Automatic voice quality measurement based on efficient combination of multiple features

Ji-Yeoun Lee, Sangbae Jeong, and Minsoo Hahn
Information and Communications University
119, Munjiro, Yuseong-gu, Daejeon
305-732, Korea
Email: {jyle278, sangbae, mshahn}@icu.ac.kr

Hong-Shik Choi
Department of Otorhinolaryngology
Yongdong Severance Hospital
Yonsei University College of Medicine, Seoul, Korea
Email: hschoi@yumc.yonsei.ac.kr

***Abstract*—This work proposes higher-order statistics (HOS)-based features to improve classification performance of voice quality measurement. They are means and variances of skewness and kurtosis which show meaningful differences in normal, breathy, and rough voices. Jitter, shimmer, and harmonic to noise ratio (HNR) are implemented as conventional features. The performances are measured by classification and regression tree (CART) analysis. Specifically, the CART-based method by utilizing both conventional and HOS-based features is shown to be an effective for voice quality measurement, with an 89.7% classification rate.**

***Index Terms*—breathiness, roughness, voice quality measurement, higher-order statistics.**

I. Introduction

Pathological voices due to laryngeal disease have been described by their perceptual impression: hoarseness, breathiness, roughness, and so on. Several methods have been introduced to assess the perceptual qualities of pathological voices, and some voice specialists have begun to use them [1]. Nevertheless, it is one of the more controversial themes in vocal evaluation because there is a poor correlation between evaluators. For this purpose, a quantitative definition of the perceptual qualities must be established based on acoustic and physiological correlations [2].

The "GRBAS" scale is now widely used by voice specialists. It consists of five scales: "grade of hoarseness (G)", "roughness (R)", "breathiness (B)", "asthenicity (A)" and "strained quality (S)" [2]. "Roughness" and "breathiness" are very widely available among GRBAS scales [1]. The researches in this field have been reported that the acoustic features that are correlated with the perceptivity are jitter related to the R scale, shimmer connected with the B scale, and harmonic to noise ratio (HNR) related to the G scale [3]. The correlation between perceptual classifications and above features is shown in [4] and discriminant analysis was introduced to measure the voice quality [4].

This paper presents the characteristics of normal, breathy, and rough voices based on the analysis of various features and the classification performance of the voice qualities. The focus is to propose new features for the performance improvement in voice quality measurement. New features based on a higher-order statistics (HOS) analysis are investigated, which are means and variances of skewness and kurtosis. As the conventional features, jitter, shimmer, and HNR are implemented. The classification and regression tree (CART) analysis are used to combine multiple features and to measure their performances.

II. Conventional features

Over the past few years a considerable number of studies have been focused on the extraction of acoustic features for the objective judgment of pathological voices. Among acoustic features, the important ones are pitch, jitter, shimmer, and HNR. Since these features are based on the fundamental frequency, a very reliable pitch detection algorithm is essential to measure voicing irregularities. In this paper, it is extracted to use the autocorrelation function (ACF) which is applied in multi-dimensional voice program (MDVP), one widely used program [5].

According to pitch period based on ACF, jitter, shimmer, and HNR$_{Yumoto}$ are implemented in this paper. Jitter(%) is defined as in Eq. (1). It is a measure of cycle-to-cycle fluctuations in the fundamental period, $T0$, of vocal fold vibration [1].

$$Jitter(\%) = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}(|T0_n - T0_{n+1}|)}{\frac{1}{N}\sum_{n=1}^{N-1}T0_n} \times 100 \quad (1)$$

where N is the number of samples.

Shimmer(%) is defined as in Eq. (2). It relates to cycle-to-cycle variation in waveform amplitude, $A0$ [1].

$$Shimmer(\%) = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}(|A0_n - A0_{n+1}|)}{\frac{1}{N}\sum_{n=1}^{N-1}A0_n} \times 100 \quad (2)$$

Pathological voices are characterized by a smaller harmonic to noise ratio than healthy ones. This is due to the non-regularity of the vibration of the vocal cords and is measured by HNR$_{Yumoto}$ like Eq. (3) [1].

978-1-4244-1748-3/08/$25.00 © 2008 IEEE

$$HNR_{Yumoto} = 100 \times log_{10}(\frac{E_p}{E_{ap}}) \quad (3)$$

where E_p is the energy of periodic components and E_{ap} is the energy of aperiodic components.

III. Proposed features

A speech signal, $x(n)$, which may be normal, breathy, or rough voice, can be expressed as in (4) [6].

$$x(n) = s(n) + w(n) \quad (4)$$

where $s(n)$ is a non-Gaussian signal produced by vibration of the vocal folds and $w(n)$ is a Gaussian noise which can be assumed to be zero in normal voices and not to be zero in breathy and rough voices.

Breathy voices result when the vocal folds fail to close completely in each phonation cycle, and a steady stream of air rushes audibly through the glottis and resonance cavities. They are characterized by turbulence noise and audible escape of air through the glottis due to insufficient closure [1-2]. The degree of the noise, which is directly related to the perceived breathiness of voice, can be modeled by $w(n)$.

Rough voices are distinguished by a noisy, rasping, and unmusical tone. It indicates a psych-acoustic impression of aperiodic noise, presumably related to some kind of irregular vocal fold vibration [1-2]. The perceived degree of the noise can also be modeled by $w(n)$. According to the characteristics of the voices, $s(n)$ of rough voices may be have larger variation in the pitch period, breaks in pitch generation, and the presence of sub-harmonic components than $s(n)$ of breathy voices [1].

On the other hand, $s(n)$ of most normal voices have periodicity and stability. 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 [2].

Recently, the application of HOS to speech processing has been primarily motivated by their inherent Gaussian suppression and phase preservation properties [6-7]. Works in this area are based on the assumption that speech has HOS properties that are distinct from those of Gaussian noise [7]. Therefore, when HOS analysis is applied to pathological voices, unstable and discontinuous statistics of $x(n)$ may be estimated because HOS analysis is blind to Gaussian processes. In some cases, $s(n)$ and $w(n)$ may be modeled simultaneously through HOS analysis. In normal voice, HOS of only non-Gaussian measurements may be extracted because a Gaussian noise can be assumed to be zero [6-7].

Among various HOS statistics, the normalized skewness, γ_3, and the normalized kurtosis, γ_4, are widely used as characteristic features. They are defined as (5) [7].

$$\gamma_3 = \frac{\sum_{n=1}^{N}(x_n - \mu)^3}{(N-1)\sigma^3}, \qquad \gamma_4 = \frac{\sum_{n=1}^{N}(x_n - \mu)^4}{(N-1)\sigma^4} \quad (5)$$

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

For objective voice quality measurement, the proposed HOS-based features are as follows: $\overline{\gamma_3}$, $\overline{\gamma_4}$, $\gamma_3{}^{(v)}$, and $\gamma_4{}^{(v)}$. They are estimated in a sentence and have their roots in frame-based γ_3 and γ_4. Eq. (6) and (7) indicate the sentence-based means and the variances of γ_3 and γ_4 extracted in raw samples.

$$\overline{\gamma_3} = \frac{1}{T}\sum_{t=1}^{T}\gamma_{3t}, \qquad \overline{\gamma_4} = \frac{1}{T}\sum_{t=1}^{T}\gamma_{4t} \quad (6)$$

$$\gamma_3{}^{(v)} = \frac{1}{T}\sum_{t=1}^{T}(\gamma_{3t} - \overline{\gamma_3})^2, \quad \gamma_4{}^{(v)} = \frac{1}{T}\sum_{t=1}^{T}(\gamma_{4t} - \overline{\gamma_4})^2 \quad (7)$$

where γ_{3t} and γ_{4t} are γ_3 and γ_4 in the t^{th} frame, respectively and T is the number of the frames.

IV. CART algorithm

CART analysis is a common method to build statistical models founded on tree-based techniques. One of the most important characteristics of the CART is that the optimal decision tree contains the rules which are easily readable by humans compared to other classification and regression methods such as vector quantization (VQ) and neural networks (NNs). Decision tree contains a binary question about some feature at each node. The leaves of the tree contain the best prediction based on the training data [8].

To improve the performance of voice quality measurement, there have been many studies on feature extraction. However, each feature does not always guarantee the reliable performance in the various kinds of environments. Therefore, it may be necessary to use these features together to ensure the robustness in various conditions. This paper focuses on the efficient combination method of the multiple features for voice quality measurement. Statistical approach can be considered as a solution to effectively combine the multiple features. We use the CART algorithm for the classification of normal, breathy, and rough voices using multiple features.

V. Experiments and results

The Japan Society of Logopedics and Phoniatrics distributed a DVD-ROM database of 65 speakers based on GRBAS scale. Among them, we only used 53 pathological voices that are definitely divided into roughness and breathiness. If rough voice is a higher grade than breathy one, it is supposed to rough voice; otherwise, breathy voice. If the grade of rough and breathy voice is same, then the voice is ruled out in our experiments. These perceptual grades were determined by the juries composed of Japanese speech and language therapist (SALT). 30 normal Koreans voices were also added after careful examination by a group of speech experts. Hence, our database were composed of 30 voices with Korean normal (G0 voices), 30 pathological voices with roughness (R), and 23 pathological voices with breathiness (B). Since we were interested only in pathologies which affect the vocal folds,

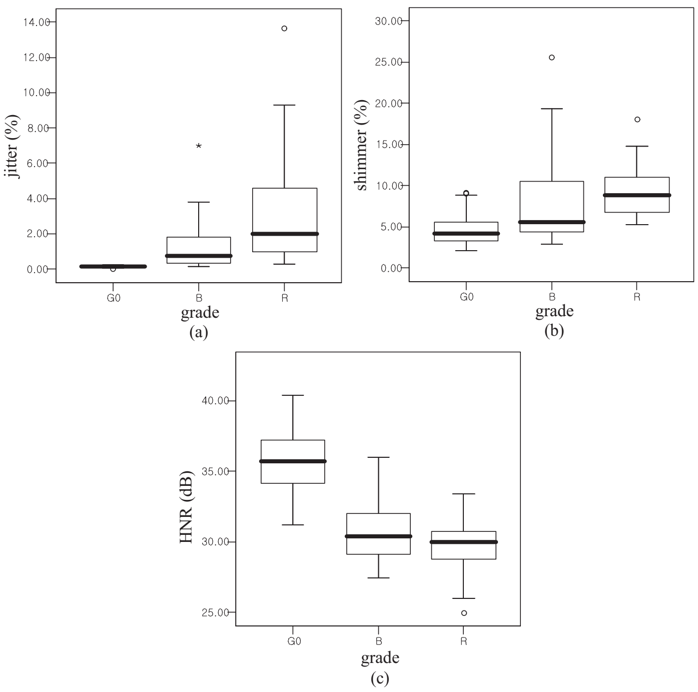

Fig. 1. Distributions of conventional features (G0: normal voices, B: breathy voices, R: rough voices)

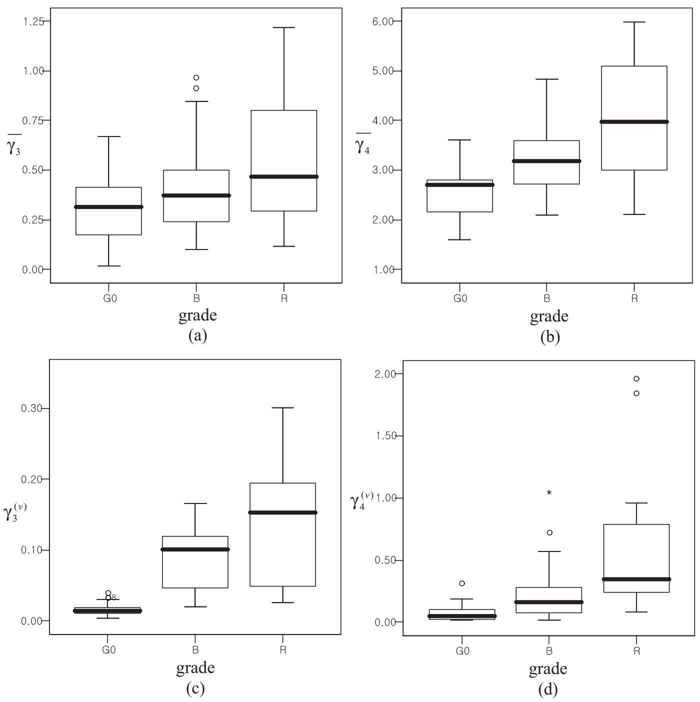

Fig. 2. Distributions of HOS-based features (G0: normal voices, B: breathy voices, R: rough voices)

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 for training and testing sets, respectively. The speakers were randomly selected from the database to build each set for a 10-fold cross-validation scheme [2].

A. Distributions of conventional features

Fig. 1(a), (b), and (c) show the distributions of jitter(%), shimmer(%), and HNR(dB), respectively. The box plots provide better visualization in normal, breathy, and rough voices. They are made by minimum, first quartile, median, third quartile, maximum values, and outliers of the conventional features. Correspondences between perceptual definition of voice quality and objective measurement are found in Fig. 1. The distributions between pathological and normal voices have a definite threshold. Rough voices that generally correspond to aperiodicity and noise have a tendency to higher values and more broad distribution than normal and breathy voices in jitter of Fig. 1(a). Breathy voices that coincide with turbulence noise and loudness weakness tend to be a broad range in shimmer of Fig. 1(b) and have more harmonic components than rough voices in HNR of Fig. 1(c).

B. Distributions of HOS-based features

First of all, γ_3 and γ_4 are extracted in each raw voice samples of 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 in a sentence. Absolute values of γ_3 are then used for the calculation of the HOS-based features. Fig. 2 shows their distributions in normal, breathy, and rough voices. In Fig. 2(a), $\overline{\gamma_3}$ of pathological voices tend to be larger values than that of normal voices. Specifically, $\overline{\gamma_3}$ values of rough voices are larger than those of others. In $\overline{\gamma_4}$ of Fig. 2(b), pathological voices can be considered to have a leptokurtic distribution ($\overline{\gamma_4} > 3$) and normal voices, a platykurtic ($\overline{\gamma_4} < 3$). $\overline{\gamma_4}$ of rough voices spreads out rather widely with large values and have more leptokurtic distribution. And the variances of pathological voices have a tendency to have larger values than those of normal voices in ${\gamma_3}^{(v)}$ and ${\gamma_4}^{(v)}$ of Fig. 2(c) and (d). Specifically, rough voices show larger variation than normal and breathy voices in both ${\gamma_3}^{(v)}$ and ${\gamma_4}^{(v)}$. In general, many outliers are found in pathological voices. Based on the above observations, we might insist that HOS analysis is more appropriate to discern the signals characterized by an irregularity of the speech production mechanism.

C. CART experiments

This part suggests the decision tree to combine the conventional and the HOS-based features.

As the first experiment, the CART algorithm is used to analyze the conventional features of jitter, shimmer, and HNR. Then, the performance is averagely 81% in distinguishing normal, breathy, and rough voices. Next, when the HOS-based features are only used for formation of the decision tree, the performance is averagely 80.3%. Two experiments show similar performances in voice quality measurement. Finally, the conventional and HOS-based features together are used to generate the decision tree. The accuracy is averagely 89.7%. The optimal decision tree formed by jitter, shimmer, HNR,

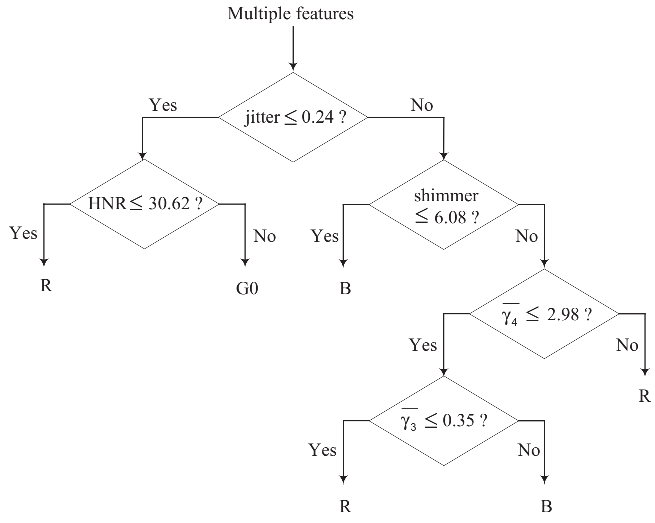

Fig. 3. Optimal decision tree formed by multiple features (G0: normal voices, B: breathy voices, R: rough voices)

$\overline{\gamma_3}$, $\overline{\gamma_4}$, ${\gamma_3}^{(v)}$, and ${\gamma_4}^{(v)}$ is shown in Fig. 3. We confirm that characteristics of the conventional and HOS-based features shown in Fig. 1 and Fig. 2 are reflected at each tree node. It can be said that the conventional features which have been used for a long time still have an important effect on the performance. It is also believed that $\overline{\gamma_3}$ and $\overline{\gamma_4}$ among the HOS-based features are useful for the classification of breathy and rough voices.

The confusion matrix is presented in Table 1. It is formed based on the decision tree shown in Fig. 3. Each matrix cell indicates how many instances with the corresponding actual class label were predicted by the model to have the corresponding predicted class label. The diagonal numbers indicate the performance of correctly classified signals. The off diagonal elements are associated with the performance of misclassifications. From the Table 1, we can observe that although the accuracy of groups such as B and R is not so high, good performance is shown between pathological and normal voices. A small part of breathy and rough voices are not classified as any of the defined classes and are designated as unclassified. Actually, mis-classification is inevitable because we are not considered for the "A" and "S" factors which may affect the performance in "GRBAS" scale.

TABLE I
CONFUSION MATRIX (G0: NORMAL VOICES, B: BREATHY VOICES, R: ROUGH VOICES)

		Predicted		
	(%)	G0	B	R
Actual	G0	100	0	0
	B	8.7	82.6	8.7
	R	0	13.3	86.6

VI. Conclusion

In this paper, novel features utilizing the HOS analysis have been introduced to improve the classification performance of normal, breathy, and rough voices. Firstly, we have analyzed the characteristics of the conventional features, such as jitter, shimmer, and HNR. As new HOS-based features, means and variances of skewness and kurtosis are suggested. They are estimated in raw voice samples and are calculated in a sentence. We have also analyzed their characteristics. A close correlation between the HOS-based features and voice quality measurement has been demonstrated. For the performance measurements of the multiple features, the CART algorithm has been implemented. Especially, the CART analysis based on the conventional and HOS-based features has been proposed to effective combination method of the multiple features. The optimal decision tree is obtained by jitter, shimmer, HNR, means of skewness and kurtosis. The experiments have demonstrated that the CART algorithm which uses the conventional and HOS-based features together can provide the highest classification performance, at 89.7%. This is very important, since CART analysis has been shown to be more appropriate for combining multiple features.

As the future work, our proposed algorithm should be tested with a larger database, especially for breathy and rough voices, to improve the accuracy of the system, and it has to be tested using continuous speech. In actual clinical circumstances, it will be tested for the application of a monitoring system for patients. Finally, researches will be investigated to provide an objective assessment of the voice quality according to GRBAS scales.

References

[1] R.D. Kent, M.J. Ball, Voice quality measurement, 1st Ed., Thomson Learning, 2000.
[2] N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, F. Cruz-Roldan, "Automatic Assessment of Voice Quality According to the GRBAS Scale," Proc. of the 28th EMBS Conf., pp. 2478-2481, 2006.
[3] M. Frohlich, D. Michaelis, H. Werner Strube, "Acoustic "breathiness measures" in the description of pathologic voices," Proc. of ICASSP Conf., Vol. 2, pp. 937-940, 1998.
[4] Ping Yu, M. Ouaknine, J. Revis, 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.
[5] Li Hui, Bei-qian Dai, Lu Wei, "A Pitch Detection Algorithm Based on AMDF and ACF," Proc. of ICASSP Conf., Vol. 1, pp. 377-380, 2006.
[6] Jesus B. Alonso, Jose de Leon, Itziar Alonso, Miguel A. Ferrer, "Automatic Detection of Pathologies in the Voice by HOS Based Parameters," EURASIP Journal on Applied Signal Processing, Vol. 4, pp. 275-284, 2001.
[7] E. Nemer, R. Goubran, S. Goubran, "Robust voice activity detection using higher-order statistics in the LPC residual domain," IEEE Trans. Speech and Audio Processing, Vol. 9, pp. 217 -231, 2001.
[8] S. Gey, E. Nedelec, "Model selection for CART regression trees," IEEE Trans., Information Theory, Vol. 51, pp. 658-670, 2005.