomposed into its base (spectral) and coefficient (temporal) vectors using nonnegative matrix factorization (NMF) method. Four features were extracted from the components with low frequency structure, and five features were derived from the bases with high frequency composition. The features were extracted from the decomposed vectors based on the spectral and temporal characteristics of the normal and pathological signals. In this study, we performed K-means clustering to the proposed feature vectors, and we achieved an accuracy rate of 98.6% for the MEEI voice disorders database, including 161 pathological and 51 normal speakers.

Appendices

A. Matching Pursuit TFD

Matching pursuit (MP) was proposed by Mallat and Zhang [21] in 1993 to decompose a signal into Gabor atoms, g_{γ_i}, with a wide variety of modulated frequency (f_i) and phase (ϕ_i), time shift (p_i) and duration (s_i) as shown in

$$g_{\gamma_i}(t) = \frac{1}{\sqrt{s_i}} g\left(\frac{t - p_i}{s_i}\right) \exp\left[j(2(\pi f_i t + \phi_i))\right], \tag{A.1}$$

where γ_i represents the set of parameters (s_i, p_i, f_i, ϕ_i). The MP dictionary is consisted of Gabor atoms with durations (s_i) varying from 2 samples to N (length of the signal $x(t)$), and it therefore is a very flexible technique for non-stationary signal representation. At each iteration, the MP algorithm chooses the Gabor atom that best fits to the input signal. Therefore, after I iterations, MP procedure chooses the Gabor atoms that best fit to the signal structure without any preassumption about the signal's stationarity. Components with long stationarity properties will be represented by long Gabor atoms, and transients will be characterized by short Gabor atoms.

At each iteration, MP projects the signal into a set of TF atoms as follows:

$$x(t) = \sum_{i=0}^{I-1} \left\langle R_x^i, g_{\gamma_i} \right\rangle g_{\gamma_i}(t) + R_x^I, \tag{A.2}$$

where $\langle R_x^i, g_{\gamma_i} \rangle$ is the expansion coefficient on atom $g_{\gamma_i}(t)$, and R_x^I is the decomposition residue after I decomposition. At this stage, the selected components represent coherent structures and the residue represents incoherent structures in the signal. The residue may be assumed to be due to random noise, since it does not show any TF localization. Therefore, the decomposition residue in (A.2) is ignored, and the Wigner-Ville distribution (WVD) of each I components is added in the following:

$$\hat{V}(t, f) = \sum_{i=0}^{I-1} \left| \left\langle R_x^i, g_{\gamma_i} \right\rangle \right|^2 W g_{\gamma_i}(t, f), \tag{A.3}$$

where $W g_{\gamma_i}(t, f)$ is the WVD of the Gabor atom $g_{\gamma_i}(t)$, and $\hat{V}(t, f)$ is called the MP-TFD. Wigner distribution is a powerful TF representation; however when more than one component is present in the signal, the TF resolution will be confounded by cross-terms. Nevertheless, when we apply the Wigner distribution to single components and add them up, the summation will be a cross-term free TFD.

B. Analysis of sharpness feature

In order to demonstrate the behavior of feature SH_w, we assume that the base vector, w_i, has two components at frequencies samples m_1 and m_2 with energies of α and β, respectively

$$w_i(m) = \alpha\delta(m - m_1) + \beta\delta(m - m_2), \tag{B.1}$$

$|W(\nu)|$ (21) is calculated as

$$|W(\nu)| = \sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\cos(2\pi(m_1 - m_2)\nu)}. \tag{B.2}$$

$|W(\nu)|$ is independent to the parameter ν only when $m_1 \approx m_2$, or when the energy ratio of the components in (B.1) is too small (either $\beta/\alpha \approx 0$ or $\alpha/\beta \approx 0$). In this case, when we calculate the Fourier transform of $|W(\nu)|$ as shown in (21), $|W(\kappa)|$ is non-zero only at small values of κ (say $\kappa < m_0$, where m_0 is a small number). Hence, SH_{w_i} as it is calculated in (22) results in a small feature. From the other side, $|W(\nu)|$ is dependent on the parameter ν when both the components in (B.1) are strong ($\beta/\alpha \approx R, R \neq 0$). In this case, the Fourier transform of $|W(\nu)|$ is not negligible at $\kappa > m_0$, and SH_{w_i} results in lager values.

From the above explanation, we conclude that the small values of SH_{w_i} represent pathological formants, in which the components' energies are very small compared to the energy of the main frequency ($\beta/\alpha \approx 0$ or $\alpha/\beta \approx 0$), and the large values of SH_{w_i} show the strong formants in speech ($\beta/\alpha \approx R, R \neq 0$).

References

[1] R. T. Sataloff, *Professional Voice: The Science and Art of Clinical Care*, Raven Press, New York, NY, USA, 1991.

[2] P. Carding and A. Wade, "Managing dysphonia caused by misuse and abuse," *British Medical Journal*, vol. 321, pp. 1544–1545, 2000.

[3] E. J. Wallen and J. H. L. Hansen, "A screening test for speech pathology assessment using objective quality measures," in *Proceedings of International Conference on Spoken Language Processing (ICSLP '96)*, vol. 2, pp. 776–779, Philadelphia, Pa, USA, October 1996.

[4] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, 2006.

[5] T. Ananthakrishna, K. Shama, and U. C. Niranjan, "k-means nearest neighbor classifier for voice pathology," in *Proceedings of the IEEE India Conference INDICON*, pp. 352–354, Indian Institute of Technology, Kharagpur, India, 2004.

[6] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology (EMBS '02)*, vol. 1, pp. 182–183, Houston, Tex, USA, 2002.

[7] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[8] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 36–47, 2008.

[9] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.

[10] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, pp. 469–485, 2000.

[11] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 3, pp. 421–430, 2005.

[12] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[13] L. Cohen and T. E. Posch, "Positive time-frequency distribution functions," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 1, pp. 31–38, 1985.

[14] S. Krishnan, R. M. Rangayyan, G. D. Bell, and C. B. Frank, "Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 6, pp. 773–783, 2000.

[15] B. Ghoraani and S. Krishnan, "Quantification and localization of features in time-frequency plane," in *Proceedings of Canadian Conference on Electrical and Computer Engineering (CCECE '08)*, pp. 1207–1210, May 2008.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS '01)*, pp. 556–562, 2001.

[17] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[18] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[19] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[20] M. Eye and E. Infirmary, *Voice Disorders Database, Version 1.03*, Kay Elemetrics Corporation, Lincoln Park, NJ, USA, 1994.

[21] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.