DETECTING PATHOLOGICAL SPEECH USING CONTOUR MODELING OF HARMONIC-TO-NOISE RATIO

Jung-Won Lee, Samuel Kim, and Hong-Goo Kang

Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

ABSTRACT

This paper proposes a new feature extraction method for automatically detecting pathological voice in a normal conversation scenario. Unlike conventional approaches that utilize the static harmonic-to-noise ratio (HNR) characteristics of sustained vowel, the proposed method considers the dynamic movements of articulatory organs depending on the types of phonations. Assuming those movements reflect the health status of subjects, the proposed method utilizes the characteristics of HNR contour within a single sentence-level speech signal. Experimental results show that the proposed method reduces the classification error rate by 35.2 % (relative) compared to the conventional method.

Index Terms— pathological speech, harmonic-to-noise ratio, dynamic characteristic, continuous speech

1. INTRODUCTION

Various approaches have been studied to automatically detect pathological voice with speech signals. The state of the art approach typically measures perturbation from sustained vowels, such as in the Multi-Dimensional Voice Program software package [1] and Praat [2], and cepstral/spectral-based parameters [3]. However, the analysis methods often rely on single vowel phonation of sustaining several seconds because they are simple. The metrics for sustained vowels include a perturbation measurement of fundamental frequency and amplitude, harmonics-to-noise ratio (HNR), and nonlinear chaotic measurements [4, 5, 6, 7, 8].

However, most clinicians regard continuous conversational speech as more informative than sustained vowel phonation because continuous speech contains abundant information, such as rapid adjustment of articulatory organs related to the initiation of voicing, which may not manifest during sustained vowel phonation. Thus, it is desirable and potentially more appropriate to investigate continuous speech for diagnosing pathological subjects. Researches on continuous speech have not been much investigated partly because of the inherent feature variations. For example, pitch variation or amplitude perturbation may occur by phonetic variations not by pathological reasons [9]. Mel-frequency cepstral coefficients (MFCCs), which are widely used in speech signal processing applications [10, 11], and nonlinear chaotic measures [8] also significantly vary depending on phonemes. Studies devoted to vocal aperiodicities, such as signal-to-noise ratio, have been conducted [1, 3, 12, 13, 14], but they only used static features, i.e., average value of estimated vocal aperiodicities computed in utterances.

The characteristics of vocal folds' vibrations keep changing across different types of phonations, such as onsets, offsets, transient or weak voiced regions. The supra-laryngeal impedance also varies especially during obstruents, and the larynx continually moves up and down in the neck. It is well known that adjustments across different phonations in continuous speech are challenging for pathological subjects [15]. This results in different dynamic characteristics between normal and pathological subjects; pathological subjects may fail to perform the necessary adjustments, while normal subjects can maintain rapid changes. Therefore, we hypothesize that dynamic characteristics of vocal folds' vibration patterns would be helpful for classifying normal and pathological speech.

This paper proposes an HNR contour modeling method that utilizes global statistics of localized features. Among a set of HNR contour related features, key features that can identify normal and pathological groups are selected by measuring information gain between features and subject groups. Classification of normal and pathological speech is also conducted using a support vector machine (SVM). Experimental results show that the proposed method significantly improves classification accuracy. Compared to the performance obtained from the conventional method using HNR only, the proposed system using HNR contour modeling reduces error rates by 35.2 %.

Section 2 introduces dynamic characteristics of the HNR contour, and describes the proposed HNR contour modeling method. In Section 3, the performance of the proposed algorithm is compared with that of the conventional approach after conducting feature analysis. The conclusion follows in Section 4.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. FEATURE EXTRACTION

2.1. Harmonic-to-noise ratio

HNR is defined as the energy ratio between periodic and aperiodic components:

$$\mathrm{HNR}(l) = 20\log\left(\frac{\sum_{m=m_i}^{m_j} ||S(m,l)| - |N(m,l)||}{\sum_{m=m_i}^{m_j} |N(m,l)|}\right), \quad (1)$$

where $S(m,l)$ and $N(m,l)$ are short-time Fourier transform of original signal and aperiodic components, respectively. l and m denote the frame index and frequency bin index. Aperiodic components $N(m,l)$ are regarded as the residuals of long-term predictive analysis (LTP) [12, 13, 15]. In the LTP step, the current analysis frame of length L is predicted by a lagged frame of the same length such that

$$\hat{s}(k) = \beta s(k-T), \quad (2)$$

where $s(k)$ is the current speech sample, T is the prediction lag with $T_{\min} \leq |T| \leq T_{\max}$, and β is the long-term prediction coefficients. $T_{\max}$ and $T_{\min}$ are fixed to 25ms and 2.5ms, considering human pitch range. The optimal long-term prediction coefficient is derived by minimizing prediction error energy E, i.e.,

$$E = \sum_{k=0}^{L-1} e^2(k) = \sum_{k=0}^{L-1} [s(k) - \beta s(k-T)]^2, \quad (3)$$

which yields

$$\beta = \frac{\sum_{k=0}^{L-1} s(k)s(k-T)}{\sqrt{\sum_{k=0}^{L-1} s^2(k) \ \sum_{k=0}^{L-1} s^2(k-T)}} \ . \quad (4)$$

β is bounded to 1. The frame length L is set to 80 samples (5ms).

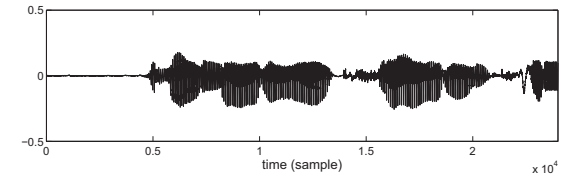
To extract the HNR contour, HNR is computed at every 10 ms based on the technique given in Eg. 1. The voiced segments are detected by considering the absolute value of normalized cross correlation [16].

Typically, normal voices show relatively strong harmonic structures up to 4 kHz. In the case of the pathologic voice, however, its spectrum includes higher noise levels than normal one with a deteriorated harmonic structure even at lower frequencies. Therefore, HNRs at limited frequency bands are beneficial for discriminating pathologic voices from normal ones [4, 6, 17]. This paper uses a bandpass filter whose frequency band is between 2kHz and 4kHz, which is set based on our prior experimental results [18].

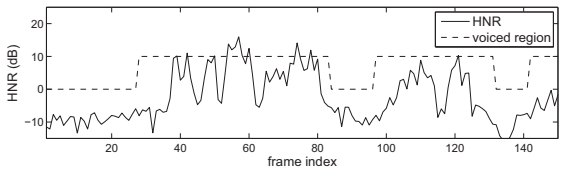
2.2. Dynamic characteristics of the HNR contour

In continuous speech, the vibration pattern of the vocal folds keeps changing because the mechanism of voice production varies depending on the type of phonemes, e.g., onsets, offsets, transient or weak voiced regions. Since the capability of voice production mechanism between normal and pathological person is different, their speech show different dynamic characteristics. For example, in the HNR contour shown in Fig. 1, normal speech has more prominent peaks and valleys than pathological speech. Therefore, it is expected that the dynamic characteristics of the HNR contour must be a good feature for discriminating normal and pathological speech.

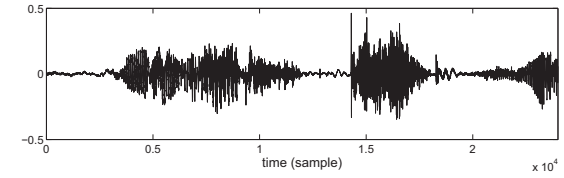
This paper attempts to capture the characteristics of the HNR contour; in particular, the *attack* and *release* patterns are modeled in the intensity level of the HNR contour (see Fig. 2 for an example). Attack is often described as an energy level increase when a sound event starts, while release is an energy level decrease region where a sound event ends. To parameterize the region of attack and release, positive and negative peaks (local maxima and local minima, respectively) are detected at first using the HNR contour of given speech signal in voiced frames. Then, the periods from negative peaks to positive peaks are considered as attacks and the periods from



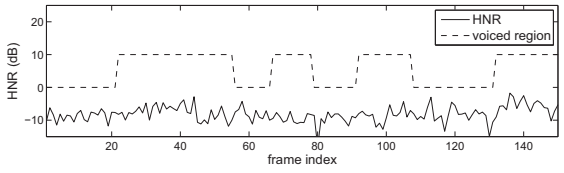
(a) Normal speech



(b) HNR contour for normal speech



(c) Pathological speech



(d) HNR contour for pathological speech

Fig. 1. HNR contour of example utterances from normal ((a) and (b)) and pathological speech ((c) and (d)).

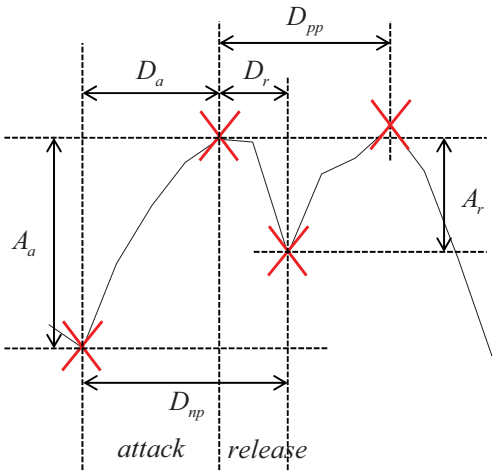


Fig. 2. Modeling the HNR contour using the concept of *attack* (a) and *release* (r). The term D_a, A_a, D_r, A_r, D_{pp}, and D_{np} indicate duration and amplitude of attacks and releases, durations between positive or negative peaks, respectively.

positive peaks to negative peaks as releases.

To represent the contour, fourteen parameters are used: duration, amplitude, slope (the ratio of amplitude change and duration), and convexity of attacks and releases (8 features), durations between positive or negative peaks (2 features), and convexity between negative peaks (1 feature) (see Fig. 2 for an example). Convexity is calculated as the sum of the difference between each point and the linear interpolation between the start and end values of the segment, i.e.,

$$convexity = \frac{\sum_{n=n_1}^{n_2} [s(n) - h(n)]}{n_2 - n_1} \quad (5)$$

where n_1 and n_2 are respectively the start and end times of the HNR contour, $s(n)$ is the HNR value at frame n, and $h(n)$ is the linear interpolated function,

$$h(n) = \frac{s(n_2) - s(n_1)}{n_2 - n_1}(n - n_1) + s(n_1) \quad (6)$$

for $n_1 \leq n \leq n_2$, and $n_1 \leq n_2$, respectively. The time derivatives of the HNR contour (HNR delta) during attacks and releases (2 features) and HNR values themselves (1 feature) are included. Once the parameters reflecting attacks and releases are obtained, a set of statistics of those fourteen parameters, such as mean ($E[\cdot]$), standard deviation ($\sigma[\cdot]$), interquartile range ($Iqr[\cdot]$), maximum ($Max[\cdot]$), median ($Med[\cdot]$), minimum ($Min[\cdot]$), and four percentile values (10th ($Q_{10}[\cdot]$), 25th ($Q_{25}[\cdot]$), 75th ($Q_{75}[\cdot]$) and 90th ($Q_{90}[\cdot]$)) are calculated to yield 140 features.

3. PERFORMANCE EVALUATION

3.1. Database

The voice recordings consist of utterances from pathological and normal speech collected in Samsung Medical Center, Seoul, Korea. The database contains readings of a passage (about 8 seconds) in Korean, recorded by 2379 pathological (1155 female, 1224 male), and 235 normal (105 female, 130 male) subjects. The recordings of pathological speech were obtained by considering a great range of speech system pathologies (polyp, nodule, Reinke's edema, cyst, contact granuloma, unilateral vocal cord paresis, sulcus, atrophy, etc.). The data samples were recorded in different sessions in a sound treated booth using a standardized recording protocol. The sampling frequency is downsampled to 16 kHz.

3.2. Feature analysis

Fig. 3 illustrates box-and-whisker plots of various features to represent the characteristics of normal and pathological speech. As we discussed earlier, in continuous speech, the characteristics of vocal folds' vibrations keep changing across different types of phonations, which is challenging for pathological subjects. Therefore, the HNR contour of normal speech contains more prominent peaks and valleys than

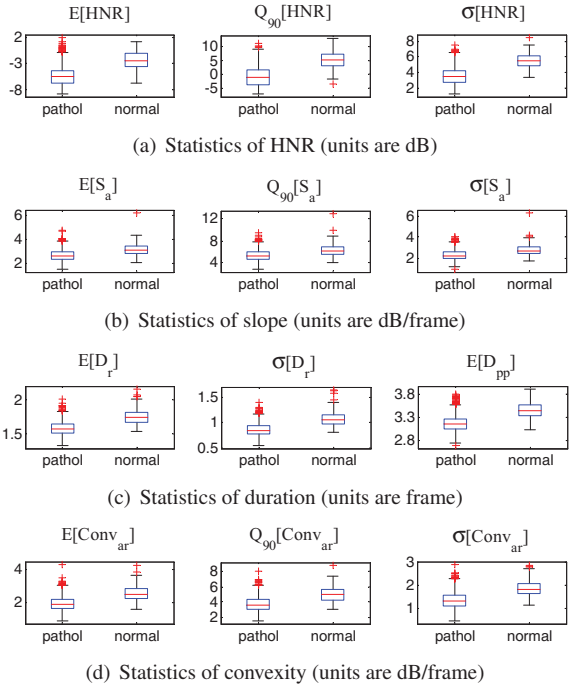


Fig. 3. Box-and-whisker plots of some significant HNR contour features: (a) HNR value themselves, (b) slope, (c) duration, and (d) convexity. The term S_a, $Conv_{ar}$ indicate slope of attack and convexity between negative peaks, respectively.

Table 1. Top twelve features according to information gain. The outside and inside brackets in the description of feature indicate statistics and parameters, respectively.

rank	feature	info. gain	rank	feature	info. gain
1	σ[HNR]	0.166	7	$E[D_{pp}]$	0.136
2	Q_{90}[HNR]	0.158	8	$E[D_{nn}]$	0.134
3	Max[HNR]	0.153	9	$\sigma[D_r]$	0.134
4	Iqr[HNR]	0.152	10	$Q_{90}[A_r]$	0.133
5	$E[D_r]$	0.148	11	$\sigma[A_a]$	0.133
6	Q_{75}[HNR]	0.137	12	E[HNR]	0.132

that of pathological speech. This results in greater values for static features (such as mean, high rank of percentile, and dispersion for HNR in Fig. 3 (a)) as well as dynamic features (such as those for slope in Fig. 3 (b)). In addition, pathological subjects have difficulties in maintaining harmonics to be monotonically increasing or decreasing. Note that normal subjects usually have greater values for duration related features (Fig. 3 (c)). The other observation is that convexity related features for normal subjects are greater than those for pathological ones (Fig. 3 (d)).

The information gain between selected features and speaker groups is examined to see which feature can be inferred with high confidence to discriminate normal and pathological subjects. In this paper, information gain is calculated by measuring mutual information between individual feature and classes with the WEKA data mining software [19]. Table 1 shows the top twelve features that show the maximal information gain. Overall, some features using dispersion, mean, and high rank of percentile in HNR values, amplitude, and duration are selected along with the average value of HNR which is regarded as a conventional feature. Results show that global statistics of HNR values have greater values than duration or amplitude related features in terms of information gain.

3.3. Classification tasks

Experiments for classifying normal and pathological speech are conducted to evaluate the performance of the proposed HNR contour related features. The 10-fold cross validation is used to reduce the influence of training tokens. Discrimination between normal and pathological subject is conducted using SVM with a radial basis function kernel. The distance of SVM output is used to obtain equal error rate (EER) which is the rate at which both missed detection and false alarm error are equal.

Fig. 4 illustrates the EERs using the HNR contour features by accumulating the feature one by one according to the order of information gain. For each combination, the best EERs were selected by varying the diverse sigma value of the RBF kernel. Table 2 shows the comparisons of EERs for

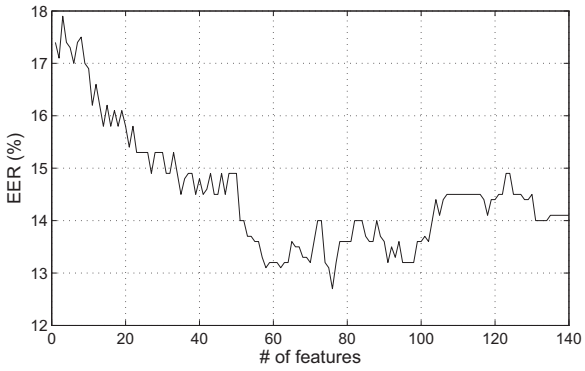


Fig. 4. EERs for the HNR contour features according to feature combination.

the conventional method using HNR only and HNR contour modeling. Performance in all the combinations using measurements of HNR contour modeling is always better than the one for the conventional method using HNR only. When the number of features was 76, it showed the best performance (EER = 12.7 %). When more than 76 features were used, the performances were not better than the best one due to the redundancy between features or the low information gain of the added features. In the case of best performance, relative error for classifying normal and pathological speech is improved by 35.2 % compared to the conventional method using HNR only.

Table 2. EERs (in %) for conventional method using HNR only and HNR contour modeling.

	HNR only	Proposed features
EER	19.6	12.7

4. CONCLUSION

This paper proposed an HNR contour modeling method to discriminate normal and pathological voices using sentence-level read speech signal. We captured the dynamic characteristics of the HNR contour by modeling local characteristics and their global statistics. Experimental results showed that the proposed HNR contour modeling outperformed the conventional method only using the HNR feature. The results support the hypothesis that the dynamic characteristics of vocal folds' vibration patterns in continuous speech are useful in discriminating normal and pathological speech.

Since the characteristics of vocal folds' vibrations vary across different types of phonation and the mechanism of voice production, studies on investigating the relative effectiveness of localized HNR contour variation depending on the type of phones must be a very interesting topic in the future.

5. REFERENCES

[1] D. D. Deliyski, "Acoustic model and evaluation of pathological voice production," *3rd Conference on Speech Communication and Technology EUROSPEECH'93,* pp. 1969–1972, 1993.

[2] P. Boersma, and D. Weenink, Praat: doing phonetics by computer. 5.3.04 ed. [Computer program]. Available at: http://www.praat.org/.

[3] S. N. Awan, N. Roy, M. E. Jette, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V.," *Clinical Linguistics & Phonetics,* vol. 24 pp. 742–758, 2010.

[4] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. Biomed. Eng.,* vol. 53, no. 3, pp. 468–477, Mar. 2006.

[5] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech Lang. Hearing Res.*, vol. 43, pp. 469–485, 2000.

[6] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Computer Methods and Programs in Biomedicine*, vol. 91, pp. 36–47, 2008.

[7] Y. Maryn, P. Corthals, P. V. Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," *J. Voice*, vol. 24, pp. 540–555, 2010.

[8] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-de-María, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Trans. Audio, speech, Lang. Process.,* vol. 17, pp. 1186–1195, 2009.

[9] G. Constantinescu, D. Theodoros, T. Russell, E. Ward, S. Wilson, and R. Wootton, "Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application," *International Journal of Language & Communication Disorders*, vol. 45, pp. 630–644, 2010.

[10] J. I. Godino-Llorente, Ruben Fraile, N. Saenz-Lechon, V. Osma-Ruiz, and P. Gomez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomed Signal Process Control*. vol. 4, pp. 176–182, 2009.

[11] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. 2nd Joint EMBS/BMES Conf.,* vol. 1, pp. 182–183, 2002.

[12] F. Bettens, F. Grenez, and J. Schoentgen, "Estimation of vocal dysperiodicities in connected speech by means of distant-sample bidirectional linear predictive analysis," *J. Acoust. Soc. Am.,* vol. 117, pp. 328–337, 2005.

[13] A. Alpan, Y. Maryn, A. Kacha, F. Grenez and J. Schoentgen, "Multi-band dysperiodicity analyses of disordered connected speech," *Speech Communication,* vol. 53, pp. 131–141, 2011.

[14] Y. Qi, R.E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *J. Acoust. Soc. Am.,* vol. 105, pp. 2532–2535, 1999.

[15] A. Kacha, and F. Schoentgen, "Estimation of dysperiodicities in disordered speech," *Speech Communication,* vol. 48, pp. 1365–1378, 2006.

[16] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 497–518, 1995.

[17] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech Hearing Res.*, vol. 36, pp. 224–266, 1993.

[18] J. W. Lee, H. G. Kang, S. Kim, and Y. Lee, "Detecting pathological speech using local and global characteristics of harmonic-to-noise ratio," *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct. 2013.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, volume 11, 2009.