Detection of Pathological Voice Using Linear Discriminant Analysis

Ji-Yeoun Lee(ICU), SangBae Jeong(ICU),
Hong-Shik Choi (Yonsei Univ.), Minsoo Hahn(ICU)

<차 례>

<Abstract>

Detection of Pathological Voice Using Linear Discriminant Analysis

Ji-Yeoun Lee, SangBae Jeong, Hong-Shik Choi, Minsoo Hahn

Nowadays, mel-frequency cesptral coefficients (MFCCs) and Gaussian mixture models (GMMs) are used for the pathological voice detection. This paper suggests a method to improve the performance of the pathological/normal voice classification based on the MFCC-based GMM. We analyze the characteristics of the mel frequency-based filterbank energies using the fisher discriminant ratio (FDR). And the feature vectors through the linear discriminant analysis (LDA) transformation of the filterbank energies (FBE) and the MFCCs are implemented. An accuracy is measured by the GMM classifier. This paper shows that the FBE LDA-based GMM is a sufficiently distinct method for the pathological/normal voice classification, with a 96.6% classification performance rate. The proposed method shows better performance than the MFCC-based GMM with noticeable improvement of 54.05% in terms of error reduction.

* Keywords: Pathological voice detection, Gaussian mixture model, Linear discriminant analysis.

1. Introduction

Nowadays, people are much interested in vocal health. When speech impediment happens to the social life, they keenly realize how much the speech is important as a means of communication. Therefore, the researches to objectively detect pathological voices without professional doctors and medical instruments have been done in the biomedical engineering. Especially, acoustic analysis has been known to be an effective tool for objective measurement of the degree deviation between pathological and normal voice patterns. Using this tool, we can establish an objective evaluation before an application of medical treatments [1]-[9].

Many algorithms to calculate the acoustic parameters for the objective judgment of pathological voice have been developed. Among the acoustic parameters, the important parameters are the pitch, the jitter, the shimmer, the harmonics to noise ratio (HNR), and the normalized noise energy (NNE), etc. Enough correlations between the parameters and the pathological voice detection have been demonstrated by [1]-[3] based on the fundamental frequency. However, it is not easy to correctly estimate the fundamental frequency in pathological voices.

In the recent years, pattern classification algorithms such as the Gaussian mixture model (GMM), the neural networks (NNs), the vector quantization (VQ) and the characteristic parameter such as the mel frequency cesptral coefficients (MFCCs) become more popular for the voice damage detection [6]-[8]. Especially, the GMM and the MFCCs become generally accepted as the most useful methods for the detection of voice impairments as in [9]. Therefore, in this paper, we regard the MFCC-based GMM method as baseline algorithm and propose an efficient method to detect pathological voices in terms of performance improvement. In the first place, our study examines the effectiveness of the mel frequency-based filterbank energies as fundamental parameters of the feature extraction using the fisher discriminant ratio (FDR). And then performance of the MFCC-based GMM algorithm for the construction of the baseline system is measured. Finally, a new approach with the linear discriminant analysis (LDA) transformation of the filterbank energies (FBE) is suggested. Our experiments and analyses verify the effectiveness of the parameters extracted from the FBE-LDA transformation compared to the MFCCs.

This paper is organized as follows. Chapter 2 shows the previous works to detect the voice impairments. Chapter 3 analyzes the mel frequency-based filterbank energies. Chapter 4 describes an effective method such as the LDA for the performance improvement. The proposed procedure, the experiments, and improved results are explained to Chapter 5. Finally, Chapter 6 is for conclusion.

2. Previous Works

A large number of works have been reported in automatic detection and classification of pathological voices in terms of acoustic analyses, parametric and non-parametric feature extractions, pattern classification methods and statistical methods [1]-[9].

Hadjitodorov et al. [10] described a system to use the acoustic analysis of pathological voices. Based on the glottal cycles measured by a cross-correlation detector, the shimmer, the jitter, the harmonics-to-noise ratios and other widely used acoustic parameters are calculated. Classification was achieved by the LDA and the NN clustering. They used 53 normal and 638 pathological voices directly collected for test experiments. The system accuracy was 96.1%. Despite such a high level of accuracy, the reliability of the study may be questioned due to the limited experimental material information presented, such as the types of the diseases and the recording conditions and cross validation experiment not to implement.

Dibazar et al. [11] presented the best results with database distributed by Kay Elemetrics in the pathological voice detection. They used the parameters extracted by a multi-dimensional voice program (MDVP), MFCCs, and measures of pitch dynamics. They presented a best accuracy of 98.3% with a hidden markov model (HMM) classifier. However, in their paper, it is not easy to find the details how their experiments are carried out.

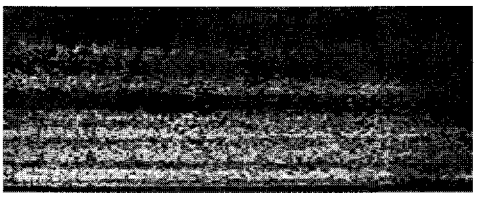
Godino et al. [9] reported several papers using database distributed by Kay Elemetrics. They adopted the integrated methods of the GMMs and the MFCCs to detect pathological voices and used 200 speakers (53 normal and 147 pathological voices) in 2006. Results were presented with the confidence intervals and the confusion matrices. A best accuracy of 94.07% was obtained by the cross validation scheme. However, they utilized rather a small size database and certain kinds of pathologies such as the hyperfunction and the A-P squeezing.

Saenz et al. [12] presented an overview of the previous work using database of Kay Elemetrics and other ones. The intention of this paper is to compare the efficiency to be made with previous approaches. They described the methodological requirements that should be satisfied to allow comparisons with other systems when the system is designed for pathological voice detection .

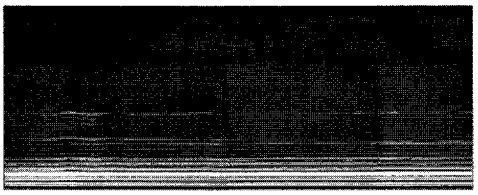
3. Effectiveness of Mel Frequency-based Filterbank Energies

<Figure 1> shows spectrograms of a typical /ah/ sound uttered by a pathological and a normal speaker. The pathological voice in <Figure 1>(a) exhibits aperiodic and a noise-like broadband spectrum. This is generally because the movement of the vocal folds is not

balanced and an incomplete closure may appear in glottal cycles. On the other hand, most normal voices have periodic and discrete peaks in spectrum as shown in <Figure 1>(b). Since they are produced without trauma to the vocal folds and larynx, they have good voice quality and sound more pleasant [3][5]. After all, the spectrogram, which shows distinct difference as shown in <Figure 1>, can be an important clue for the classification of pathological and normal voices. The mel frequency-based filterbank energies can express the characteristic of spectrogram in frequency domain [9].



time (msec.)
(a) pathological voice



time (msec.)
(b) normal voice

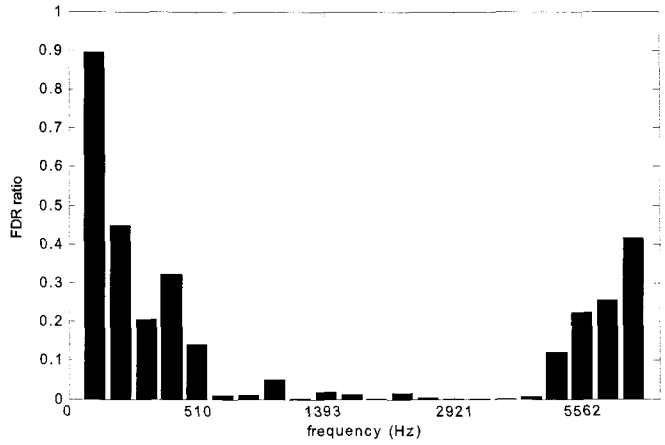
<Figure 1> Spectrograms of typical /ah/ sound

In this paper, mel frequency-based filterbank energies are used as the fundamental parameters for the feature extraction in pathological voices. This part demonstrates the effectiveness of the mel frequency-based filterbank energies with the analysis of the FDR. The FDR has been widely used as a class separability criterion and the standard for the feature selection in speaker recognition applications [9]. It is defined as in (1).

$$F_i = \frac{(\mu_{iC} - \mu_{i\bar{C}})^2}{\sigma_{iC}^2 + \sigma_{i\bar{C}}^2} \tag{1}$$

where μ, σ^2, C, $\bar{C}$ represent class mean, class variance, the normal voice, and the pathologic al voice, respectively.

This ratio selects the features which maximize a scatter between the classes. The higher the value of F_i, the more important the feature is. It means that the feature i has a low



<Figure 2> FDR plot using the 22nd mel frequency-based filterbank energies

variance in regard of the inter-class variance and the feature is suitable to discriminate the classes. <Figure 2> shows the FDR of the mel frequency-based filterbank energies with the 22nd dimension according to the frequency. The usefulness of the mel frequency-based filterbank energies to classify pathological and normal voices is found in comparatively low and high frequency bands. The largest value indicating the 1st formant appears in the low frequency band below 700 Hz. It shows that the 1st formant is the important feature to distinguish pathological voice from normal one. Also the high frequency band above 5 kHz can be used as a discriminant feature. It shows the tendency that noises increase at the high frequency band due to an inefficient movement of the vocal folds [13]. Those results suggest that the 1st formant and the high frequency noise are the important information to classify pathological and normal voices. Finally, this mel frequency-based filterbank energies are converted back to the MFCCs using the discrete cosine transform (DCT) [14]. On the other hand, they are transformed into discriminant feature vectors through the FBE-LDA transformation. Through the FDR analysis, we can confirm that the use of the mel frequency-based filterbank energies is suitable for our purpose.

4. Linear Discriminant Analysis

The LDA aims at finding the best combination of classes to improve the discrimination among the feature vector classes. It is implemented by the transformation matrix of the feature vector classes. The transformation is defined as the method to maximize between-class

separability and to minimize within-class variability [6][15].

$$W = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{n_k}(x_{kn}-\mu_k)(x_{kn}-\mu_k)^t \quad (2)$$

$$B = \frac{1}{N}\sum_{k=1}^{K} n_k(\mu_k-\mu)(\mu_k-\mu)^t \quad (3)$$

where

W : within-class covariance matrix,

B : between -class covariance matrix,

N : the total number of training patterns,

K : the number of classes,

n_k : the number of training patterns of the k^{th} class

μ_k : the mean of the k^{th} class, μ : the overall mean.

The transformation matrix is formed by the eigenvectors corresponding to the predominant eig envalues, the largest eigenvalues of the matrix $W^{-1}B$, in the classes.

The LDA can be implemented in two forms: class-independent and class-dependent transformations. The class-independent method maximizes the ratio of the class covariances across all classes simultaneously. This defines the single transformation matrix in K classes. The class-dependent method implemented in this paper maximizes the ratio of the class covariances for each class separately. This forms the K transformation matrixes, each corresponding to one class [6]. However, in case of two classes, class-independent and class-dependent methods are essentially the same.

5. Experiments and Results

5.1. Database

A disordered voice database distributed by Kay Elemetrics was used in our experiments. It included 53 normal and 657 pathological speakers with a wide variety of organic, neurological, traumatic, and psychogenic voice disorders. All of the recordings and clinical information related to the disordered voice samples were structured to CD-ROM in detail [16]. Since we were only interested in pathologies which affect the vocal folds, the experiment was

carried out for the sustained vowel /ah/ phonation (1-3 sec.). All voice data were down-sampled to 16 kHz. 70% and 30% of the data were used separately for training and test sets in the MFCC-based GMM and the LDA experiments. Then, the 30-fold cross-validation scheme was used to estimate the classifier performance [9]. The ratio of gender and age was randomly selected from the database to build each set for the experiments.

5.2. Overall block diagram

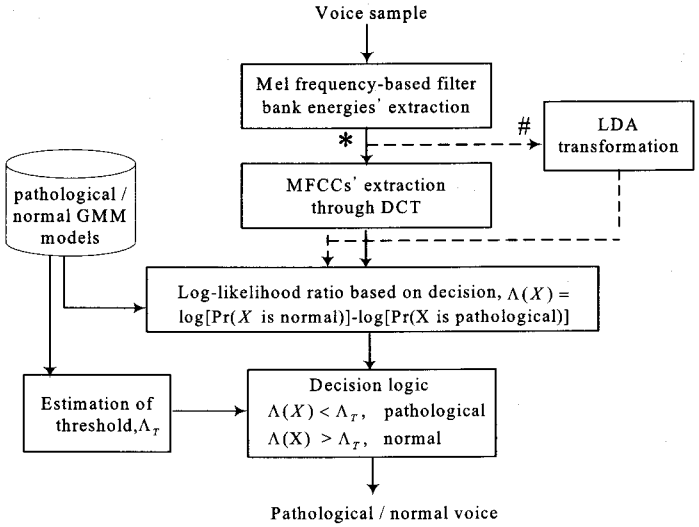
<Figure 3> presents the overall block diagram of our pathological/normal voice classification procedure. Firstly, the mel frequency-based filterbank energies are estimated. They are the important baseline feature vectors utilized in our procedure. Their analyses are implemented by two methods of the feature extraction to compare the performances: (*) MFCCs extraction through the DCT and (#) extraction of the feature vectors through the FBE-LDA transformation. In training process, Gaussian models of the pathological/normal voices are trained with an expectation- maximization (EM) algorithm to determine the model parameters such as mean vectors, covariance matrices and mixture weights in advance. And then, the log-likelihood ratio is estimated as a threshold, Λ_T, and the equal error rate (EER) is applied to evaluate the performance of GMMs in test procedure. In test process, the log-likelihood ratio, $\Lambda(X)$, estimated by the pre-trained GMMs parameters is compared with Λ_T. The voice is considered to be normal if $\Lambda(X) > \Lambda_T$, otherwise, pathological.

5.3. Baseline performance (MFCC-based GMM)

The GMM initialization is performed by the Linde-Buzo-Gray (LBG) algorithm [17]. Covariance matrices are diagonal. The GMMs are trained using 2, 4, 8, 16, and 32 mixtures. The MFCCs dimension as the feature vector is 12. It is obtained from the DCT with the mel frequency-based filterbank energies ranging from the 22^{nd} to the 42^{nd}. The static vectors are only used because the temporal derivatives of the MFCCs have no discriminant ability compared with the MFCCs [9]. <Table 1> shows the average EERs and 95% confidence intervals (CIs) according to the number of the Gaussian mixtures and the number of the mel frequency-based filterbank energies. The definition of a 95% CI can be defined as in (4) [9].

$$CI = \pm \frac{Z_p \sigma}{\sqrt{N}} \tag{4}$$

where Z_p is the value derived from the normal distribution (1.96 for a 95% CI), σ is the



<Figure 3> Overall classification procedure

populated standard deviation, and N is the sample size.

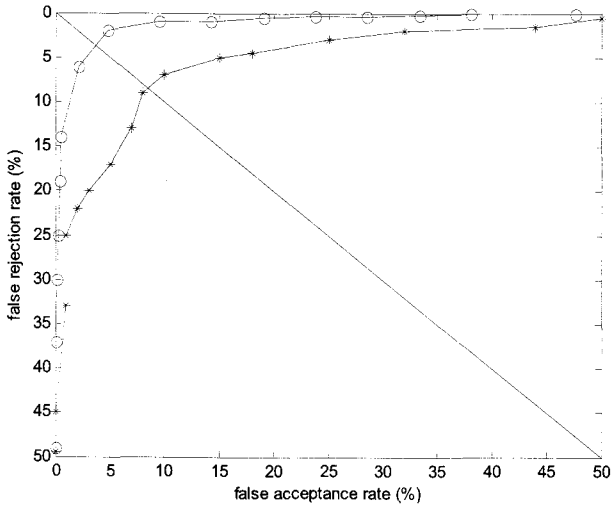
Although the performances along with the reduction of the dimension are fairly similar, it can be said that a larger number of mixtures tends to improve the performance. When the Gaussian mixtures are 16 and the DCT changes the 26th dimensional vector of the mel frequency-based filterbank energies into the 12nd MFCCs, the best performance of the classification between pathological and normal voice, 92.6%, is obtained. Then the ROC curve is shown in <Figure 4>.

<Table 1> Average EER±CI (%) of MFCC-based GMM

	Mixture 2	Mixture 4	Mixture 8	Mixture 16	Mixture 32
Filterbank 22	9.4±1.5	9.0±1.7	8.5±1.6	8.3±1.9	8.6±1.7
Filterbank 26	9.8±1.6	9.8±1.5	8.3±0.9	7.4±1.0	8.5±1.2
Filterbank 30	10.1±1.2	9.7±1.1	9.0±0.0	9.5±1.2	8.5±1.1
Filterbank 34	9.5±1.3	9.6±1.1	9.5±1.2	9.4±1.8	8.9±1.5
Filterbank 38	9.5±1.1	9.9±1.2	9.7±1.0	8.5±1.8	8.4±1.3
Filterbank 42	10.2±1.1	9.9±0.9	9.4±1.2	9.3±1.4	9.0±1.4

5.4. Performance of proposed FBE LDA-based GMM

For the GMMs experiments, the same scheme to that of the MFCC-based GMM algorithm is applied to the FBE LDA-based GMM method. <Table 2> shows the average



<Figure 4> Smoothed ROC curves. (—✱— : MFCC-based GMM, —○— : FBE LDA-based GMM)

EERs and CIs of the FBE LDA-based GMM method according to the number of the Gaussian mixtures and the number of the mel frequency-based filterbank energies. In comparison with that of the MFCC-based GMM algorithm, it shows fairly similar trends. When the number of Gaussian mixtures is 16 and the LDA transformation transform the 22 dimensional vectors of the mel-frequency filterbank energies into 12 dimensional vectors, the best EER performance is 3.4%. The ROC curve for the best performance is shown in Fig.4. In conclusion, the performance is approximately improved by 4.0% through the FBE-LDA method. It shows better performance than the MFCC-based GMM with noticeable improvement of 54.05% in terms of error reduction. It can be said that the proposed LDA-based method are more effective for the pathological voice detection than the conventional MFCC-based GMM algorithm.

<Table 2> Average EER±CI (%) of FBE LDA-based GMM

	Mixture 2	Mixture 4	Mixture 8	Mixture 16	Mixture 32
Filterbank 22	4.5±1.5	4.4±1.5	4.0±1.4	3.4±1.5	4.0±1.7
Filterbank 26	5.3±1.6	5.1±1.6	4.3±1.4	4.2±1.6	4.2±1.5
Filterbank 30	4.9±1.7	4.5±1.4	4.6±1.4	4.4±1.3	4.4±1.4
Filterbank 34	4.8±1.6	4.4±1.5	4.2±1.5	4.2±1.4	4.2±1.4
Filterbank 38	4.7±1.7	4.8±1.7	4.7±1.6	4.1±1.6	4.2±1.6
Filterbank 42	5.3±1.7	4.5±1.6	4.8±1.5	4.4±1.6	4.4±1.4

6. Conclusion

The objective of our study is to implement the FBE-LDA transformation to provide effectively discriminant feature vector classes for the pathological voice detection. And it is to compare the performances by utilizing the MFCCs and the FBE-LDA transformation with the mel frequency-based filterbank energies. We analyzed the mel frequency-based filterbank energies using the FDR and implemented the GMM detector with feature vectors through the DCT and the FBE-LDA transformation. Especially, there is a strong correlation between pathological voice detection and the GMM method through the FBE-LDA approach. The best performance is 96.6% when the filterbank of the 22^{nd} dimension is reduced to the 12^{nd} dimension through the FBE-LDA transformation and the number of mixtures is 16. Then, average EER and CI are obtained from the 30-fold cross-validation scheme. The proposed FBE-LDA method outperforms the well-known MFCC-based GMM method. The amount of the improvement is 54.05% in an error reduction sense.

The future works may include the application of our technique in real environments and the study in the pathological type classification.

References

[1] D. Michaelis, M. Frohlich, H. W. Strobe, "Selection and combination of acoustic features for the description of pathological voices", *Journal of the Acoustical Society of America*, Vol. 103, No. 3, pp. 1628-1639, 1998.

[2] Y. Qi, R. E. Hillman, C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices", *Journal of the Acoustical Society of America,* Vol. 105, No. 4, pp. 2532-2535, 1999.

[3] M. N. Vieira, "On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures", *Journal of the Acoustical Society of America,* Vol. 111, No. 2, pp. 1045-1055, 2002.

[4] J. H. L. Hansen, L. Gavidia-Ceballos, 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] D. G. Childers, K. S. Bae, "Detection of laryngeal function using speech and electroglottographic data", *IEEE Transactions on Biomedical Engineering*, Vol. 39, No. 1, pp. 19-25, 1992.

[6] T. Xiong, V. Cherkassky, "A combined SVM and LDA approach for classification", *Proc. IJCNN*, Vol. 3, pp. 1455-1459, 2005.

[7] M. M. Tanabian, P. Tierney, B. Z. Azami, "Automatic speaker recognition with formant Trajectory tracking using CART and neural networks", *Proc. Canadian Conference on ECE*, pp. 1225-1228, 2005.

[8] M. Rosa, J. C. Pereira, M. Grellet, "Adaptive estimation of residual signal for voice pathology diagnosis", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 5, pp. 96-104, 2000.

[9] J. I. Godino-Llorente, S. Aguilera-Navarro, P. Gomez-Vilda, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters", *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 10, pp. 1943-1953, 2006.

[10] S. Hadjitodorov, P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal disease screening", *Medical Engineering and Physics*, Vol. 24, No. 6, pp. 419-429, 2002.

[11] A. A. Dibazar, S. Narayanan, T. W. Berfer, "Feature analysis for automatic detection of pathological speech", *Proc. IEEE EMBS/BMES*, Vol. 1, pp. 182-183, 2002.

[12] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, P. Gomez-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.

[13] R. D. Kent, M. J. Ball, *Voice Quality Measurement*, Singular Thomson Learning, 1999.

[14] K. I. Molla, K. Hirose, "On the effectiveness of MFCCs and their statistical distribution properties in speaker identification", *Proc. IEEE Symposium on VECIMS*, pp. 136-141, 2004.

[15] S. Olivier, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition", *Proc. ICASSP*, Vol 1, pp. 125-128, 1995.

[16] Kay Elemetrics Corp, *Disordered voice database, ver.1.03*, 1994.

[17] Y. Linde, A. Buzo, R. Gray, "An algorithm for vector quantizer design", *IEEE Transaction on Communications*, Vol. 28, No. 1, pp. 84-94, 1980.

접수일자: 2007년 11월 9일
게재결정: 2007년 12월 13일

▶ Ji-Yeoun Lee : Corresponding author
Address: Information and Communications University, 119, Munjiro, Yuseong-gu, Daejeon, 305-732, Korea
Affiliation: Speech and Audio Information Lab.
Telephone: +82-42-866-6196
E-mail: jyle278@icu.ac.kr

▶ SangBae Jung
Address: Information and Communications University, 119, Munjiro, Yuseong-gu, Daejeon, 305-732, Korea
Affiliation: Speech and Audio Information Lab.
Telephone: +82-42-866-6196
E-mail: sangbae@icu.ac.kr

▶ Hong-Shik Choi
Address: Yongdong Severance Hospital, Yonsei University, College of Medicine, Seoul, Korea
Affiliation: Department of Otorhinolaryngology, Institute of Logopedics and Phoniatrics
Telephone: +82-42-866-6196
E-mail: hschoi@yumc.yonsei.ac.kr

▶ Minsoo Hahn
Address: Information and Communications University, 119, Munjiro, Yuseong-gu, Daejeon, 305-732, Korea
Affiliation: Speech and Audio Information Lab.
Telephone: +82-42-866-6123
E-mail: sangbae@icu.ac.kr