Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 85286, 9 pages
doi:10.1155/2007/85286

Research Article

Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology

Kumara Shama,[1] Anantha krishna,[1] and Niranjan U. Cholayya[2]

[1] *Department of Electronics and Communication Engineering, Manipal Institute of Technology, 576104 Manipal, India*
[2] *Department of Biomedical Engineering, Manipal Institute of Technology, 576104 Manipal, India*

Received 5 April 2005; Revised 5 January 2006; Accepted 13 January 2006

Recommended by Douglas O'Shaughnessy

Acoustic analysis of speech signals is a noninvasive technique that has been proved to be an effective tool for the objective support of vocal and voice disease screening. In the present study acoustic analysis of sustained vowels is considered. A simple k-means nearest neighbor classifier is designed to test the efficacy of a harmonics-to-noise ratio (HNR) measure and the critical-band energy spectrum of the voiced speech signal as tools for the detection of laryngeal pathologies. It groups the given voice signal sample into pathologic and normal. The voiced speech signal is decomposed into harmonic and noise components using an iterative signal extrapolation algorithm. The HNRs at four different frequency bands are estimated and used as features. Voiced speech is also filtered with 21 critical-bandpass filters that mimic the human auditory neurons. Normalized energies of these filter outputs are used as another set of features. The results obtained have shown that the HNR and the critical-band energy spectrum can be used to correlate laryngeal pathology and voice alteration, using previously classified voice samples. This method could be an additional acoustic indicator that supplements the clinical diagnostic features for voice evaluation.

1. INTRODUCTION

Diseases that affect the larynx cause changes in the patient's vocal quality. Early signs of deterioration of the voice due to vocal malfunctioning are normally associated with breathiness and hoarseness of the produced voice. The first tool used to detect laryngeal pathology is subjective analysis of the speech. Trained physicians perform a subjective evaluation of the patient's voice, which is followed by laryngeoscopy that may cause discomfort to the patient. A complementary technique could be acoustic analysis of the speech signal, which is shown to be a potentially useful tool to detect voice disease. This noninvasive technique is a fast low-cost indicator of possible voice problems.

Any change in the anatomical structure because of pathology in turn results in physiological function that alters the vocal output [1–7]. The analysis methods found in the literature are mainly based on the periodicity of vocal fold vibration and the turbulence in the glottal flow resulting from malfunctioning of the vocal folds [8–17]. The periodicity perturbations are associated with the measurement of jitter and shimmer. Jitter is the variation between the successive fundamental periods and shimmer is the variation between successive magnitudes of the signal from cycle to cycle. The turbulence in the glottal flow is usually quantified by the noise components in the voiced speech spectrum. In this study we focus on the vocal noise for the analysis of vocal fold pathology.

Researchers have extensively used the vocal noise for the evaluation of pathologic voice. Many noise features have been used which are designed to quantify the relative noise components in a speech signal. The prominent ones are the harmonics-to-noise ratio (HNR), the normalized noise energy (NNE), and the glottal-to-noise-excitation ratio (GNE). Yumoto et al. [11] proposed the HNR as a measure of hoarseness. But the estimation of HNR is based on the assumption that a long stationary data segment is available for analysis, which may not be realistic as the speech is highly nonstationary. Kasuya et al. [12] proposed NNE as a novel and effective acoustic measure to evaluate noise components in pathologic voices. They have devised an adaptive comb filtering method operating in the frequency domain to estimate noise components and NNE from a sustained vowel phonation. A fixed length (seven times the fundamental pitch

period) voiced segment is used for the analysis. Manfredi [13] used an adaptive window, whose length is adapted according to the fundamental pitch period for the analysis. The adaptive NNE proposed by them is particularly useful for complete word utterances. Michaelis et al. [16] have proposed a new acoustic measure called GNE for the objective description of voice quality. This parameter is related to the breathiness in the voiced speech and it indicates whether a given voice signal originates from the vibration of the vocal folds or from the turbulent noise generated in the vocal tract.

In this paper, we extract two different sets of features from the acoustic analysis of voiced speech and further use them to correlate laryngeal pathology and voice alteration on a previously classified database of voice samples. The first feature set is the energy ratio of harmonics to noise components (HNR) in the voiced speech signal at four different frequency bands and the second set of features is based on the energy spectrum at critical-band spacing [18]. A k-means nearest neighbor classifier [19] is used separately on these sets of features to test their efficacy as tools for the detection of laryngeal pathology. As the same classifier is used on the two feature sets independently, we get two different sets of classification results. As we have used a preclassified database of voices, this allows us to make a comparison between the efficacies of the two sets of features apart from their individual efficiencies.

2. MATERIALS AND METHODS

2.1. Database

In the present study, we wanted to understand if HNR and critical-band energy spectrum could be used as effective tools for the classification of normal and pathologic voices. A prior-labeled database is helpful in such a study to correlate the results obtained. We have taken the speech signals from such a database distributed by Kay Elemetrics Corporation. This CD ROM database of acoustic records originally developed by Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab. [20] contains over 1400 voice signals of approximately 700 subjects. Included are the sustained phonation and running speech samples from patients with a wide variety of organic, neurological, traumatic, and psychogenic disorders, as well as from 53 normal subjects. We have used voice samples of sustained phonation of the vowel /a/. The recordings were made in a controlled environment and data were available at sampling frequencies of 25 KHz or 50 KHz. We have down sampled all the voice signals to a sampling frequency of 16 KHz. The normal voice records are about 5 seconds long, whereas the pathologic voice records are about 3 seconds long. 53 normal and 163 pathologic voice signals have been used in our study as shown in Table 1. Approximately 50 percent of the signals of each group were considered for training (to estimate the prototype) and the remaining for testing.

Table 1: Details of the voice signals used in the study.

Laryngeal disease	No. of samples
Normal	53
Adductor	13
Paralysis	53
Cyst	04
Leukoplakia	20
Vocal fold polyp	13
Polyp degenerative	17
Vocal fold edema	30
Vocal nodule	13

2.2. Estimation of HNR

One of the important characteristics of voiced speech is the well-defined harmonic structure. The source for the voiced speech is often modeled as quasiperiodic glottal pulses. But in reality, even the sustained vowel phonation consists of some random part mainly due to turbulence of airflow through the glottis (anterior and/or posterior glottis) and due to pitch perturbations. A windowed segment $s(n)$ of the voiced speech signal is therefore assumed to have a periodic component $p(n)$ and a random component $w(n)$, represented as

$$s(n) = p(n) + w(n), \quad n = 0, 1, \ldots, M-1, \tag{1}$$

where M is the length of the analysis window. The two components cannot be directly separated because the random component may have energy in the entire speech spectrum. But one can get an estimate of the random component by decomposing speech into periodic and random components. We have used a method similar to the one proposed by Yegnanarayana et al. [21] for the decomposition of the speech into periodic and aperiodic components. The method involves an initial approximation of the periodic and the random components using the harmonicity criterion. This is followed by an iterative reconstruction of the random component in the region labelled as "periodic" based on discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) pairs.

2.2.1. Identification of harmonic and noise regions

The first step in the signal decomposition algorithm is to derive a first approximation of periodic and aperiodic components in the frequency domain. The spectrum of a windowed voiced speech segment is schematically shown in Figure 1. An N point DFT of a Hamming windowed segment of length M of the voiced speech is assumed. The harmonic peak region P_i has a width of $2N/M$ on either side of the peak frequency k_i corresponding to the ith harmonic of the fundamental frequency. $2N/M$ is the approximate bandwidth of the Hamming window. This region contains both periodic and aperiodic energy. In the harmonic dip region D_i, it is assumed that the periodic components have no energy and the entire energy is due to random components. In order to obtain nonempty dip region with d points, the window length

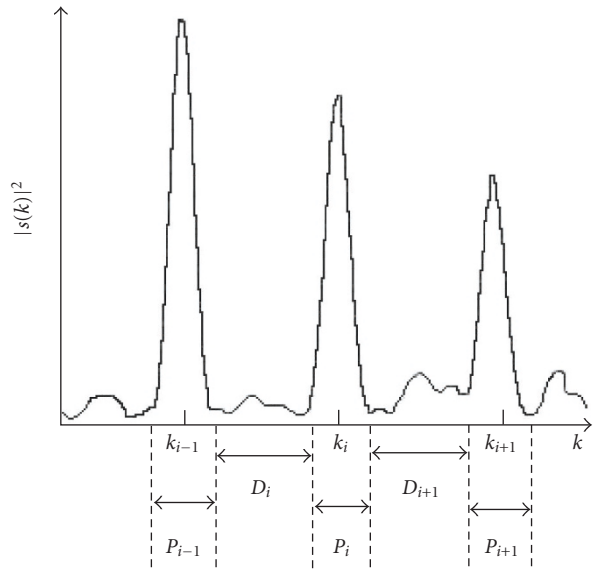

Figure 1: Schematic representation of the spectrum of a windowed voiced speech segment.

should satisfy [13]

$$M \geq \frac{4N}{f_0 NT - (d+1)}, \tag{2}$$

where f_0 is the fundamental frequency of phonation and T is the sampling interval. Thus with a nonempty dip region, one can identify the harmonic region and noise region as

$$\begin{aligned} P_i &= \left\{ k \mid k_i - \frac{2N}{M} \leq k \leq k_i + \frac{2N}{M} \right\}, \\ D_i &= \left\{ k \mid k_{i-1} + \frac{2N}{M} \leq k \leq k_i - \frac{2N}{M} \right\}, \end{aligned} \tag{3}$$

where k = frequency number. A peak-searching algorithm is used to initially locate the harmonic peak frequencies k_i. This algorithm determines the spectral peaks by searching for the peaks in the intervals centered at each multiple of the fundamental frequency f_0. The fundamental frequency is estimated using the method described in Section 2.2.2 below.

2.2.2. Estimation of f_0

Sufficient subglottal air pressure and vocal fold adduction produce oscillation of the vocal folds and therefore voiced sounds when the vocal fold tissues are pliable. The rate of vibration is the fundamental frequency. The glottis opens and closes, resulting in quasiperiodic flow of air. The instant of closure of the glottis is referred to as the glottal closure instant (GCI). During each period of voiced speech, a GCI occurs. To detect this, Wendt and Petropulu [22] used a wavelet function having a derivative property. When the speech signal is filtered by this function, maxima will occur at every GCI. For many phonation cases, normal and abnormal, the vocal folds do not come all the way together, and there is no glottal closure. However, there can be a more prominent flow reduction within a cycle, and therefore a greater acoustic excitation at that time in the cycle. Many of the pathological voices will not have closure, but will have stronger excitation moments somewhere in the cycle. Such voiced speech signals also exhibit prominent peaks when filtered through the wavelet filtering function at these stronger excitation moments. Thus the time elapsed between two adjacent maxima of the filtered signal represents the pitch period of the signal at that moment. We propose an extension to this method to estimate the pitch.

To construct a filtering function, the wavelet with the derivative property described by Mallat and Zhong [23] is combined with the bandwidth property of the wavelet transform at different scales. Let $\psi(t)$ be the mother wavelet with derivative property. The functions

$$\psi_k(t) = 2^{k/2}\psi(2^k t), \qquad \phi_k(t) = 2^{k/2}\phi(2^k t) \tag{4}$$

represent Haar wavelet and scaling functions, respectively, at scale k. Here $\phi(t)$ is a lowpass function and is the conjugate mirror filter of $\psi(t)$, which is a highpass function. As the approximate range of the fundamental frequency of the voiced speech is between 60 and 500 Hz [24], the final filtering function should have the same bandwidth. Thus we construct a filtering function $\lambda(t)$ as

$$\lambda(t) = \phi_{k_a}(t) * \psi_{k_b}(t), \tag{5}$$

where $*$ denotes convolution. The scales k_a and k_b are given by

$$\begin{aligned} k_a &= \left\lceil \log_2 \frac{F_s}{500} \right\rceil, \\ k_b &= \left\lceil \log_2 \frac{F_s}{60} \right\rceil, \end{aligned} \tag{6}$$

where F_s is the bandwidth of the input speech signal.

The speech signal is passed through this filter. The filtered signal shows dominant peaks at the GCIs. The peaks of the filtered signal are detected using a peak detection algorithm which identifies the peaks by detecting the points where the slope polarity change occurs. For real speech, the filtered signal exhibits some spurious peaks which are to be eliminated by using a suitable peak correction method. Thresholding the strength and the proximity of the adjacent peaks [25] is used in the peak correction algorithm. That is, in the first stage of correction, a peak is validated only if its amplitude is above a threshold. The threshold is fixed at 25 percent of the average peak amplitude. In the second stage, the average distance D_a between the two adjacent peaks is first estimated. Every peak whose distance with its adjacent peak is shorter than $0.5D_a$ or longer than $2D_a$ is then eliminated. This two-stage peak correction algorithm eliminates the spurious peaks and identifies only the correct peaks. The average distance between the consecutive peaks is then found to compute the pitch period and hence the fundamental frequency f_0.

2.2.3. Estimation of harmonic and noise energies

By estimating the signal energies in the identified harmonic and noise regions (Section 2.2.1), one can get only an approximate harmonics-to-noise ratio. The energy in a noise region is assumed to be due to noise components only, but in the harmonic region, the energy is a superposition of harmonic and noise components. The noise energy can be estimated by signal extrapolation methods. In this paper, we have used an iterative algorithm developed by Yegnanarayana et al. [21] to reconstruct the noise components. The algorithm is based on bandlimited signal extrapolation proposed by Papoulis [26]. The noise component is reconstructed by iteratively moving from the frequency domain to the time domain and vice versa. For an M-length signal, an N-point ($N > M$) DFT is first obtained. The iterations begin with zero values in the frequency region identified as the harmonic region and actual DFT values in the noise region. An inverse DFT is then obtained and the first M points of the resulting signal are retained. An N-point DFT is again obtained and the harmonic region is forced to zero. The IDFT is computed and this procedure is repeated for a few iterations. It is shown [21] that for a finite duration signal with known noise samples, the reconstructed noise component converges to the actual noise component in the mean-square sense, as the iterations grow. In fact, after a number of iterations (about 8 to 10), the noise components are reconstructed with negligible error. After reconstructing the noise components, the harmonic components are obtained by time domain subtraction. From these components the harmonics-to-noise energy ratio in the required frequency bands is estimated.

2.3. critical-band energy spectrum

The effect of noise on speech has been found to change the spectral characteristics. Marked differences are found in the distribution of energy at critical-bands between clean and noisy speech signals [27]. This difference factor was effectively used to differentiate the clean speech from the speech added with noise. We extend this idea to differentiate pathologic voices from the normal ones, as the voiced speech of subjects with vocal fold pathology has additional noise components caused mainly by the incomplete closure of the glottis and improper vibration pattern of the vocal folds. We have used energy spectra at critical-bands because the center frequency and bandwidths of the critical-bands roughly correspond to the tuning curves of human auditory neurons. The human auditory system is assumed to perform a filtering operation, which partitions the audible spectrum into critical-bands [28]. Twenty one critical-bands described in Table 2. [27] have been used in this work. Thus the proposed automated analysis mimics the human perceptual analysis of voice pathology. These 21 bands cover the frequency range from 1 to 7.7 KHz. The bandwidths at lower critical-bands are narrower and they progressively increase as the center frequency increases.

Table 2: Upper-edge frequencies, lower-edge frequencies, center frequencies, and bandwidths for 21-channel filter-bank with *critical-band spacing*.

Band	Lower-edge frequency (Hz)	Upper-edge frequency (Hz)	Center frequency (Hz)	Bandwidth (Hz)
1	1	100	50	100
2	100	200	150	100
3	200	300	250	100
4	300	400	350	100
5	400	510	450	110
6	510	630	570	120
7	630	770	700	140
8	770	920	840	150
9	920	1080	1000	160
10	1080	1270	1170	190
11	1270	1480	1370	210
12	1480	1720	1600	240
13	1720	2000	1850	280
14	2000	2320	2150	320
15	2320	2700	2500	380
16	2700	3150	2900	450
17	3150	3700	3400	550
18	3700	4400	4000	700
19	4400	5300	4800	900
20	5300	6400	5800	1100
21	6400	7700	7000	1300

We have adopted a filter bank approach for the estimation of energy. Sixth-order Butterworth bandpass filters are used to obtain the 21 band filter bank. The filter bank approach is preferred due to its simple and inexpensive implementation. This approach is particularly suitable when a small set of parameters describing the spectral distribution of energy has to be derived. The outputs from a bank of 21 bandpass filters typically provide a very efficient spectral representation.

In the next section, we describe the extraction of the features and the design of the classifier.

2.4. Feature estimation

2.4.1. Features based on HNR

One of the important characteristics of normal voiced speech is that it exhibits a good harmonic structure even up to about 4 KHz. In contrast, the pathologic voices exhibit higher noise levels and the noise is distributed across the entire speech spectrum. The pathologic voices may have good harmonic structure at low frequencies, and at higher frequencies the harmonic energy decreases with the increase in noise energy. This is evident from Figure 2 where the log magnitude spectra of the estimated harmonic component and noise components for a segment of speech corresponding to sustained vowel /a/ uttered by both a normal and a pathologic subject

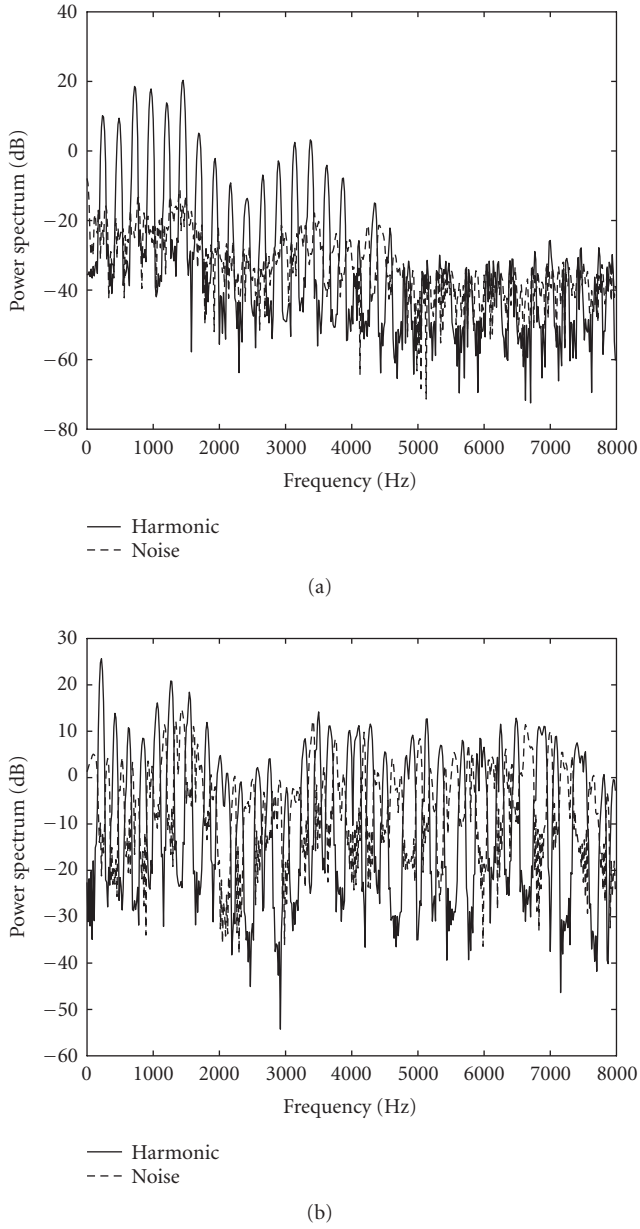

Figure 2: Power spectra of the estimated harmonic and noise components for the vowel segment /a/ corresponding to (a) a normal subject and (b) a pathologic subject.

are shown. The harmonic and the noise components are obtained by decomposing the segment of the speech signal using the method discussed in Section 2. The normal voice shows a regular harmonic structure up to about 4 KHz with relatively low noise energy. In the case of the pathologic voice, the spectrum shows higher noise levels with deteriorated harmonic structure even at lower frequencies. The harmonics-to-noise energy ratio (HNR) at different frequency bands can therefore be used for discriminating pathologic voices from normal ones. In this study, we have used HNRs at four different frequency bands as the features for the classification as shown in Table 3. These frequency bands are the

Table 3: The frequency bands in which the HNR values are estimated.

Band number	Lower-edge frequency (Hz)	Upper-edge frequency (Hz)	Center frequency (Hz)
1	300	620.8	460.4
2	620.8	1248.5	925.6
3	1248.5	2658	1971.3
4	2658	5500	4079

standard bands used in many speech-processing applications [27] and have logarithmic spacing that would approximate the frequency response of human ear. We have experimented with more than 4 frequency bands and no significant improvement in the results was found. Using frequencies above 5.5 KHz also had no significant effect on the results because both the normal and pathologic voices show low HNR above this frequency.

The speech recordings corresponding to the sustained vowel /a/ are sampled at 16 KHz and digitized with 16- bit resolution. The data are then segmented into overlapping segments of length 1023 samples. This particular choice of the segment length is based on the following issues. The accuracy of the extrapolation algorithm for the decomposition of the voice signal into harmonic and noise components is poor for low-pitched voices, as the numbers of sample points available in the harmonic dip region for the extrapolation are fewer. At lower pitch, to have nonempty dip regions, the frame length needs to be higher (see (2)). At the same time, the data window at the higher pitch frequencies spans a large number of pitch cycles. The pitch of the voice samples used in the current study was in the range 90 Hz to 220 Hz. Thus we found the segment length of 1023 points adequate. This also suits the requirements of the iterative procedure based on DFT and IDFT used for the decomposition of speech where we have used 2048 point DFTs.

For each segment, the HNR at the four frequency bands are estimated by the method described in Section 2. These 4 HNRs are then averaged over all the segments. The averaged HNR values form the feature vector for the classifier.

2.4.2. Features based on energy spectrum

The voiced speech data (sustained phonation of vowel /a/) are uniformly divided into 20 ms frames. Each frame is filtered through the 21-channel filter-bank, whose center frequencies and bandwidths are taken according to critical-band spacing. These 21-bands cover a frequency range of 1 to 7.7 KHz. Energies of each of the 21-filter outputs are computed and normalized to the total energy. This normalized energy spectrum is used as a feature vector in this study.

Figure 3 shows an example of normalized energy spectra for normal and pathologic voice signals. Here we have plotted normalized energy (which is the sum of both harmonic and noise energies) versus the frequency bands. It is observed that for the healthy voices considered in the study, most of the energy content is accumulated in critical-bands 5 through

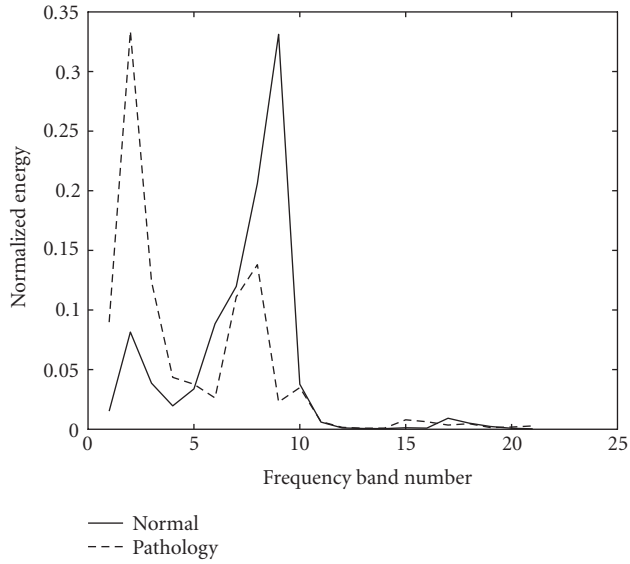

Figure 3: Normalized energy spectrum for normal and pathologic voice samples.

10, which correspond to the frequency range of 400 Hz to 1270 Hz, whereas the pathologic voice does not show such a pattern. Pathologic voices exhibited energy distributions such that considerable energy is seen in lower bands also (critical-bands 1 through 4). It is also evident from Figure 2 where one can see the pathologic voice having large harmonic and noise energy at lower frequencies though the harmonic energy falls rapidly at higher frequencies with the increase in noise energy. However some pathologic voices show a significant amount of energy at higher frequency bands also.

2.5. Classifier

This section describes the design of a classifier to classify the given voice signal to the normal or pathologic class, based on the estimated acoustic features. The distribution functions for these features are unknown and hence nonparametric methods of classification are necessary. There are several techniques available, which include fitting an arbitrary density function to a set of samples, histogram techniques, and kernel or window techniques [29]. Apart from these, there are several nearest neighbor techniques, which do not explicitly use any density functions.

2.5.1. Nearest neighbor classification

This method assigns an unknown sample signal to that class having most similar or nearest sample signal in the reference set or training set of signals. The nearest sample signal is found by using the concept of distance or metric. We have used Euclidean distance as the metric. The Euclidean distance in n-dimensional feature space, which is the usual distance between the two points $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$ is defined by

$$D_e(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}. \tag{7}$$

In the present work, a simple k-means nearest neighbor classifier has been used. This is a variant of the nearest neighbor technique. Here a prototype is computed from the reference set of sample signals and a given test sample signal is classified as belonging to the class of the closest prototype. The prototype is computed as the mean of feature vectors corresponding to signals in the reference set belonging to a particular class. The prototype, referred to as a centroid vector, is computed separately for both normal and pathologic voice signals. This averaging process represents the training phase of the classifier.

2.5.2. Classification based on HNR

Let HNR_{ij} denote the harmonics-to-noise ratio at the ith frequency band for the jth sample signal with $i = 1, 2, 3, 4$. Then the centroid vector is

$$\mathrm{HNR}_i{}^c = \frac{1}{k} \sum_{j=1}^{k} \mathrm{HNR}_{ij}, \tag{8}$$

where c = nc (normal class) or pc (pathologic class) and k = number of sample signals in the reference set belonging to class c.

Two such centroid vectors are computed, one for normal voices and the other for pathologic voices. For the test sample signal, we calculate the Euclidean distance parameter D between the HNR feature vector corresponding to the test sample signal and the centroid vector. Thus we have two distance measures:

$$\begin{aligned} D_{\mathrm{nc}} &= \sqrt{\sum_{i=1}^{4} \left(\mathrm{HNR}_i{}^t - \mathrm{HNR}_i{}^{\mathrm{nc}}\right)^2}, \\ D_{\mathrm{pc}} &= \sqrt{\sum_{i=1}^{4} \left(\mathrm{HNR}_i{}^t - \mathrm{HNR}_i{}^{\mathrm{pc}}\right)^2}, \end{aligned} \tag{9}$$

where $\mathrm{HNR}_i{}^t$ is the ith component of the HNR vector for the test sample signal, $\mathrm{HNR}_i{}^{\mathrm{nc}}$ and $\mathrm{HNR}_i{}^{\mathrm{pc}}$ are the ith components of the centroid vector corresponding to normal and pathologic classes, respectively. D_{nc} and D_{pc} are the distances between the test vector and the corresponding centroid vectors.

The nearest neighbor rule is then applied to assign the test sample signal to normal or pathologic class. The rule is if $D_{\mathrm{pc}} < D_{\mathrm{nc}}$, then the test sample is considered as pathologic, otherwise as normal.

2.5.3. Classification based on energy spectrum

We define spectral distance SD as the Euclidean distance between the feature vector (normalized energy values at the 21-band critical-bands) corresponding to the test sample signal

and that of the centroid vector as

$$\mathrm{SD} = \sqrt{\sum_{i=1}^{21} \left(\mathrm{EB}_i{}^t - \mathrm{EB}_i{}^c\right)^2}, \tag{10}$$

where $\mathrm{EB}_i{}^t$ denotes the ith normalized filter-bank energy output of the test sample and $\mathrm{EB}_i{}^c$ denotes the corresponding energy of the centroid vector. For any given test sample, the two spectral distances, one corresponding to the normal centroid and the other corresponding to the pathologic centroid, are estimated as

$$\begin{aligned} \mathrm{SD}_n &= \sqrt{\sum_{i=1}^{21} \left(\mathrm{EB}_i{}^t - \mathrm{EB}_i{}^{\mathrm{nc}}\right)^2}, \\ \mathrm{SD}_p &= \sqrt{\sum_{i=1}^{21} \left(\mathrm{EB}_i{}^t - \mathrm{EB}_i{}^{\mathrm{pc}}\right)^2}, \end{aligned} \tag{11}$$

respectively, where $\mathrm{EB}_i{}^{\mathrm{nc}}$ and $\mathrm{EB}_i{}^{\mathrm{pc}}$ denote the ith components of the centroid vectors corresponding to normal and pathologic cases, respectively. Based on the above spectral distance measures, the given test sample is classified into the normal class if $\mathrm{SD}_n \leq \mathrm{SD}_p$ or into the pathology class otherwise.

3. PERFORMANCE EVALUATION AND RESULTS

The following parameters were used to evaluate the performance of the classifier.

(1) True positive (TP): the classifier detected pathology when pathology was present.

(2) True negative (TN): the classifier detected normal when normal voice was present.

(3) False positive (FP): the classifier detected pathology when normal voice was present (false acceptance).

(4) False negative (FN): the classifier detected normal when pathology was present (false rejection).

(5) Sensitivity (SE): likelihood that pathology will be detected given that it is present.

(6) Specificity (SP): likelihood that the absence of pathology will be detected given that it is absent.

(7) Accuracy: the accuracy with which the classifier is able to classify the given sample to the correct group.

$$\begin{aligned} \mathrm{SE} &= 100 \cdot \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{SP} = 100 \cdot \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \text{accuracy} &= 100 \cdot \frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TN}+\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}. \end{aligned} \tag{12}$$

The results are depicted in Table 4. These results were calculated based on the number of samples used for testing.

Table 4: Results.

Features (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
HNR	94.94	92.31	94.28
Energy spectrum	91.14	96.15	92.38

4. DISCUSSIONS

The HNR based features provided lower false rejection and thus higher sensitivity than the critical-band energy-spectrum-based feature set. In fact, 4 pathologic cases were rejected falsely out of 79 test cases by the first classifier, whereas 7 of them were falsely rejected by the other classifier. Though significant difference in percentile specificity was seen, the two sets of features provided low false acceptance. The large difference (about 4%) in the specificity was because the number of normal subjects used in the study was small. 26 normal subjects were used for testing the classifiers; the classifier based on HNR features misclassified two of them while the other misclassified one of them. It was observed that for all the samples that were misclassified, there was a large amount of overlap between the features (HNR and energy spectrum) and the two corresponding estimated prototypes (centroids).

The frequency bands used for the estimation of HNR cover frequencies up to 5.5 KHz, whereas the critical-band energy spectrum stops at 7.7 KHz. This does not alter the results significantly as seen in Table 4. This is also evident from Figures 2 and 3, which show that there is no significant spectral energy in the voiced speech above about 5 KHz. The low harmonic energy above 5 KHz results in low HNR for both normal and pathologic cases. Hence using HNR above 5 KHz will not improve the classifier efficiency.

We have considered mainly vocal fold pathologies and normal voices in this study. The method works well for all these cases. The prototypes for individual pathologic cases were not considered because of small sample sizes and hence a comparison of the performance of the classifier in separating individual pathologic cases from normal is not reported in this paper. We have tried interpathology classification using these features, but the results were poor.

The results shown in Table 4 appear to be promising in separating the normal from pathologic voice samples. These results are comparable to those reported by several other research studies [30–33]. In [30], a voice analysis system was developed for the screening of laryngeal diseases using four different types of classifiers based on time and cepstral domain parameters derived from the speech signal of sustained phonation of the vowel /a/. Overall classification accuracy of 93.5% was reported with a test data set consisting of 50 normal and 150 pathologic subjects. In [31], automatic detection of pathologies in voice was done based on "classic" parameters, that is, shimmer, jitter, energy balance, spectral distance, and newly proposed higher-order statistics (HOS)-based parameters. Classification scores of 94.4% and 98.3%, respectively, were obtained using speech data from 100 healthy and 68 pathologic speakers. Though the results are superior to ours, the method is computationally more complex as 5 vowels are analyzed for each speaker and neural network classifiers are used. In more recent studies found in the literature [32, 33], data from the Kay-Elemetrics disordered voice database have been used for the separation

of pathological voices from normal ones. This is the same database that we used in the present study. In [32], a multilayer perceptron network was used on mel-frequency cepstral coefficients (MFCC) to achieve a classification rate of 96%. As in our study, the sustained vowel phonation /a/ was used but the classification was done on a different set of pathologic voice samples (53 normal and 82 pathologic cases). In another recent study [33], a joint time frequency approach was proposed for the discrimination of pathologic voices. Continuous speech data from 51 normal and 161 pathologic speakers were analyzed and overall classification accuracy of 93.4% was reported using linear discriminant analysis (LDA). The method proposed by us in this paper has the advantage that the k-means nearest neighbor classifiers are easy to implement with minimum computational cost. Though the critical-band energy-spectrum-based classifier has comparatively less accurate results, the parameterization is simpler and does not require the estimation of the pitch and noise.

It is well known that laryngeal pathology can lead to a voice disorder. However, all voice disorders are not due to laryngeal pathology. Acoustical variations with normal laryngeal structure and functions, as well as normal acoustical parameters with variation in the laryngeal organs, have been reported in the literature [34, 35]. The results presented here are from an explorative study to look at the efficacy of HNR and energy spectrum at critical-band spacing as diagnostic tools. Both methods described in this paper may give false results in the case of normal voice produced by altered laryngeal function and "pathological" sounding voices because of some muscular imbalance due to behavioral causes or style settings for artistic purposes. However, such cases can be eliminated while recording, by a suitable screening procedure.

5. CONCLUSIONS

A simple k-means nearest neighbor classifier is designed for the classification of pathologic voices. The harmonics-to-noise ratio and energy spectrum at critical-band spacing of speech signals are demonstrated as tools for the differential classification of laryngeal pathology versus normal voice. This can be used as a tool to supplement the perceptual evaluation of speech for the detection of suspected laryngeal pathologies. The method has the advantage that a comparatively shorter length of speech data is sufficient for the analysis. The HNR-based classifier makes use of 4 frequency bands, while the energy spectrum based classifier makes use of 21. The 4 bands used in the first classifier as well as the 21 bands used in the second classifier correspond to the frequency response of auditory neurons of the human ear. Choice of only 4 frequency bands in the first classifier reduces the dimensionality from 21 to 4 when compared to the second classifier. Though the first method has the advantage of working on reduced dimensional features, the computational gain is used up by the need for the extraction of fundamental frequency and the estimation of noise components, which are computationally expensive. For the pathologic voices, estimation of fundamental frequency (f_0) is difficult and for very breathy, almost aphonic voices, the filtered speech may not have dominant peaks or the peaks may be comparable to noise peaks leading to erroneous pitch estimation. In such cases the energy-spectrum-based classifier is preferred, though this method is comparatively less accurate.

REFERENCES

[1] I. R. Titze, *Principles of Voice Production*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1994.

[2] M. Hirano, S. Hibi, R. Terasawa, and M. Fujiu, "Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia," *Journal of Phonetics*, vol. 14, pp. 445–456, 1986.

[3] S. B. Davis, "Acoustic characteristics of laryngeal pathology," in *Speech Evaluation in Medicine*, J. Darby, Ed., pp. 77–104, Grune and Stratton, New York, NY, USA, 1981.

[4] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 300–313, 1998.

[5] O. Fujimura and M. Hirano, *Vocal Fold Physiology-Voice Quality Control*, Singular, San Diego, Calif, USA, 1995.

[6] R. J. Baken and R. F. Orlikoff, *Clinical Measurements of Speech and Voice*, Singular Thomson Learning, San Diego, Calif, USA, 2000.

[7] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, AITBS, New Delhi, India, 1995.

[8] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 373–383, 1996.

[9] D. G. Childers, "Signal processing methods for the assessment of vocal disorders," *The Journal of Biomedical Engineering Society of India*, vol. 13, pp. 117–130, 1994.

[10] N. B. Pinto and I. R. Titze, "Unification of perturbation measures in speech signals," *The Journal of the Acoustical Society of America*, vol. 87, no. 3, pp. 1278–1289, 1990.

[11] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics to noise ratio as an index of the degree of hoarseness," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[12] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1329–1334, 1986.

[13] C. Manfredi, "Adaptive noise energy estimation in pathological speech signals," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 11, pp. 1538–1543, 2000.

[14] M. de Oliveira Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 1, pp. 96–104, 2000.

[15] F. Plant, H. Kessler, B. Cheetham, and J. Earis, "Speech monitoring of infective laryngitis," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96)*, vol. 2, pp. 749–752, Philadelphia, Pa, USA, October 1996.

[16] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal to noise excitation ratio-a new measure for describing pathological voices," *Acustica - Acta Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[17] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.
[18] Anantha krishna, K. Shama, and U. C. Niranjan, "*k*-Means nearest neighbor classifier for voice pathology," in *Proceedings of IEEE India Annual Conference (INDICON '04)*, pp. 232–234, IIT-Kharagpur, India, December 2004.
[19] E. Zwicker and H. Fastl, *Psycho-Acoustics: Facts and Models*, Springer, Berlin, Germany, 1999.
[20] Kay Elemetrics Corp, Disordered Voice Database Model 4337, Version 1.03, Massachusetts Eye and Ear Infirmary Voice and Speech Lab, 2002.
[21] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.
[22] C. Wendt and A. Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS '96)*, vol. 2, pp. 45–48, Atlanta, Ga, USA, May 1996.
[23] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, 1992.
[24] T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2002.
[25] S. H. Chen and J. F. Wang, "Noise-robust pitch detection method using wavelet transform with aliasing compensation," *IEE Proceedings*, vol. 149, no. 6, pp. 327–334, 2002.
[26] A. Papoulis, *Signal Analysis*, McGraw-Hill, New York, NY, USA, Int. edition, 1984.
[27] G. K. Parikh and P. C. Loizou, "The effects of noise on the spectrum of speech," a M.S. thesis presented to the faculty of Telecommunication Engineering, University of Texas at Dallas, August 2002.
[28] W. A. Yost, *Fundamentals of Hearing*, Academic Press, New York, NY, USA, 3rd edition, 1994.
[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Analysis*, John Wiley & Sons, New York, NY, USA, 2002.
[30] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases," *IEEE Engineering in Medicine and Biology Magazine*, vol. 16, no. 4, pp. 74–82, 1997.
[31] J. B. Alonso, J. de Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 4, pp. 275–284, 2001.
[32] J. I. Godino-Llorente and P. Gomez-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.
[33] K. Umapathi, 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.
[34] D. R. Boone, *The Voice and Voice Therapy*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.
[35] J. A. Koufman and P. D. Blalock, "Functional voice disorders," in *Oto Laryngological Clinics of North America. Voice Disorders*, vol. 24, no. 5, pp. 1059–1073, Philadelphia, Pa, USA, October 1991.

Kumara Shama was born in 1965 in Mangalore, India. He received the B.E. degree in 1987 in electronic and communication engineering and M.Tech. degree in 1992 in digital electronics and advanced communication, both from Mangalore University, India. Since 1987 he has been with Manipal Institute of Technology, MAHE, Manipal, India, where he is currently a Reader in the Department of Electronics and Communication Engineering and also pursuing his Ph.D. thesis in speech processing application in medicine.

Anantha krishna was born in 1976 in Kasaragod, India. He received the M.S. degree in 1998 in electronic science from Mangalore University, India, and M.Tech. degree in 2004 in computer cognition technology, from Mysore University, India. He was a Lecturer at Mangalore University from 1998 to 2002. Since 2004 he is with Manipal Institute of Technology, MAHE, Manipal, India, where he is currently a Lecturer in the Department of Electronics and Communication Engineering.

Niranjan U. Cholayya was born in 1964 in Sholapur, India. He received the Ph.D. degree in electrical science from Indian Institute of Science, Bangalore, India in 1993. He has been working with Manipal Institute of Technology, MAHE, Manipal, India, where he is currently an Adjunct Professor in Biomedical Engineering Department. He is a Senior Member of IEEE and past Secretary of Biomedical Engineering Society of India. His research interests are signal and image processing applications in medicine.