비고정 구간 길이 음향 튜브를 이용한 성도 모델링

논 문
59-6-19

Vocal Tract Modeling with Unfixed Sectionlength Acoustic Tubes(USLAT)

김 동 준*
(Dong-Jun Kim)

Abstract - Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The vocal tract is modeled as a chain of cylinders of varying cross-sectional area in linear prediction acoustic tube modeling. In this modeling the most common implementation assumes equal length of tube sections. Therefore, to model complex vocal tract shapes, a large number of tube sections are needed. This paper proposes a new vocal tract model with unfixed sectionlengths, which uses the reduced lattice filter for modeling the vocal tract. This model transforms the lattice filter to reduced structure and the Burg algorithm to modified version. When the conventional and the proposed models are implemented with the same order of linear prediction analysis, the proposed model can produce more accurate results than the conventional one. To implement a system within similar accuracy level, it may be possible to reduce the stages of the lattice filter structure. The proposed model produces the more similar vocal tract shape than the conventional one.

Key Words : Vocal Tract Modeling, Unfixed Sectionlength Acoustic Tubes(USLAT), Lattice Filter, Area Function

1. 서 론

Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The vocal tract is the most important component in the speech production process[1].

For the determination of the vocal tract shapes, X-ray techniques have been used as direct methods. Several indirect methods for computing the vocal tract area function from measured acoustic data that avoid the drawbacks of X-ray techniques have been suggested by many researchers[2-4]. There are two methods for acoustically estimating the vocal tract area function : (1) Linear prediction acoustic tube(LPAT) method and (2) Lip impulse resonance method. In LPAT method, which has many advantages, the vocal tract is modeled as a chain of cylinders of varying cross-sectional area. The most common implementation assumes equal length of tube sections. Therefore, to model complex vocal tract shapes, a large number of tube sections are required[5].

This paper proposes a new vocal tract model which is a kind of acoustic tube model with unfixed sectionlengths. The new vocal tract model is called unfixed sectionlength acoustic tube model, and the abbreviation is USLAT. To model the new vocal tract, a reduced lattice filter is designed. This reduced lattice filter needs to modify Burg algorithm. The USLAT vocal tract model is designed considering actual vocal tract shapes. Using this model, vocal tract area estimation is executed.

The advantages of the USLAT model are as follows. First, redundancy of conventional vocal tract model which use equal tube sections can be eliminated in this proposed model. Second, when the conventional and the USLAT models are implemented with the same order, the USLAT model can produce more accurate results than the conventional one. Third, to implement a system within similar accuracy level, it may be possible to reduce the stages of the lattice filter structure. And finally, the USLAT vocal tract model can coincide well with the speech production process.

* 정 회 원 : 청주대 공대 전자정보공학부 교수・공박
E-mail : djkim@cju.ac.kr
접수일자 : 2010년 4월 22일
최종완료 : 2010년 5월 18일

2. Vocal Tract Modeling

2.1. Unfixed Sectionlength Acoustic Tube Model

Fig. 1 shows the conventional acoustic tube model[2].

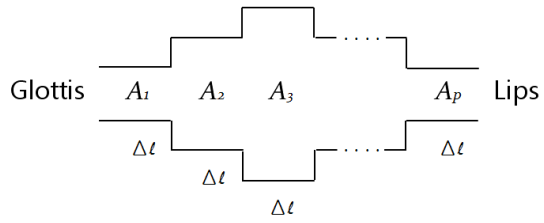


Fig. 1 Conventional acoustic tube model

In this figure, the left side corresponds to glottis and this is the excitation source of vocal tract filter. The right side is lips. A represent the vocal tract area of the relevant section and $\triangle l$ is sectionlength of this conventional acoustic tube model.

The proposed new vocal tract model is shown in Fig. 2.

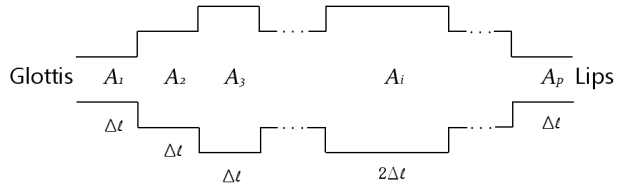


Fig. 2 Proposed acoustic tube model

This model is based on the actual vocal tract shapes. If any adjacent sections, i and $i+1$ have little difference in real vocal tract area, these two sections can be combined into a section with twice $\triangle l$. In this study, the USLAT is composed of eight sections with equal sectionlength and one section with twice sectionlength. So, the model order of this structure is nine.

2.2. Estimation of Vocal Tract Area Function

The sampling frequency and the number of sections is constrained by equation (1).

$$F_s = \frac{Mc}{2l} \tag{1}$$

Where, F_s, M, c and l means sampling frequency, linear prediction order, sound velocity(340m/sec) and vocal tract length, respectively. As long as the sampling frequency is constant, the vocal tract length is assumed to be fixed[6].

Some amount of pre-emphasis is necessary to eliminate the source characteristic and the lip radiation characteristic.

$$y[n] = s[n] - \mu s[n-1] \tag{2}$$

Where, $s[n]$, $y[n]$ represents raw speech signal, pre-emphasized signal and μ is pre-emphasis factor. The source characteristic is modeled by about -12dB/oct and the lip radiation characteristic is +6dB/oct. For steady-state vowels, the reasonable choice of μ is 1.

In the case of vowel-type speech generation in which the source is located at the glottis with no side branch, there exist two possibilities for boundary conditions at the lips and glottis for sound-wave propagation through the vocal tract. One was used by Nakajima and Wakita and the other was by Atal[2,3]. The Wakita's condition seems to give reasonable results in the form of area function.

The vocal tract area of i-th section is computed by equation (3) and (4).

$$\gamma_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i} \tag{3}$$

$$A_{i+1} = A_i \frac{1+\gamma_i}{1-\gamma_i} \tag{4}$$

In this equation, γ_i is reflection coefficients, and these coefficients are extracted by Linear Prediction(LP) analysis[6].

2.3. Modified Burg Algorithm

In the lattice section for delay 1, the p-th reflection coefficients, γ_p, the forward and backward errors, e_p^+, e_p^- can be computed by the equation (5) and (6)[6].

$$\gamma_p = \frac{2\sum_{n=p}^{N-1} e_{p-1}^+(n) e_{p-1}^-(n-1)}{\sum_{n=p}^{N-1} [e_{p-1}^+(n)^2 + e_{p-1}^-(n-1)^2]} \tag{5}$$

$$\begin{aligned} e_p^+(n) &= e_{p-1}^+(n) - \gamma_p e_{p-1}^-(n-1) \\ e_p^-(n) &= e_{p-1}^-(n-1) - \gamma_p e_{p-1}^+(n) \end{aligned} \tag{6}$$

But, in the combined section, the corresponding lattice section must have delay 2. So, conventional equations (5) and (6) are modified to equations (7) and (8).

$$\gamma_p = \frac{2\sum_{n=p}^{N-1} e_{p-1}^+(n) e_{p-1}^-(n-2)}{\sum_{n=p}^{N-1} [e_{p-1}^+(n)^2 + e_{p-1}^-(n-2)^2]} \tag{7}$$

$$\begin{aligned} e_p^+(n) &= e_{p-1}^+(n) - \gamma_p e_{p-1}^-(n-2) \\ e_p^-(n) &= e_{p-1}^-(n-2) - \gamma_p e_{p-1}^+(n) \end{aligned} \tag{8}$$

3. Experiments and Results

Fig. 3 shows the whole procedure of this speech analysis in block diagram.

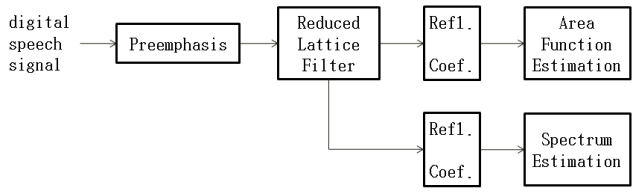
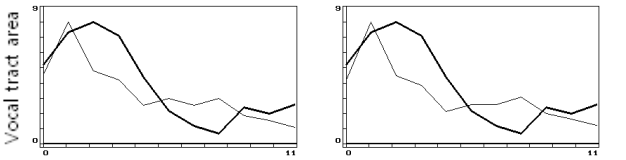


Fig. 3 Procedure of speech analysis

Five Korean vowels /a/, /e/, /i/, /o/ and /u/ are used. These speech data are filtered using low pass filter. The cutoff frequency of this low pass filter is about 4.7kHz, and the slope is -34dB/oct. These filtered speech is digitized in 10kHz sampling frequency.

Fig. 4 represents the estimated vocal tract area function for vowel /a/.



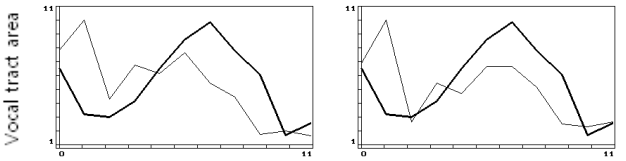
(a) Conventional model (order:10)
(b) Modified model (order:10)

Fig. 4 Vocal tract area function for vowel /a/

In this figure, the bold line is real X-ray data of Fant's experiments[1] and the fine line is the estimated data.

The Fig. (a) is the result of conventional acoustic tube model of order 10, and the Fig. (b) is the result of proposed model of order 9. In each figure, the left side corresponds to lips and the right is glottis.

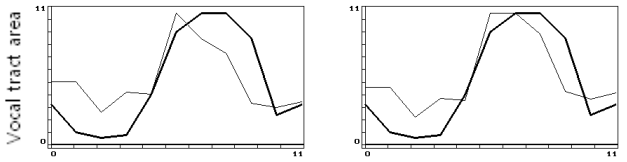
In Fig. 4, the modified model show the slightly better result in the back portion. But, in Fig. 5, the modified method shows more similar shapes than the conventional one.



(a) Conventional model (order:10)
(b) Modified model (order:10)

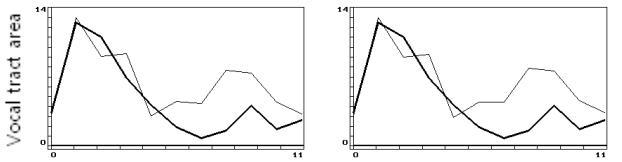
Fig. 5 Vocal tract area function for vowel /e/

Fig. 6 also produces the same results as the previous figures. In Fig. 7 and 8, though the order of the modified model is lower than the conventional one, the modified method shows nearly equal results with the conventional one.



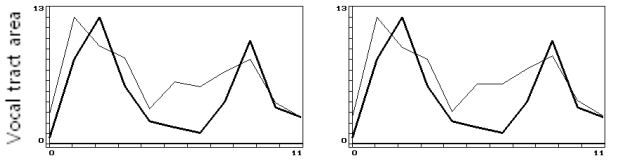
(a) Conventional model (order:10)
(b) Modified model (order:10)

Fig. 6 Vocal tract area function for vowel /i/



(a) Conventional model (order:10)
(b) Modified model (order:10)

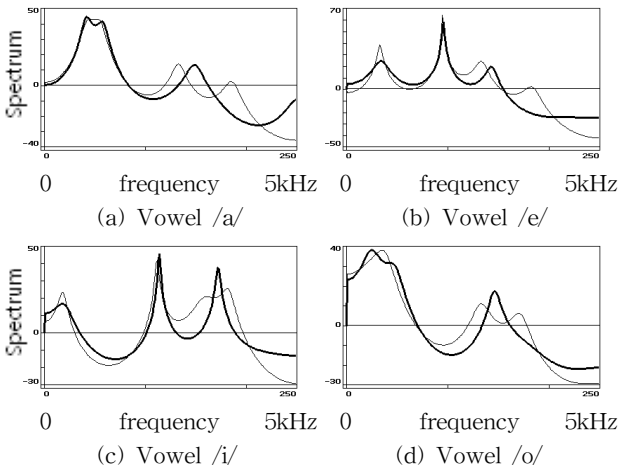
Fig. 7 Vocal tract area function for vowel /o/

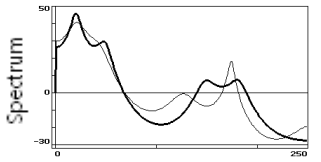


(a) Conventional model (order:10)
(b) Modified model (order:10)

Fig. 8 Vocal tract area function for vowel /u/

AR spectra with same order and with different order are estimated. The spectra for five vowels with same order are shown in Fig. 9. The fine line is the results of the conventional model and the bold line represents those of the USLAT model.





(e) Vowel /u/

Fig. 9 Spectra for five vowels with same order

In this figure, the first and the second formants of the conventional model are sometimes lumped together, but the USLAT model can divide them accurately.

The spectra with different order are shown in Fig. 10. Where the fine line is the results of the conventional model and the bold line represents those of the USLAT model.

Here, it can be found that the first and second formants of the USLAT model are as accurate as those of the conventional one, but the third and the forth are somewhat biased or lumped together.

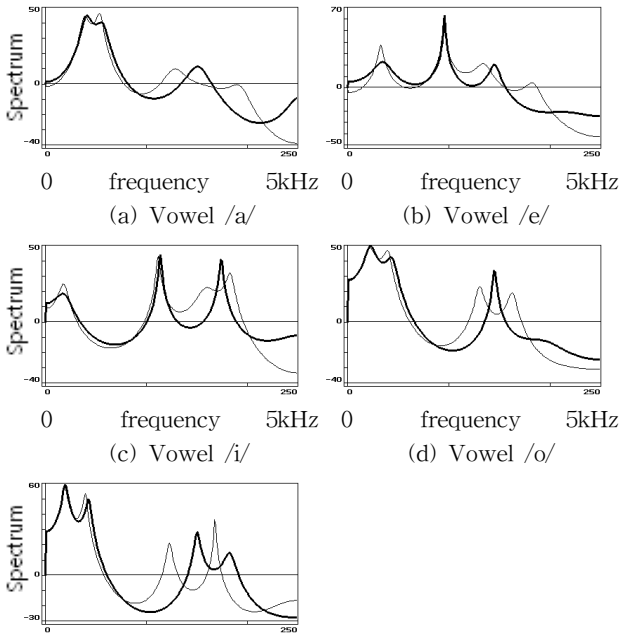


Fig. 10 Spectra for five vowels with different orders(USLAT : 9, conventional : 10)

4. Conclusion

In this study, a new vocal tract model with unfixed sectionlengths is proposed. And, using this model, speech analyses, such as, area function estimation and spectrum estimation are performed.

The obtained results are as follows.

First, the vocal tract area functions of the USLAT model of order 9 are more similar to real vocal tract areas than those of the conventional vocal tract model of order 10. Specially, in back vowels, outstanding similarity of the vocal tract area functions was found using the USLAT model. Second, in the AR spectra of each model order 9, the first and the second formants of the conventional model are sometimes lumped together, but the USLAT model can divide them accurately. Third, in the AR spectra of the USLAT model of order 9 and the conventional model of order 10, the first and the second formants of the USLAT model are as accurate as those of the conventional one, but the third and the forth are somewhat biarsed or lumped together. And finally, redundancy of conventional vocal tract model which use equal tube sections can be eliminated in this proposed model.

To implement a system for a certain degree of accuracy, it will be possible to reduce the stages for zero reflection coefficients of the filter structure.

참 고 문 헌

[1] G. Fant : Acoustic Theory of Speech Production, Mouton, 1970.

[2] H. Wakita, “Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms,” IEEE Trans. Acoust., Speech, Signal Processing, Vol. AU-21, No. 5, Oct. 1973.

[3] B. S. Atal, “Speech analysis and synthesis by linear prediction of speech wave,” J. Acoust. Soc. Am, Vol. 41, pp. 65(A), 1970.

[4] J. Schroter, J. N. Larar, and M. M. Sondhi, “Speech Parameter Estimation using a Vocal Tract/Cord Model,” IEEE Int. Conf. on Acoustics. Speech. and Signal Processing, pp. 308-311, 1987.

[5] J. D. Markel, A. H. Gray : Linear Prediction of Speech, Springer-Verlag·Berlin·Heidelberg·New York, 1976.

[6] T. F. Quatieri : Discrete-Time Speech Signal Processing, Principles and Practice, Prentice Hall, 2002.

[7] E.P. Neuburg, W.R. Bauer, “On the Source-Filter Model of the Vocal Tract,” IEEE Int. Conf. on Acoustics. Speech. and Signal Processing, pp. 1609-1612, 1986.

[8] H. Fusisaki, M. Ljungqvist, “Estimation of Voice Source and Vocal Tract Parameters Based on ARMA Analysis and a Model for the Glottal Source Waveform,” IEEE Int. Conf. on Acoustics. Speech. and Signal Processing, pp. 637-640, 1987.

[9] A. M. de L. Araújo, F. Violaro, “Formant Frequency Estimation Using a MEL Scale LPC Algorithm,” IEEE Int. Conf. on Acoustics. Speech. and Signal Processing, pp. 207-212, 1998.

저 자 소 개

김 동 준 (金 東 浚)

1963년 4월 14일 생. 1988년 연세대학교 전기공학과 졸업. 1990년 동 대학원 전기공학과 졸업(석사). 1994년 동 대학원 전기공학과 졸업(공박). 현재 청주대학교 이공대학 전자정보공학부 교수

Tel : 043-229-8460

Fax : 043-229-8460

E-mail : djkim@cju.ac.kr