LETTER

Pathological Voice Detection Using Efficient Combination of Heterogeneous Features

Ji-Yeoun LEE[†a)], ***Member***, **Sangbae JEONG**[†], ***and*** **Minsoo HAHN**[†], ***Nonmembers***

SUMMARY Combination of mutually complementary features is necessary to cope with various changes in pattern classification between normal and pathological voices. This paper proposes a method to improve pathological/normal voice classification performance by combining heterogeneous features. Different combinations of auditory-based and higher-order features are investigated. Their performances are measured by Gaussian mixture models (GMMs), linear discriminant analysis (LDA), and a classification and regression tree (CART) method. The proposed classification method by using the CART analysis is shown to be an effective method for pathological voice detection, with a 92.7% classification performance rate. This is a noticeable improvement of 54.32% compared to the MFCC-based GMM algorithm in terms of error reduction.

key words: *pathological voice detection, heterogeneous feature combination, mel-frequency filter bank energies, higher-order statistics, pattern classification algorithm*

1. Introduction

Over the past few years, a considerable number of studies have focused on the extraction of acoustic parameters for the objective and automatic judgment of pathological voices. Many acoustic features have been utilized in the time and frequency domains. Among acoustic parameters, the most important ones are pitch, jitter, shimmer, harmonics-to-noise ratio (HNR), and the normalized noise energy (NNE). These parameters are based on the fundamental frequency. Correlations between these parameters and pathological voice detection have been demonstrated [1]. However, it is not easy to estimate the fundamental frequency in a pathological voice correctly.

Recently, the emergence of attractive pattern classification algorithms such as Gaussian mixture models (GMMs), neural networks (NN), and vector quantization (VQ) has received greater attention [2]–[4]. GMMs have been reported as a very successful classification method for pathological detection [3]. In this paper, GMMs, linear discriminant analysis (LDA), and classification and regression tree (CART) analysis are used to combine heterogeneous features and to measure their performances. Then, we propose a new feature combination method using the features from the completely different domains to improve the pathological/normal voice classification performance. That is, the main issue is how the heterogeneous inputs can be combined to show good performance in pathological/normal voice classification. We use auditory-based features of the speech signal, e.g. mel-frequency filter bank energies and mel frequency cepstrum coefficients (MFCCs). Higher-order statistics (HOS) coefficients are also used as time-domain features. The HOS methods to repress Gaussian noises and maintain some of the non-Gaussian information may be more valuable for the discriminant modeling between normal and pathological voices [5], [6].

2. HOS Analysis

A voiced speech signal, $x(k)$, can be expressed as (1).

$$x(k) = s(k) + w(k) \tag{1}$$

where $s(k)$ is a non-Gaussian signal generated by the vocal folds oscillation and $w(k)$ is an additive noise that cannot be ignorable for pathological voices.

$s(k)$ of pathological voices can be characterized by the large pitch period variation. It occurs because the vocal folds movement is not balanced and an incomplete closure may appear in glottal cycles. It also explains energy increases for high-frequency noisy components due to the aerial turbulence [6]. The degree of hoarseness caused by increased high-frequency noises can be modeled by $w(k)$. On the other hand, $s(k)$ of normal voices is rather periodic and stable. It is comparatively easy to estimate the pitch-related statistics. After all, the variation difference of the statistics extracted from $s(k)$ can be an important cue for the pathological and normal voice classification [3], [6].

Generally, the random noises, $w(k)$, can be modeled as a Gaussian distribution, while $s(k)$ generated in a sustained vowel sound /ah/ used in this paper can be modeled as a non-Gaussian. The HOS analysis is well-known that it can estimate the non-Gaussian statistics rather successfully in a random process [6]. Namely, when it is applied to pathological voices, the unstable and discontinuous components of $s(k)$ can be easily estimated because the HOS analysis can be blind to Gaussian noises.

Among various HOS statistics, the 3rd- and the 4th-order cumulants called the normalized skewness, γ_3, and the normalized kurtosis, γ_4, are widely used as characteristic parameters. They can be defined as (2) [5].

$$\gamma_3 = \frac{\sum_{n=1}^{N}(x_n - \mu)^3}{(N-1)\sigma^3}, \quad \gamma_4 = \frac{\sum_{n=1}^{N}(x_n - \mu)^4}{(N-1)\sigma^4} \tag{2}$$

Manuscript received July 11, 2007.
Manuscript revised October 9, 2007.

[†]The authors are with Information and Commun. Univ., 119 Munjiro, Yuseong-gu, Daejeon, 305–732, Korea.
a) E-mail: jyle278@icu.ac.kr

DOI: 10.1093/ietisy/e91–d.2.367

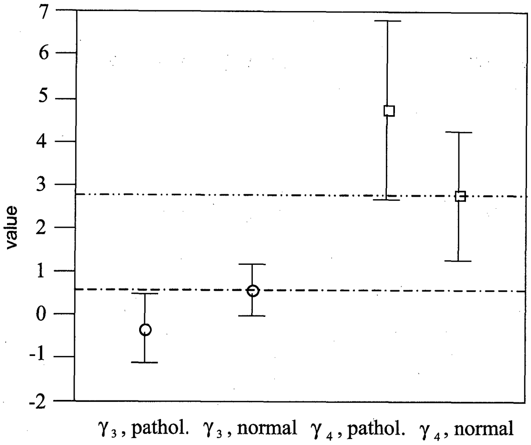

Fig. 1 Distributions of γ_3 and γ_4. (circle: mean of γ_3, square: mean of γ_4, vertical lines: standard deviations of γ_3 and γ_4)

where x_n is the n^{th} sample value and N is the number of the samples while μ and σ represent the mean and the standard deviation, respectively.

Figure 1 shows the distributions of γ_3 and γ_4 extracted from raw pathological and normal voices distributed by Kay Elemetrics. In the distributions of γ_3, pathological speakers tend to be skewed left while normal speakers, right. Therefore, the sign information of γ_3 is important for our study. In the distributions of γ_4, pathological voices can be considered to have a leptokurtic distribution ($\gamma_4 > 3$) and normal voices, a platykurtic ($\gamma_4 < 3$). And pathological voices obviously tend to show larger variations than normal voices in γ_3 and γ_4 distributions. Based on the above observations, we might insist that the HOS analysis is more appropriate for discerning the signals characterized by an irregularity of the speech production mechanism.

3. Proposed Feature Combination Methods

3.1 LDA-Based Methods

The LDA has been sucessfully used in many applications, such as data classification and compression. It aims at finding the optimal transformation by simultaneouly minimizing the within-class distance and maximizing the between-class distance. Thus, it achieves maximum discrimination among the feature vector classes. The optimal trasformation can be readily estimated by computing the eigen decomposition on the scatter matrices [2]. To improve the performance, an important issue is how to transform or project the raw feature data into a new feature space in which the classes are easier to distinguish and a more robust decision can be found. When the LDA is applied as a pre-processing step for pathological voice detection, it may be possible to find the optimal transformation to discriminate between pathological and normal voices. Therefore, it is hoped that the LDA will provide better performance than the MFCC-based GMM algorithm.

In this approach, the different acoustic features in the frequency and time domains are combined by the LDA-based method. Two kinds of LDA-based methods are suggested to investigate classification performance. The first one utilizes the feature vectors consisting of the mel-frequency filter bank energies, γ_3, and γ_4 extracted by each analysis frame as (3). This is designated as the frame-based LDA method.

$$\vec{g} = [FBE_1, FBE_2, FBE_3, \ldots, FBE_P, \gamma_3, \gamma_4] \qquad (3)$$

where FBE_k is the k^{th} mel-frequency filter bank energies and P is the total number of FBEs.

In our experiments, the number of the mel-frequency filter bank energies is varied from 22 to 42 in order to find the optimal dimension. Therefore, the dimension of $\vec{g}$ ranges from 24 to 44. Then, it is reduced to the 12th-order feature vector by the LDA transformation. Finally, the transformed feature vector is used to train GMMs by the expectation-maximization (EM) algorithm. In testing, the log-likelihood ratio evaluated by the LDA-transformed feature vector and the GMMs are used to classify normal and pathological voices.

The second method utilizes the GMM log-likelihoods, γ_3, and γ_4 extracted and averaged in each sentence as (4). This is designated as the sentence-based LDA method. The 3rd-order feature vector is directly reduced to the 1st-order scalar value by the LDA transformation. The definition of the log-likelihood ratio can be found in [3].

$$g' = \left[\overline{LL}, \overline{\gamma_3}, \overline{\gamma_4}\right] \qquad (4)$$

where $\overline{LL}$ is the average log-likelihood of the MFCC-based GMM, $\overline{\gamma_3}$ is the average γ_3, and $\overline{\gamma_4}$ is the average γ_4.

3.2 Sentence-Based CART Method

CART analysis is a common method for building statistical models founded on tree-based techniques. The analysis is powerful because it can deal with incomplete data, multiple types of features, and a decision tree that contains rules which are readable by humans [4]. Since some features enable good decisions in each frame or sentence and some do not for pathological voice detection, it is necessary to design a rule to make the final decision use the multiple inputs in the classifiers at the same time.

In this method, the heterogeneous features are the log-likelihood estimated by the MFCC-based GMM algorithm, γ_3 , and γ_4 extracted and averaged by each sentence as (4). The decision tree of the CART analysis implements their combination for the experiments.

4. Experiments and Results

A disordered voice database produced by Kay Elemetrics was used in our experiments [7]. A subset of 53 normal and 600 pathological speakers was formed from the above database. The acoustic samples were the sustained phonation of the vowel /ah/. To maintain the balance of the number

Table 1 Average EER ± CI (%) of MFCC-based GMM algorithm.

Number of filter banks	Number of mixtures				
	2	4	8	16	32
22	24.2±2.1	21.7±2.0	20.5±2.2	18.1±2.0	18.4±1.8
26	22.4±1.9	20.8±1.2	19.3±1.3	18.9±1.2	20.3±1.6
30	20.7±1.4	21.4±0.8	19.4±1.0	18.5±1.0	20.1±1.4
34	21.1±1.4	19.1±0.9	19.2±0.8	16.2±0.5	18.5±0.6
38	21.4±1.8	20.2±1.5	20.7±1.5	20.6±1.6	21.3±1.6
42	20.7±1.6	22.1±1.6	20.3±1.7	19.8±2.1	18.6±1.9

of speakers, the voice signals uttered by 547 normal Koreans were added after careful examination by a group of speech experts. The recording conditions were similar to those of the database distributed by Kay Elemetrics. Each utterance was down-sampled to 16 kHz. 70% and 30% of the data was used for training and testing, respectively. The speakers were randomly selected from the database to build each set for a 30-fold cross-validation scheme [3].

4.1 MFCC-Based GMM Algorithm

The GMMs with 2, 4, 8, 16, and 32 mixtures are trained by the EM algorithm after the estimation of their initial parameters by the Linde-Buso-Gray (LBG) algorithm. The order of the extracted MFCCs is the 12th-order. Only the static vectors are utilized, because the temporal derivatives of the MFCCs have little discriminating ability compared to the MFCCs [3].

Table 1 shows the average equal error rates (EERs) and the 95% confidence intervals (CIs) of the MFCC-based GMM algorithm according to the number of the Gaussian mixtures and the number of the mel frequency-based filter bank energies. The definition of a 95% CI can be found in [3]. The CIs can be used to describe how reliable the average EERs are. In Table 1, although the performances along with the dimension reduction of filter banks are fairly similar, it can be said that a higher number of mixtures tends to improve the performance. On the other hand, an increase in the number of the mel-frequency filter bank energies has little effect. When the number of Gaussian mixtures is 16 and the DCT changes the 34 dimensional vectors of the mel-frequency filter bank energies into 12 dimensional MFCC vectors, the best performance of the average EERs, 16.2%, is obtained. The receiver operating characteristic (ROC) curve is shown in Fig. 2.

4.2 LDA-Based Methods

For the GMMs and the MFCCs, the same scheme to that of the MFCC-based GMM algorithm is applied to the frame-based LDA method. In Table 2, the performance of the frame-based LDA method is shown. In comparison with that of the MFCC-based GMM algorithm, it shows fairly similar characteristics. When the number of Gaussian mixtures is 16 and the LDA transformation changes the 36 dimensional vectors of the mel-frequency filter bank energies into 12 dimensional vectors, the best EER performance is 14.2%. Then, the ROC curve is shown in Fig. 2.

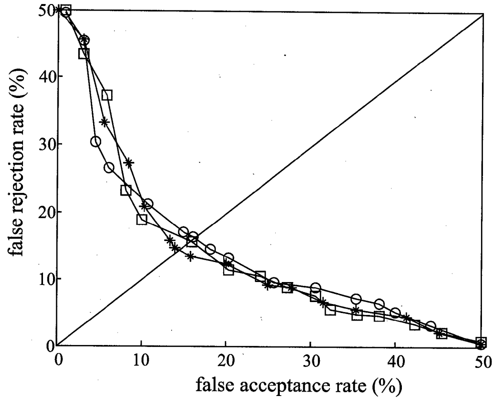

Fig. 2 Comparative ROC curves. (–□–: sentence-based LDA, – ∗ –: frame-based LDA, – ○ –: MFCC-based GMM)

Table 2 Average EER ± CI (%) of frame-based LDA method.

Number of filter banks + HOS	Number of mixtures				
	2	4	8	16	32
24	20.2±1.7	18.7±1.5	18.1±1.5	16.4±1.4	17.7±1.5
28	18.6±1.4	17.9±1.6	18.5±1.5	17.7±1.3	16.4±1.4
32	20.1±1.7	19.4±1.4	19.4±1.4	15.2±1.3	17.4±1.5
36	18.2±1.6	17.6±1.4	15.3±1.5	14.2±1.4	15.2±1.5
40	18.9±1.6	19.2±1.6	17.9±1.4	16.9±1.6	16.2±1.6
44	20.4±1.5	19.5±1.5	18.1±1.4	16.7±1.5	17.8±1.5

In the sentence-based LDA method, the log-likelihoods used in the training process are decided in the conditions that show best performance for the MFCC-based GMM algorithm. The average EER performance is 15.8%. The ROC curve is shown in Fig. 2. In conclusion, the frame-based LDA method outperforms both the MFCC-based GMM algorithm and the sentence-based LDA method. It can be said that the proposed LDA-based methods are more effective for pathological voice detection than the conventional MFCC-based GMM algorithm.

4.3 Sentence-Based CART Method

The optimal decision tree formed by the log-likelihood ratio estimated from the MFCC-based GMM algorithm, γ_3, and γ_4 is shown in Fig. 3. We confirm that the HOS statistics characteristics shown in Fig. 1 are reflected at each tree node. And normal voices tend to have negative log-likelihood ratios while pathological voices, positive ones. The average performance is 92.7% in distinguishing pathological and normal voices. Table 3 shows a brief summary of results for all methods. We can confirm that the proposed feature combination methods outperform the MFCC-based GMM algorithm. The best performance, 92.7%, is obtained when the sentence-based CART analysis combines the heterogeneous inputs. It is obvious that the effective combination of the heterogeneous features shows better performance than using the homogeneous feature by about 2 to 10%.

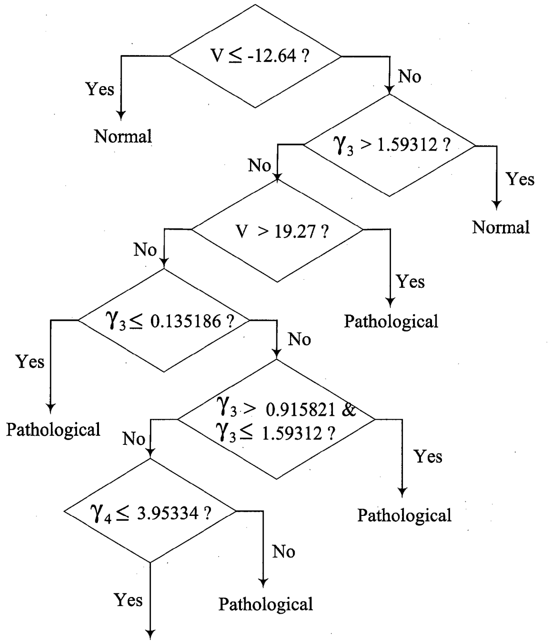

Fig. 3 Decision tree obtained by sentence-based CART analysis. (V: log-likelihood ratio of MFCC-based GMM, γ_3: normalized skewness, γ_4: normalized kurtosis)

Table 3 Comparative performance evaluation.

		Average performance ± CI (%)
Baseline algorithm	MFCC-based GMM	83.8 ± 0.5
Proposed algorithm	frame-based LDA	85.8 ± 1.4
	sentence-based LDA	84.2 ± 1.3
	CART-based	92.7 ± 0.4

5. Conclusion

This paper presents a novel method to combine the various feature representations for pathological voice detection. In several studies, many feature parameters have shown the statistically significant differences between the pathological and normal voices [1], [3], [5], [6]. It can be necessary to measure the performance through combination of heterogeneous features. Therefore, in this study we have introduced some effective combinations of heterogeneous features. The performance of the feature combinations has been measured in the MFCC-based GMM, LDA-based methods, and the CART-based classifier. The experiments have demonstrated that the CART-based method can provide the highest classification performance, at 92.7%. The amount of the improvement is 54.32% compared to the conventional MFCC-based GMM algorithm from the aspect of error reduction. This is very important, since CART analysis has been shown to be more appropriate for combining heterogeneous features. CART analysis might also be usefully applied to complement existing pathological voice detection methods in the clinic. Our future work may include the application and the analysis of our technique in real environments and the study of the pathological type classification.

References

[1] D. Michaelis, M. Forhlich, and H.W. Strobe, "Selection and combination of acoustic features for the description of pathological voices," J. Acoust. Soc. Am., vol.103, no.3, pp.1628–1639, 1998.

[2] T. Xiong and V. Cherkassky, "A combined SVM and LDA approach for classification," Proc. IEEE IJCNN, vol.3, pp.1455–1459, 2005.

[3] J.I. Godino-Llorente, S. Aguilera-Navarro, and P. Gomez-Vilda, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," IEEE Trans. Biomed. Eng., vol.53, no.10, pp.1943–1953, 2006.

[4] M.M. Tanabian and P. Tierney, "Automatic speaker recognition with formant trajectory tracking using CART and neural networks," Canadian Conference on ECE, pp.1225–1228, 2005.

[5] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," IEEE Trans. Speech Audio Process., vol.9, no.3, pp.217–231, 2001.

[6] J.B. Alonso, J. de Leon, I. Alonso, and M.A. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," EURASIP J. Applied Signal Processing, vol.4, pp.275–284, 2001.

[7] Kay Elemetrics Corp., "Disordered voice database," Ver. 1.03, 1994.