LETTER

Objective Pathological Voice Quality Assessment Based on HOS Features

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

SUMMARY This work proposes new features to improve the pathological voice quality classification performance. They are the means, the variances, and the perturbations of the higher-order statistics (HOS) such as the skewness and the kurtosis. The HOS-based features show meaningful differences among normal, grade 1, grade 2, and grade 3 voices classified in the GRBAS scale. The jitter, the shimmer, the harmonic-to-noise ratio (HNR), and the variance of the short-time energy are utilized as the conventional features. The performances are measured by the classification and regression tree (CART) method. Specifically, the CART-based method by utilizing both the conventional features and the HOS-based ones shows its effectiveness in the pathological voice quality measurement, with the classification accuracy of 87.8%.

key words: *pathological voice quality assessment, higher-order statistics, classification and regression tree, GRBAS*

1. Introduction

The accurate assessment of the pathological voice quality has attracted great attentions in the field of speech disorders for many years, motivating a large amount of research based on acoustical, aerodynamic, and physiological measurements [1]–[3]. In particular, all the studies made on the objective analyses demonstrate the need for combining different measurements in order to cope with the so widely varying nature of the voice and to increase the analysis reliability [2], [3]. In [2], a discriminant analysis was performed to detect the correlation between the jury classification and the combinations of the features along with the G component of the 4-point GRBAS scale ranging from 0 for normal to 3 for severely disordered voices. The experimental results showed that a nonlinear combination of six features produced the best accuracy of 86% among published reports: vocal range, Lyapunov coefficient (LC), maximum phonatory time (MPT), signal-to-noise ratio (SNR), estimated subglottic pressure (ESGP), and fundamental frequency (F0).

The focus of this study is the proposal of new features for the performance improvement of the pathological voice quality assessment. New features are the means, the variances, and the perturbations of the skewness and the kurtosis based on the HOS analysis. Pathological voices are classified according to the G parameter of the GRBAS scale evaluated by the speech and language therapists (SALTs). Therefore the performance is evaluated by the classification and regression tree (CART) analysis.

2. HOS Analysis

Pathological voice, $x(n)$, can be expressed as follows [4].

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

where $s(n)$ is a non-Gaussian signal generated by vibration of the vocal folds and $w(n)$, Gaussian noise.

$s(n)$ can be characterized by the large variation in the pitch period and the pitch peak amplitude, breaks in pitch generation, presence of sub-harmonic components, and distortion of the pitch pulse shape during sustained vowel phonation. This is because the movement of the vocal folds is unbalanced and an incomplete closure may appear in glottal cycles. These impact the overall degree of hoarseness represented as the G component of the GRBAS scale. So, as the G-based grade becomes higher, $s(n)$ of the voices tend to be more severely irregular, aperiodic, and unstable. And an increase of noisy components due to the aerial turbulence is modeled by $w(n)$. On the other hand, $s(n)$ of most normal voices are rather periodic and stable. They have good voice quality and sound more pleasant because they are produced without trauma to the vocal folds and larynx. In this case, $w(n)$ can be assumed to be zero [5].

Recently, the application of the HOS to speech processing has been mainly motivated by their inherent Gaussian suppression and phase preservation properties [4]. Works in this area are based on the assumption that speech has HOS properties that are different from those of Gaussian noises. Therefore, when the HOS analysis is applied to pathological voices, unstable and discontinuous statistics of $x(n)$ may be estimated because the HOS analysis is blind to Gaussian processes. In normal voices, the HOS of only non-Gaussian measurements may be extracted because $w(n)$ can be assumed to be zero [4].

Among various HOS statistics, the normalized skewness, γ_3, and the normalized kurtosis, γ_4, are widely used as the characteristic features. They can be defined as in (2) [4].

Manuscript received May 22, 2008.
Manuscript revised August 17, 2008.

[†] The authors are with Information and Commun. Univ., 119 Munjiro, Yuseong-gu, Daejeon, 305–732, Korea.

[††] The author is with the Department of Otorhinolaryngology, Institute of Logopedics and Phoniatrics, Yongdong Severance Hospital, Yonsei University College of Medicine, Seoul, Korea.

a) E-mail: jyle278@icu.ac.kr

DOI: 10.1093/ietisy/e91–d.12.2888

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

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

For the objective voice quality measurement, the proposed HOS-based features are the means and the variances, i.e., $\overline{\gamma}_3$, $\overline{\gamma}_4$, $\gamma_3^{(v)}$, and $\gamma_4^{(v)}$. They are all estimated for a sentence and in (3) and (4) the evaluation equations are presented.

$$\overline{\gamma}_3 = \frac{1}{T}\sum_{t=1}^{T}\gamma_{3,t}, \quad \overline{\gamma}_4 = \frac{1}{T}\sum_{t=1}^{T}\gamma_{4,t} \quad (3)$$

$$\gamma_3^{(v)} = \frac{1}{T-1}\sum_{t=1}^{T}(\gamma_{3,t}-\overline{\gamma}_3)^2, \quad \gamma_4^{(v)} = \frac{1}{T-1}\sum_{t=1}^{T}(\gamma_{4,t}-\overline{\gamma}_4)^2 \quad (4)$$

where t and T are the frame index and the number of frames, respectively.

To reflect the perturbation of the successive HOS features in pathological voice samples, the skewness perturbation (*SP*) and the kurtosis perturbation (*KP*) are proposed as in (5) and (6).

$$SP = \frac{1}{T-1}\sum_{t=1}^{T-1}(|\gamma_{3,t+1} - \gamma_{3,t}|), \quad (5)$$

$$KP = \frac{1}{T-1}\sum_{t=1}^{T-1}(|\gamma_{4,t+1} - \gamma_{4,t}|) \quad (6)$$

3. Experiments and Results

In 1981, the Japan Society of Logopedics and Phoniatrics distributed a DVD-ROM database of a total of approximately 65 utterances scored with the GRBAS scale. The pathological voices used in this paper were composed of 63 male and female voices of aged from 7 to 78 (mean: 45.7). We added 30 normal Korean voice data to this pathological data and finally, we have the 93 pathological-normal voice data. Among the 93 voices, 30 voices were Korean normal, 17 were associated with a voice of grade 1, 26 with a voice of grade 2, and 20 with a voice of grade 3. On the G-based grade of the GRBAS scale, a normal voice is 0 (G0), a slight, 1 (G1), a moderate, 2 (G2) and finally, a severe, 3 (G3). These perceptual grades were determined by the Japanese and the Korean SALTs. Since we were interested only in pathologies which affect the vocal folds, the experiment was carried out for the sustained vowel /a/ phonation (1–3 sec.). All voice data were down-sampled to 16 kHz with 16 bits. 70% and 30% of the data were used for the training and the testing set, respectively. The speakers were randomly selected from the database to build each set for a 10-fold cross-validation scheme.

3.1 Distributions of Conventional Features

In this paper, the jitter, the shimmer, the HNR_{Yumoto}, and $\gamma_2^{(v)}$ are utilized as the conventional features. The jitter (%) is a cycle-to-cycle frequency perturbation measure while the shimmer (%), a cycle-to-cycle amplitude perturbation one. The HNR_{Yumoto} shows the degree of the speech waveform aperiodicity and pathological voices are characterized by a smaller HNR_{Yumoto} than normal ones. The definition of the above features can be found in [5]. Since these features are based on the fundamental frequency, a high-quality reliable pitch detection algorithm is essential to measure voice irregularities accurately. In this paper, pitches are estimated by using the autocorrelation function (ACF) as appeared in [6]. Finally, as the second-order statistics (SOS)-based parameter, i.e., the variance of the variance, $\gamma_2^{(v)}$, is defined as in (7).

$$\gamma_2^{(v)} = \frac{1}{T-1}\sum_{t=1}^{T}(\gamma_{2,t} - \overline{\gamma}_2)^2 \quad (7)$$

Figures 1 (a), (b), and (c) show the box plots of the jitter (%), the shimmer (%), and the HNR_{Yumoto} (dB), respectively. They provide better visualization of G0, G1, G2, and G3 voices. The definite thresholds to classify normal (G0) and pathological (G1–G3) voices are easily found in Figs. 1 (a) and (c). In Figs. 1 (a) and (b), G3 voices show higher values and more broad distributions than others. Therefore, we can confirm that as the voices become more severe, they become more and more fluctuant in the fundamental period and the waveform amplitude. Figure 1 (c) shows that G0 and G1 voices tend to have more dominant harmonic components than the G2 and G3 ones.

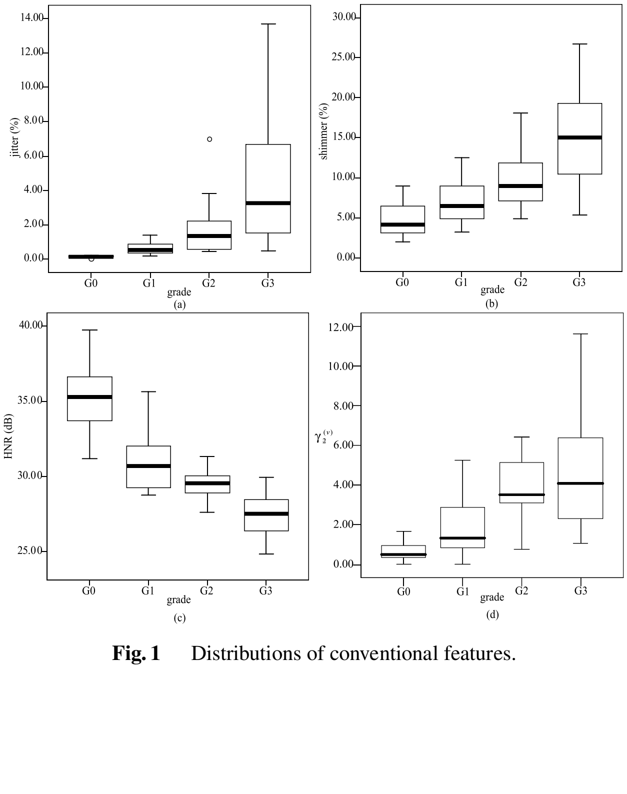

Fig. 1 Distributions of conventional features.

3.2 Distributions of HOS-Based Features

In our experiments, γ_3 and γ_4 are extracted for each 20 msec frame. Next, the means, $\overline{\gamma}_3$ and $\overline{\gamma}_4$, and the variances, $\gamma_3^{(v)}$ and $\gamma_4^{(v)}$, are calculated for a sentence. In calculating γ_3-related statistics, the absolute values of γ_3 are utilized because we confirmed that the signed and the absolute values do not make any meaningful difference. Figure 2 shows their distributions in G0, G1, G2 and G3 voices. In Fig. 2 (a), $\overline{\gamma}_3$ of pathological voices tend to have larger values than that of normal voices. Specifically, $\overline{\gamma}_3$ values of G3 voices are larger than those of others. By observing $\overline{\gamma}_4$ in Fig. 2 (b), pathological voices can be thought to have a leptokurtic distribution ($\overline{\gamma}_4 > 3$) and normal voices, a platykurtic ($\overline{\gamma}_4 < 3$). As the voices become more severe, $\overline{\gamma}_4$ spreads out rather widely with large values and have more leptokurtic distribution properties. And the variances of pathological voices tend to have larger values than those of normal voices in $\gamma_3^{(v)}$ and $\gamma_4^{(v)}$ of Figs. 2 (c) and (d). Specifically, G2 and G3 voices show larger variation than G0 and G1 voices in both $\gamma_3^{(v)}$ and $\gamma_4^{(v)}$. In general, the outliers are found in pathological voices. In Figs. 2 (e) and (f), the features to characterize the differences between frames of γ_3 and γ_4 are designated as (*SP*) and (*KP*), respectively. They also tend to have similar properties to those of $\gamma_3^{(v)}$ and $\gamma_4^{(v)}$. However, it can be seen that (*KP*) is better than $\gamma_4^{(v)}$ in classifying voice qualities. Based on the above observations, we might insist that the HOS analysis is more suitable to classify the signals characterized by an irregularity of the speech production mechanism.

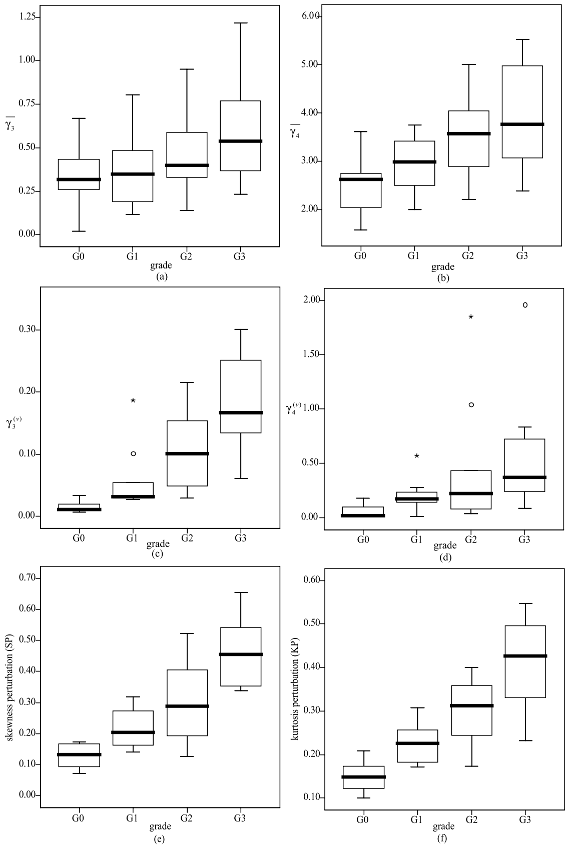

Fig. 2 Distributions of HOS-based features.

3.3 CART Experiment

In this paper, the CART method is used to combine the conventional and the HOS-based features [7]. Since some features make good decisions for the classification among G0, G1, G2 and G3 voices in the GRBAS scale and some do not, it is necessary to design a rule to make the final decision regarding the use of multiple inputs in the classifiers at the same time. Using the information of the multiple features from pathological and normal voice, the CART makes a final decision whether the current phonation is normal, a slight, a moderate, or a severe pathological voice.

Our first experiment to compare the performance between the SOS and the HOS gives the average performances are 66.2%, 77.5%, and 70.2% for $\gamma_2^{(v)}$, $\gamma_3^{(v)}$, and $\gamma_4^{(v)}$, respectively. It can be said that the utilization of the HOS-based features are more effective for the objective pathological voice quality assessment than the conventional SOS-one.

The CART algorithm is used to analyze the conventional features such as the jitter, the shimmer, and the HNR_{Yumoto}. Then, the average G0, G1, G2, and G3 voice classification performance is 83.1%. When $\gamma_2^{(v)}$ is added to the experiment as the SOS parameter, the average accuracy is 84.1%. When only the HOS-based features are used, the performance is 84.8%. And when the conventional and the HOS-based features are used together to generate the decision tree, the accuracy is 87.8%. It is the best performance among published papers, higher than the performance measured by Ping Yu et al. [2]. The optimal decision tree generated by the jitter, the shimmer, the HNR_{Yumoto}, $\gamma_2^{(v)}$, $\overline{\gamma}_3$, $\overline{\gamma}_4$, $\gamma_3^{(v)}$, $\gamma_4^{(v)}$, SP, and KP as its inputs is shown in Fig. 3. It is believed that $\gamma_3^{(v)}$ and SP among the HOS-based features are useful for the classification of G1, G2, and G3 voices.

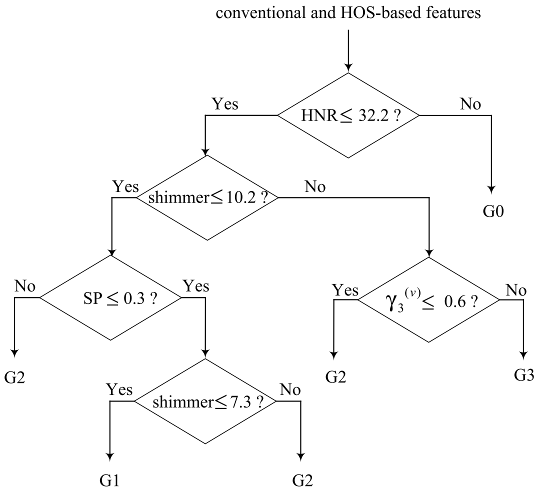

Fig. 3 Optimal decision tree formed by multiple features.

Table 1 Confusion matrix.

		Predicted				
		G0	G1	G2	G3	% correct
Manual	G0	**28**	2	0	0	93.3
	G1	0	**15**	2	0	88.2
	G2	0	2	**22**	2	84.6
	G3	0	1	2	**17**	85.0
	Total	28	20	26	19	**87.8**

Table 1 presents the confusion matrix based on the decision tree shown in Fig. 3. In this table, the final i.e., G0 versus G1 versus G2 versus G3, classification performance is 87.8% while the final G1 versus G2 versus G3, i.e., group-classification performance for pathological voices is about 85.9% while the normal-pathological, i.e., G0 versus G1+G2+G3, classification performance is about 89.6%. Although we are don't consider characteristics for the "roughness" and "breathiness" which may generally affect the G-based voice quality in the GRBAS scale, a satisfactory performance is obtained. It confirms if features to discriminate the roughness and breathiness are suggested, the best performance will be obtained in pathological voice quality assessment.

4. Conclusion

In this paper, novel features utilizing the HOS analysis have been introduced to improve the classification performance of the GRBAS scaled voices. Firstly, we analyze the characteristics of the conventional features, such as the jitter, the shimmer, the HNR_{Yumoto}, and $\gamma_2^{(v)}$. As the new HOS-based features, the means, the variances, the perturbations of the skewness and the kurtosis are estimated for voice samples. We also analyze their characteristics. A close correlation between the HOS-based features and the voice quality measurement has been demonstrated. The CART analysis based on the conventional and the HOS-based features has been executed to find an effective combination of multiple features. The optimal decision tree is obtained by the HNR_{Yumoto}, the shimmer, the perturbation, SP, and the variance, $\gamma_3^{(v)}$, of the skewness. The experiments demonstrate that the CART algorithm which uses both the conventional and the HOS-based features produces the highest reported classification performance of 87.8%.

As a future work, we plan to test our proposed algorithm with a larger database, especially for G2 and G3 voices, to improve the classification accuracy. We also plan to test the usefulness of our algorithm in real clinical circumstances. Finally, we plan to conduct more research in order to provide a rather reliable objective assessment of the voice quality according to the GRBAS scale.

References

[1] R. Sivakumar and G. Ravindran, "Automatic discrimination of abnormal subjects using the visual evoked potential spectral compoents," EURASIP Journal of Biomedicine and Biotechnology, vol.2004, no.1, pp.5–9, 2004.

[2] P. Yu, M. Ouaknine, J. Revis, and A. Giovanni, "Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements," Journal of Voice, vol.15, no.4, pp.529–542, 2001.

[3] L. Gu, J.G. Harris, R. Shrivastav, and C. Sapienza, "Disordered speech assessment using automatic methods based on quantitative measures," EURASIP Journal on Applied Signal Processing, vol.2005, no.9, pp.1400–1409, 2005.

[4] E. Nemer, R. Goubran, and S. Goubran, "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.

[5] R.D. Kent and M.J. Ball, Voice quality measurement, 1st ed., Thomson Learning, 2000.

[6] L. Hui, B. Dai, and L. Wei, "A pitch detection algorithm based on AMDF and ACF," Proc. ICASSP Conf., vol.1, pp.377–380, 2006.

[7] 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.