Detecting pathological speech using local and global characteristics of harmonic-to-noise ratio

Jung-Won Lee*, Hong-Goo Kang*, Samuel Kim* and Yoonjae Lee†
*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: jaesuk2002@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr, samuel.kim@yonsei.ac.kr
†Digital Media & Communication R&D Center, Samsung Electronics Co. Ltd., Suwon, Korea
E-mail: yj0604.lee@samsung.com

***Abstract*—This paper proposes an efficient feature extraction method for automatic diagnosis systems to detect pathological subjects using continuous speech. Since continuous speech contains slow and rapid adjustments of vocal mechanisms which relate to initiations and terminations of voicing, the proposed algorithm utilizes both localized temporal characteristics and histogram-based global statistics of harmonic-to-noise ratio (HNR) to efficiently differentiate the key features from phonetic variation. Experimental results show that the proposed method improves the classification error rate by 11.2 % (relative) compared to the conventional method using HNR.**

I. Introduction

Automatic detection of pathological speech using features that are extracted from acoustic signals is being actively studied. The acoustic analysis methods of speech signals often rely on single vowel phonation of sustaining several seconds for its simplicity. The metrics include a perturbation measurement of fundamental frequency and amplitude, and harmonics-to-noise ratio (HNR) [1], [2], [3].

However, most clinicians regard continuous speech as more informative than sustained vowel phonation because continuous speech involves slow and rapid adjustments of vocal mechanisms which relate to initiations and terminations of voicing that are not present during sustained vowel phonation. Thus, it would be desirable and potentially more appropriate to investigate continuous speech toward diagnosing pathological subjects. Researches on continuous speech have not been much investigated compared to the ones on sustained vowels partly because of the feature variations in continuous speech. For example, pitch variation or amplitude perturbation may occur by phonetic variation not by pathological reasons in continuous speech. In addition, it may not be a good idea to use Mel-frequency cepstral coefficients (MFCCs), which are widely used in the speech signal processing community [4], [5], because they also vary depending on phonemes. Studies devoted to vocal aperiodicities, such as signal-to-noise ratio, in continuous speech have been conducted [6], [7], [8], but they only used static features, i.e., average value of estimated vocal aperiodicities in utterance.

The characteristics of vocal folds' vibrations keep changing across different types of phonations, such as onsets, offsets, transient or weak voiced regions. Furthermore, the supralaryngeal impedance varies especially during obstruents, and the larynx continually moves up and down in the neck [9]. Those conditions under which adjustments across phonations must take place in continuous speech is considered to be challenging for pathological subjects. This results in the difference of dynamic characteristic, i.e., rapid changes in HNR contour of normal speech, compared to pathological speech. Therefore, dynamic characteristic would be helpful for discriminating normal and pathological speech.

To obtain dynamic characteristics of harmonicity, we propose the time derivatives of HNR contour in voiced region. By analyzing statistical distributions of both static and dynamic features of HNR in the sentence, specific regions where normal and pathological groups can be identified clearly are determined. The reliability of the proposed approach is verified by measuring Jensen-Shanon divergence of feature distributions. Classification of normal and pathological speech is also conducted using 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 both dynamic characteristics and histogram-based global statistics reduces error rates by 11.2 % relatively.

II. Feature extraction for baseline

A. Database

The voice recordings consist of utterances from pathological and normal speech collected by Samsung Medical Center, Seoul, Korea. The database contains phonation of the vowel /aa/, along with 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 data samples were recorded in different sessions in a sound treated booth using a standardized recording protocol. In this study, only a passage sample is used. The sampling frequency is downsampled to 16 kHz.

B. Harmonic-to-noise ratio

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

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

where $S(m,l)$ and $N(m,l)$ are short-time Fourier transform of original signal and aperiodic components, respectively. l and m are the frame index and frequency bin index. Aperiodic components $N(m,l)$ can be considered as the residuals of long-term predictive analysis [6]. The current analysis frame of length L is predicted by a lagged frame of the same length such that

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

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

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

which yields

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

β is bounded to be equal to or less than 1. The optimum value is the lag for which the prediction error energy becomes minimum, i.e.,

$$T_{opt} = \arg\min_{T} \left\{ \sum_{k=0}^{L-1} \left[s(k) - \beta s(k-T)\right]^2 \right\}. \tag{5}$$

The instantaneous value of the prediction error (residual signal) is calculated as follows,

$$e(k) = s(k) - \beta_{opt} s(k - T_{opt}). \tag{6}$$

The short-time Fourier transform of $e(k)$ becomes $N(m,l)$ given in (1).

C. Determination of HNR frequency bands

HNR for normal subjects is expected to be higher than that of pathological subjects. In literature, the normal voice shows a relatively strong harmonic structure up to about 4 kHz. In the case of the pathologic voice, the spectrum includes higher noise levels than normal one with deteriorated harmonic structure even at lower frequencies. Therefore, the harmonic-to-noise energy ratio (HNR) at limited frequency bands can be beneficial for discriminating pathologic voices from normal ones [1], [3], [10]. To assess the distribution distance of normal and pathological speech in limited frequency bands, Jensen-Shanon divergence is used [11]:

$$JS(p;q) = \frac{1}{2}\sum_i p_i \ln \frac{p_i}{\frac{1}{2}(p_i+q_i)} + \frac{1}{2}\sum_i q_i \ln \frac{q_i}{\frac{1}{2}(p_i+q_i)}, \tag{7}$$

where p and q are two probability distributions. In this paper, probability distributions are modeled as four-mixture Gaussians. Table I illustrates that FB2040, i.e., frequency band between 2 kHz and 4 kHz, shows the highest distance. Since the setup also shows the best classification accuracy, this paper uses the FB2040 condition for the baseline system.

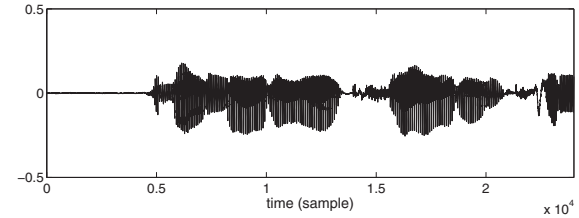
TABLE I
JENSEN-SHANON DIVERGENCE BETWEEN NORMAL AND PATHOLOGICAL DISTRIBUTION OF HNR IN VARIOUS FREQUENCY BANDS. THE TERM $FB1525$ INDICATES THE CASE OF USING FREQUENCY BAND BETWEEN 1.5 KHZ AND 2.5 KHZ.

	FB1525	FB2535	FB2040	FB2080	Full band
JS dist.	0.226	0.257	0.262	0.256	0.034

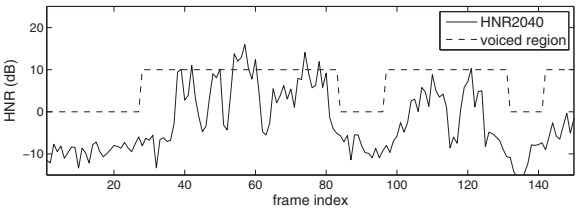
III. LOCAL AND GLOBAL CHARACTERISTIC ANALYSIS OF HNR

A. Dynamic characteristics of HNR

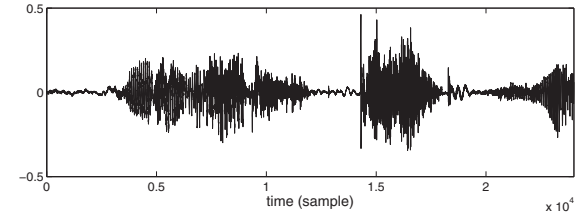
In continuous speech, the vibration of the vocal folds keeps changing depending on the types of phonemes, e.g., onsets, offsets, transient or weak voiced regions. This results in the difference of dynamic characteristic between normal and pathological speech. In the HNR contour of normal speech, there are more peaks and valleys than that of pathological



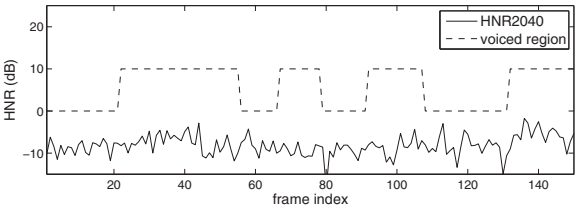
(a) normal speech



(b) HNR2040 contour for normal speech



(c) pathological speech



(d) HNR2040 contour for pathological speech

Fig. 1. HNR2040 contour of example utterances from normal ((a) and (b)) and pathological speech ((c) and (d)).

speech due to rapid changes in transition region.

Fig. 1 shows the examples of HNR contour in continuous speech. As shown in the figure, it is expected that the degree of dynamic characteristics in HNR contour can be a good feature for discriminating normal and pathological speech. From this observation, the time derivatives of HNR contour in voiced region are calculated as follows:

$$\Delta\mathrm{HNR}\,(l) = \mathrm{HNR}(l)\ - \mathrm{HNR}(l-1), \tag{8}$$

where l is frame index. Dynamic feature (HNR delta) is obtained by averaging the absolute value of time derivatives of HNR contour over the voiced speech. HNR delta for normal speech is expected to be higher than that of pathological subjects.

B. Histogram-based global statistics of HNR

Normal speech sometimes shows low HNR values even in voiced region due to various reasons such as prosody variation depending on person or phoneme characteristics that have harmonic components only in low frequency bands (e.g., nasal sound). Low HNR values in normal speech increase uncertainty with those in pathological speech, which decreases classification accuracy. Similar phenomenon can be observed while using the HNR delta feature. Although rapid changes of HNR contour in phoneme boundaries are occurred more frequent in normal speech rather than in pathological speech, the HNR contour of normal speech still changes slowly in quasi-stationary region. Left plots of Fig. 2 illustrate the histogram of HNR (top), cumulative distribution function (CDF) of HNR (middle), and HNR difference with respect to CDF (bottom) between normal and pathological subjects in all voiced frames of sentence. Similarly, right plots of Fig. 2 depict the ones for time derivatives of HNR. From Fig. 2, it is clear that the degree of overlap between normal and pathological distribution in low rank is higher than in high rank. To assess the distance of two distributions corresponding each rank, Jensen-Shanon divergence is measured again between the two groups of normal and pathological subjects. To calculate this, HNR and time derivatives of HNR obtained from segmented frames in each sentence are sorted, and are averaged after dividing into ten groups.

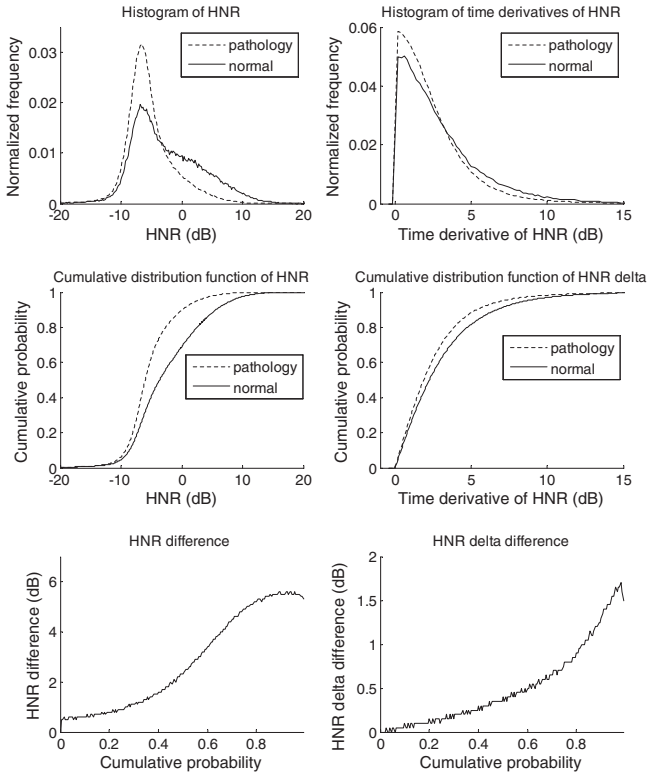
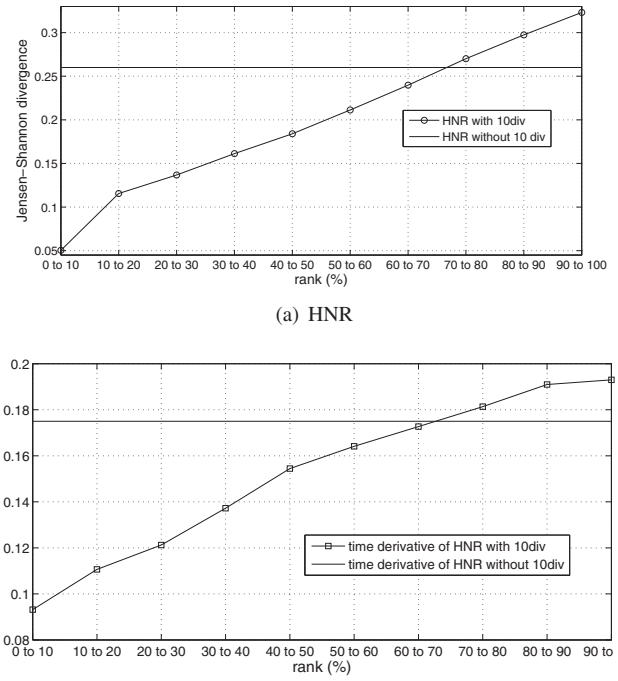


Fig. 2. Histogram (top), cumulative distribution function (middle), and difference with respect to cumulative distribution function (bottom) between normal and pathological subjects for HNR (left) and time derivatives of HNR (right).



(a) HNR

(b) Time derivative of HNR

Fig. 3. Jensen-Shanon divergence between normal and pathological distribution of (a) HNR and (b) time derivatives of HNR according to histogram rank. The term $10div$ indicate the 10 divisions of feature rank.

To assess the distribution distance of normal and pathological speech in HNR and HNR delta (top in Fig. 2), Jensen-Shanon divergence is measured. Results in Fig. 3 show that the distances of some divisions in high rank are bigger than one with using all features in both HNR and HNR delta. It means that some divisions in high rank have better capability than using all the features to classify normal and pathological speech.

IV. Performance Evaluation

A. Experimental setup

Experiments for classifying normal and pathological speech are conducted to evaluate the performance of the proposed

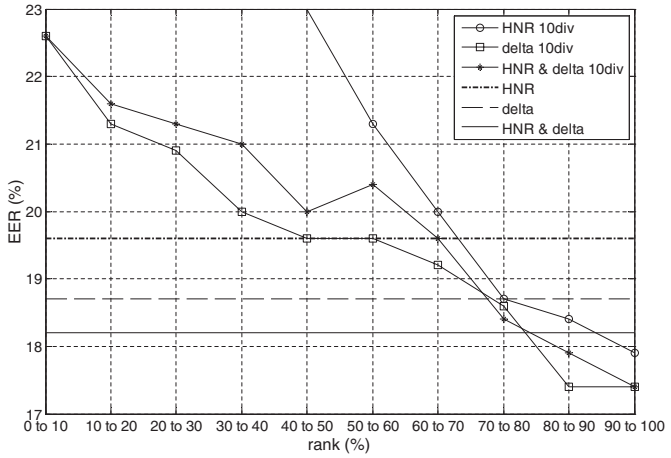


Fig. 4. EERs for static and dynamic feature of HNR according to histogram rank. Results without histogram rank are shown as horizontal line.

dynamic feature and histogram-based global statistics of HNR.

For HNR, aperiodic components are calculated from speech signals at every 2.5 ms using a 5 ms Hanning window, then HNR is extracted at every 10 ms. The voiced segments of passage have been obtained using the pitch tracking method [12]. The 10-fold cross validation is used to reduce the influence of training tokens. Discrimination between normal and pathological subject is conducted by means of SVM with a radial basis function kernel. The distance of SVM output is obtained, and in order to evaluate the classification performance of normal and pathological subjects, equal error rate (EER) is used.

B. Results and analysis

Fig. 4 and Table II show the EERs of HNR and HNR delta by varying the histogram rank. Results in either static or dynamic feature of HNR show that some divisions in high rank have better performance than using all features to classify normal and pathological speech. When using 90 to 100% rank of HNR and HNR delta features, relative error for classifying normal and pathological speech is improved by 8.7% and 7.0%,respectively, compared to using all of features. This indicates that by selecting high rank features, i.e. high HNR values in utterance for HNR and rapid changes in transition region for HNR delta, the difference between normal and pathological speech becomes prominent.

When the dynamic feature of HNR is combined with conventional static feature, relative error for classifying normal and pathological speech is improved by 7.1%, compared to the conventional method using HNR only. The performance of both static and dynamic features with high rank of feature distribution (90 to 100%) shows the relative improvement of error rate by 11.2%.

V. Conclusions and Future Work

This paper proposed a method using dynamic characteristics and histogram-based global statistics of HNR to normal and pathological speech using continuous speech. In such circumstances, the characteristics of speech signals dynamically vary across phonemes. We obtained the time derivatives of HNR and used the high rank part of feature distribution in both static and dynamic feature to extract the pathology-specific information. Experimental results showed that the proposed dynamic feature provided complementary information to conventional HNR feature.

We have limited dynamic feature as the simple time derivatives of HNR in this work. In the future, we will explore the potential of using "localized temporal characteristics". For example, we will use the duration of maintaining harmonics to be monotonically increasing or decreasing.

TABLE II
EERs (in %) in Figure 4.

	HNR	HNR delta	HNR & HNR delta
without 10div	19.6	18.7	18.2
with 10div (90 to 100%)	17.9	17.4	17.4

Acknowledgment

The authors would like to thank Dr. Young-Ik Son, Department of Otorhinolaryngology - Head and Neck Surgery, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, South Korea, for the audio recordings.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Information Theory*, vol. 37, pp. 145-151, 1991.

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