Speech Enhancement for Pathological Voice Using Time-Frequency Trajectory Excitation Modeling

Eunwoo Song*, Jongyoub Ryu† and Hong-Goo Kang*,
*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: sewplay@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
†Digital Media & Communication R&D Center, Samsung Electronics Co. Ltd., Suwon, Korea
E-mail: jongyoub.ryu@samsung.com

***Abstract*—This paper proposes a speech enhancement algorithm for pathological voices using a time-frequency trajectory excitation (TFTE) modeling. The TFTE model has a capability of delicately controlling the periodic and non-periodic excitation components by taking a single pitch based decomposition process. By investigating the difference of frequency characteristics between pathological and normal voices, this paper proposes an enhancement algorithm which can efficiently reduce the breathiness of the pathological voice while maintaining the identity of the speaker. Subjective test results are presented to verify the effectiveness of the proposed algorithm.**

I. Introduction

As the industry of U-health care with a telemedicine system has been greatly advanced, it has become more convenient for patients to receive a medical treatment even in their home. The patients who have trouble with pathological voice, however, cannot have this privilege due to difficulties in voice communication with their doctor. To allow them to conveniently use a mobile for communication, the needs for deploying speech enhancement algorithm have been increased.

Considering the specific characteristics of pathological voice, several techniques have been developed. One of them is speech enhancement for dysarthria who has neurologic disability including Parkinsons, Multiple Sclerosis, and stroke [1]. In order to enhance the voice of dysarthria, a formant synthesizer-based and a cepstrum-based approach are introduced [1][2]. Although these techniques improve intelligibility, they are limited to neurologic disorder cases, which is related to the vocal tract (or formant), rather than vocal folds (or excitation)-related disorder. On the view of the vocal folds-related disorder, a code-excited linear prediction (CELP)-based approach is introduced for a post-laryngectomised patient [3]. Since the post-laryngectomees do not have the functional capacity of the vocal folds, their voice is similar to hoarse whispering voice. To make and control their excitation be more natural, Sharifzadeh also chose an analysis-by-synthesis-based technique [3]. However, it is too difficult to universally implement the system to the other patients because its synthetic quality is artificial due to the problem on constructing excitation signals.

In order to provide more general solution to the vocal folds-related patients, this paper proposes a novel approach to enhance the characteristics of excitation signals. In general, some of the vocal folds-related patients are characterized by glottal insufficiency (GI) that includes a vocal fold atrophy, unilateral vocal fold paresis, scar defects, presbyphonia, etc. [4]. They usually suffer from their breathy, weak, or even aphonic voice caused by the incomplete glottal closure during phonation. It is important, therefore, to reduce the noisy (or non-periodic) components of the excitation signal while enhancing the harmonic (or periodic) components.

A time-frequency trajectory excitation (TFTE) modeling that was introduced in waveform interpolation speech coding area is used in this paper [5]. The TFTE can efficiently extract the voicing ratio in a unit of individual frequency bin by decomposing a single pitch based excitation signal into slowly varying and rapidly varying components. The slowly varying component of excitation (SVCE) which is a low-pass filtered version of the TFTE along the time-domain axis represents the periodic components of the excitation signal. The rapidly varying component of excitation (RVCE), on the other hand, represents the remaining noisy components. Thus, by controlling the ratio of SVCE and RVCE, the enhanced pathological voice becomes more natural voice. Note that the control mechanism is determined by taking a statistical analysis between pathological and normal voice so that the enhanced voice maintains the identity of the speaker.

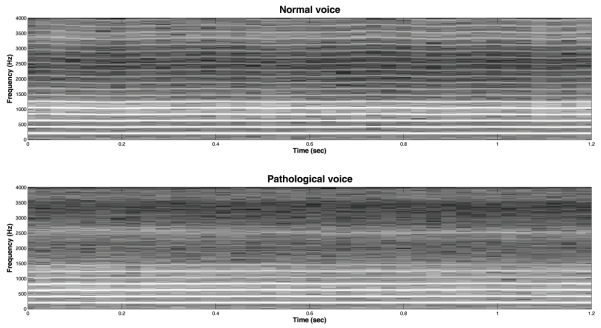
II. Characteristics of pathological voice

Fig. 1-(a) shows sustained vowel /a/. By comparing the pathological voice (bottom) from the normal voice (top), it can be observed that the harmonic structure of pathological voice is smeared especially at high frequency region. In general, this characteristic appears from the patients having functionless vocal folds, which results in breathy voice [4].

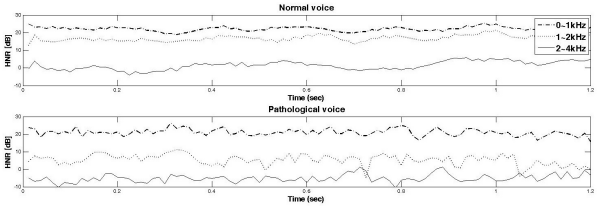
To measure a breathiness, harmonic-to-noise energy ratio (HNR) is often used as follows :

$$HNR\ [dB] = 20\log_{10}\left\{\frac{\sum_{k=k_i}^{k_l} ||S(k)| - |N(k)||}{\sum_{k=k_i}^{k_l} |N(k)|}\right\}, \quad (1)$$

where $S(k)$ is a short time Fourier transform (STFT) coefficient of original signal at k-th frequency bin, $N(k)$ is that of noise signal, and k_i, k_l denotes the initial and last frequency bin index of sub-band, respectively. In this case, the noise



(a) Spectrogram : normal (top) and pathological (bottom) voice.



(b) HNR : normal (top) and pathological (bottom) voice.

Fig. 1. Comparison between normal and pathological voice

signal $N(k)$ is obtained by bi-directional long-term prediction (LTP) method based on Betten's method [6]. Fig. 1-(b) shows the HNR measurement of normal and pathological voices. Note that the measurement is performed by several frequency bands to compare the breathiness into more detail.

From the Fig. 1-(b), the HNRs of low frequency band (0-1kHz) are somewhat similar. In the high frequency band (1-4kHz), however, the HNR of pathological voice is much smaller than that of normal voice. Another interesting observation is that the temporal variation of the HNR in the pathological voice is also greater than the one in the normal voice. This result also coincides with the analysis of [7].

Therefore, controlling the voicing ratio as well as reflecting the time-frequency characteristics is important to enhance the quality of pathological voice. To satisfy these requirements, one of the possible approaches is to use the TFTE modeling : it can represent the excitation signal on the time-frequency surface and efficiently control the voicing ratio by decomposing the excitation signal into the SVCE and RVCE.

III. Analysis of pathological voice using TFTE modeling

This section describes the characteristical differences between pathological and normal voice in the TFTE domain. The excitation signal of pathological voice is transformed into TFTE domain and it is compared with that of normal voice.

A. Time-frequency trajectory excitation

Let $u(n,\phi)$ denote a periodic function with ϕ extracted at the n-th frame, then the TFTE signal can be represented as follows :

$$u(n,\phi) = \sum_{k=1}^{P(n)/2} [A_k(n)\cos(k\phi) + B_k(n)\sin(k\phi)], \qquad (2)$$

where a phase term ϕ is defined as $\phi = \phi(m) = 2\pi m/P(n)$ with a pitch period $P(n)$, and the $A_k(n)$ and $B_k(n)$ are the k-th discrete time Fourier series (DTFS) coefficients of the single pitch based excitation signal [8].

To measure and control the voicing ratio, the TFTE signal is further decomposed into the SVCE and RVCE. In every frequency bin, the SVCE is obtained by applying a low-pass filter to the time-domain axis :

$$u_{SVCE}(n_i,\phi) = \sum_{m=1}^{M} h(m)u(n_{i-m},\phi), \qquad (3)$$

where $h(m)$ is the M-th order low-pass filter. The RVCE, on the other hand, is obtained by subtracting $u_{SVCE}(n,\phi)$ from $u(n,\phi)$:

$$u_{RVCE}(n,\phi) = u(n,\phi) - u_{SVCE}(n,\phi). \qquad (4)$$

It can be intuitively known that the energy of SVCE and RVCE reflects the voicing ratio in each frequency bin. Therefore, the decomposition of TFTE gives an advantage of delicately controlling the periodic part and noisy part of the excitation signal.

B. Analysis of pathological voice using TFTE modeling

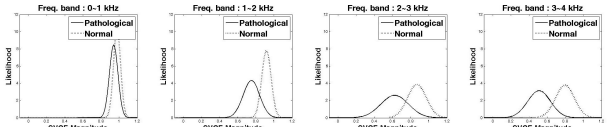
To analyze the pathological voice, the voicing ratio is measured using band-limited SVCE and RVCE magnitudes as follows :

$$SVCE_{mag} = |u_{SVCE}|, \qquad (5)$$

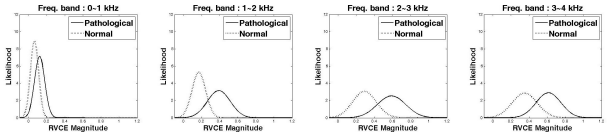
$$RVCE_{mag} = |u_{RVCE}|. \qquad (6)$$

To measure and compare the voicing ratio, the Korean vowel /a/ of 235 normal and 249 pathological voice samples are used. The pathological voice consists of sulcus, scarring, presbyphonia, and vocal fold paresis which are categorized as GI by Samsung Medical Center, Seoul, Korea.

In each frequency band, the distribution of the voicing ratio is modeled by Gaussian mixture model (GMM) with two mixtures. Fig. 2 depicts the voicing ratio distribution of $SVCE_{mag}$ and $RVCE_{mag}$. From the results, it can be observed that the pathological voice has similar characteristics to the normal voice in the low frequency band (0-1kHz). As the cut-off frequency becomes higher, however, the voicing characteristic of the pathological voice rapidly decreases compared



(a) $SVCE_{mag}$ distribution.



(b) $RVCE_{mag}$ distribution.

Fig. 2. Voicing ratio distribution between normal and pathological voice

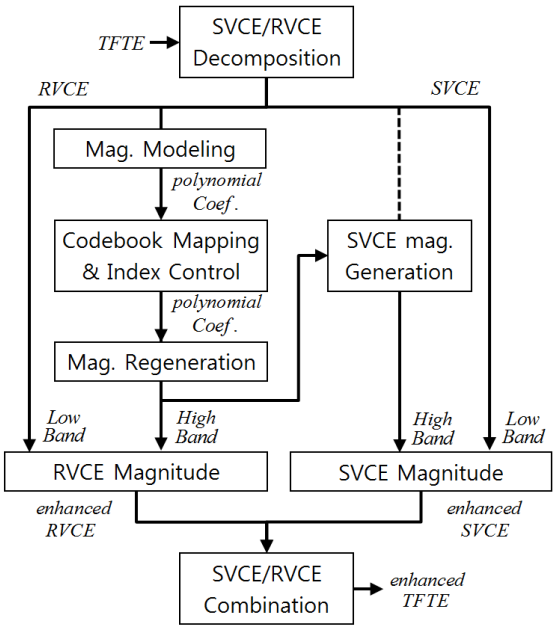


Fig. 3. Block diagram of proposed algorithm.

with that of the normal voice. From the observation, the next section describes how to enhance the high frequency band of the pathological voice while maintaining the characteristics of the low frequency band.

IV. Speech Enhancement for Pathological Voice using TFTE Modeling

This section describes the detailed algorithm to enhance pathological voice. The block diagram of the proposed algorithm is depicted in Fig. 3. To obtain the excitation signal, a 12-th order linear prediction (LP) analysis filter is applied to the input speech signal. The excitation signal is then transformed into the TFTE domain and decomposed into SVCE and RVCE. The low frequency band signal below 1 kHz is kept same because it is important to maintain the identity of the speaker. Note that its statistical characteristic is similar to normal voice. The remaining high frequency band signal, on the other hand, needs to be delicately controlled to enhance the periodic components while reducing the noisy components. After controlling, the enhanced SVCE and RVCE are combined again, then the output speech is obtained by passing the enhanced excitation signal to the LP synthesis filter.

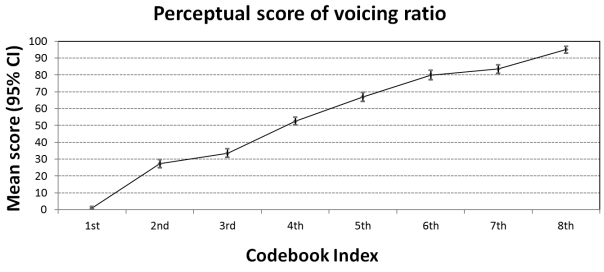


Fig. 4. Perceptual score of the voicing ratio depending on the codebook index.

A. SVCE and RVCE Modeling

Since the spectral envelope is important for determining the perceptual quality of the RVCE [8], the magnitudes of the RVCE are parameterized by the coefficients of the Legendre orthonormal polynomials [9], and they are mapped into the 8-th order codebook to make the control mechanism be easier. The codebook is modeled by the sets of the polynomial coefficients depending on the signal characteristics :

- Unvoiced case (1-st codebook)
- Transition case (2-nd, 3-rd codebook)
- Voiced case (4-th to 8-th codebook)

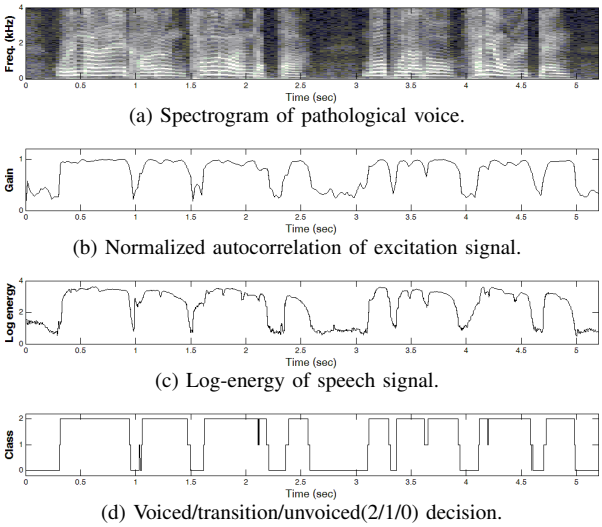
The codebook index is designed to represent how much the signal contains the periodic components. To verify the voicing ratio of the codebook perceptually, the /a/ voices of four females and four males are used for the test. Each speech is synthesized by the individual codebook index from one to eight. In the test, ten listeners are asked to make a judgment of the periodicity of the synthesized speech. Fig. 4 shows the mean score of the perceptual voicing ratio with its 95% confidence interval. Note that the perceived voicing ratio is proportional to the codebook index.

In the synthesis stage, the coefficients of the Legendre polynomial are obtained from the codebook and transformed into the RVCE magnitudes. The high frequency band (upper than 1 kHz) of the SVCE magnitude, on the other hand, is recovered by subtracting the RVCE magnitude from one in order to control the voicing ratio more efficiently. That is, if the ratio of RVCE magnitude is reduced, the portion of SVCE magnitude is increased automatically.

B. SVCE and RVCE Controlling

As mentioned above, the proposed controlling process focuses on the high frequency band signal while the original magnitudes of the SVCE and RVCE are used for the low frequency band.

In this paper, the high frequency band of pathological voice is divided into three classes such as unvoiced, transition, and



(a) Spectrogram of pathological voice.

(b) Normalized autocorrelation of excitation signal.

(c) Log-energy of speech signal.

(d) Voiced/transition/unvoiced(2/1/0) decision.

Fig. 5. Voicing ratio distribution between normal and pathological voice

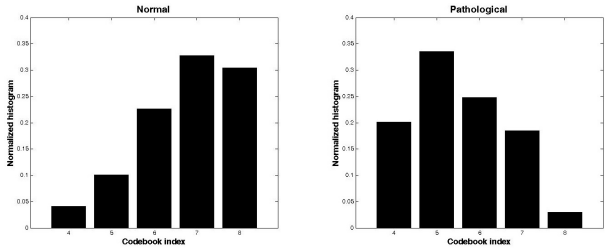


Fig. 6. Normalized histogram of codebook index in the voiced region : normal voice (left) and pathological voice (right).

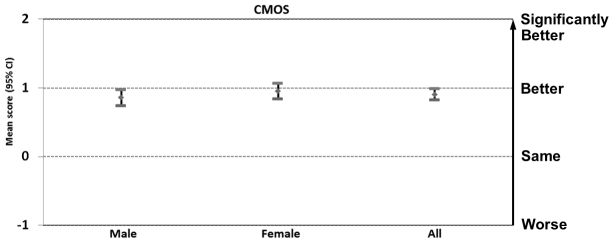


Fig. 7. Performance evaluation result

voiced case. Although more detail classification algorithm can be introduced, it is undesirable because many class candidates may bring unwanted distortion. The classification process can be achieved by using a normalized autocorrelation of the low-pass-filtered (0-1 kHz) excitation signal [10]. Note that the use of the normalized autocorrelation has an advantage that the proper detection is possible even in the pathological voice since the low frequency band signal has similar characteristics with normal voice. To provide more elaborate control, the log-energy of speech is also introduced. Fig. 5 depicts an example of the detection. Depending on the detection result, it is recommended to maintain or increase a little amount of the voicing ratio if the current SVCE and RVCE belong to the unvoiced or transition region.

In case of voiced region, on the other hand, the controlling can be efficiently achieved by increasing the codebook index since the index is proportional to the perceived voicing ratio. To determine the proper step of increasing index, the codebook indices in voiced region are collected. The recorded data with readings of a passage in Korean, which consists of 235 normal and 249 pathological voice samples from Samsung Medical Center, Seoul, Korea, are used. Fig. 6 shows a normalized histogram of codebook indices. From this result, the increasing step of codebook index is determined by two-steps since the indices of pathological voice are biased around four to six where that of normal voice are biased around six to eight.

V. Performance evaluation

This section discusses the performance of the proposed algorithm. The recorded pathological voices, which are composed of Korean sentence used in previous section, are also used for the test. The sampling frequency of each voice is set to 8 kHz. The test items consist of ten female and ten male pathological voice samples.

To evaluate the performance of the proposed algorithm, subjective listening tests are carried out. The comparison mean opinion score (CMOS) methodology is used for the test to verify the improvement of perceptual quality [11]. In the CMOS, the meaning of the zero point score is set to 'same' quality. In the test, twelve listeners are asked to make a quality judgment between the pathological and enhanced speech. Fig. 7 depicts the mean scores of CMOS test with its 95% confidence intervals. The result verifies that the proposed algorithm shows improvement of the perceived quality.

VI. Conclusions

In this paper, a speech enhancement algorithm for the pathological voice using the TFTE modeling was proposed. In order to efficiently control the periodic part and the noisy part, the TFTE was further decomposed into SVCE and RVCE. The statistical analysis between the pathological and normal voice was also included so that the proposed algorithm maintained the originality of the speaker. The result from subjective listening test verified the improvement of the perceived quality.

Acknowledgment

This research was supported by the Samsung Electronics. Advices related to analyzing the pathological voice were received by Jungwon Lee, Yonsei University, Seoul, Korea. 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] A. Kain, X. Niu, J. P. Hosom, Q. Miao, and J. van Santen, "Formant Re-Synthesis of Dysarthric Speech," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 25-30, Pittsburgh, Pa, USA, 2004.

[2] V. Lalitha, P. Prema, and L. mathew, "A Cepstrum Based Approach for Enhancement of Dysarthric Speech," in *International congress on Image and Signal Processing 2010*, vol. 7, pp. 3474-3478, 2010.

[3] H.R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of Normal Sounding Speech for Laryngectomy Patients through a Modified CELP Codec," *IEEE Trans. Biomed. Eng.*, vol.57, pp. 2448-2458, 2010.

[4] S. Hertegård, L. Hallén, C. Laurent, E. Lindström, K. Olofsson, P. Testad, and Å. Dahlqvist, "Cross-Linked Hyaluronan Used as Augmentation Substance for Treatment of Glottal Insufficiency: Safety Aspects and Vocal Fold Function," *Laryngoscope*, vol. 112, pp. 2211-2219, 2002.

[5] W. Kleijn, "Continuous Representations in Linear Predictive Coding," in *Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 215-230, 1991.

[6] F. Bettens, F. Grenez, and J. Schoentgen, "Estimation of Vocal Dysperiodicities in Disordered Connected Speech by Means of Distant-Sample Bidirectional Linear Predictive Analysis," *J. Acoust. Soc. Am*, vol.117, pp. 328-337, 2005.

[7] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated Speech Analysis Applied to Laryngeal Disease Categorization," *Computer Methods and Programs in Biomedicine*, vol.91, no.1, pp. 36-47, 2008.

[8] E. Choy, "Waveform Interpolation Speech Coder at 4kb/s," Master of Engineering, McGill Univ., Dept. Elect. Eng., Montreal, QC, Canada, 1998.

[9] E. Butkov, *Mathematical Physics*. Norwell, MA: Addision-Wesley, 1968.

[10] A. McCree and T. Barnwell, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, 1995.

[11] "Methods for subjective determination of transmission quality," ITU-T P.800, Geneva, 1996.