DEVELOPMENT OF SPEECH TRAINING SYSTEM
FOR THE HARD OF HEARING PERSON
BASED ON VOICE SYNTHESIS TECHNIQUE
USING VOCAL TRACT AREA FUNCTION

Takefumi KITAYAMA*, Hiroyuki KAMATA* and Yoshihisa ISHIDA*

*School of Science & Technology, Meiji University
1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214, JAPAN
Tel: +81-(44) 934-7307
Fax: +81-(44) 934-7312
E-Mail: kamata@isc.meiji.ac.jp
E-Mail: ishida@isc.meiji.ac.jp

ABSTRACT - In our laboratory, we have developed a speech training system called "Speech Trainer" for the deaf. This system has presented many practice items for hearing impaired people. The training results are displayed on the CRT display of personal computer. Therefore, trainers can confirm the characteristics of voice of themselves visually.
Decides, if the hearing impaired people have the ability to hear a part of frequency band, the voice synthesis technique can be adopted to practice on the speaking. In this paper, we insist the effective educational method of speech training for the hard of hearing person using the audience of the synthesis sound adopting the technique to shift the formants and the comparative training of vocal tract shape at the same time.

INTRODUCTION

For over past twenty years, we have developed a speech training system for the deaf children in our laboratory. The system is called "Speech Trainer" and has been practically used in the deaf school in Japan. The system has presented many practice items to hearing impaired people as follows: (1) vowel and consonant training using vocal tract shape, (2) intonation and stress accents training, (3) syllable and word training based on the transitions of spectra and fundamental frequency. In the analysis of voice, DSP (TMS320C25 or TMS320C32) is used for realizing the real time processing and for obtaining the variety and the expandability. The estimated results are displayed on the CRT display of personal computer. Therefore, trainers can confirm the characteristics of voice of themselves visually.
Besides, we have developed the speech training system as a tool for the teacher of deaf school. However, the hearing impaired people can not always train under teacher's guidance. Then, the system has been added the self-teaching and self-confirmation functions using techniques of voice recognition since last year. As we explain above, we have developed the system using voice analysis techniques (T.KITAYAMA, F.SUGANO, H.KAMATA & Y.ISHIDA (1995)).
However, if the hearing impaired people have the ability to hear a part of frequency band, the voice synthesis technique can be adopted to practice on the speaking. So we think that the effective educational method of speech training can be realized using the audience of the synthesis sound which formants are converted the low frequency band and the comparative training of vocal tract shape using vocal tract area function at the same time. Usually, the frequency band that the hard of hearing people can hear is low limited frequency area (less than 1[kHz]). The conventional digital hearing aid has tried to shift the all of voice information into the low frequency area based on the spectrum conversion technique and so on. In this paper, we propose a new technique to shift the formants using the result of vocal tract area function estimation, and the technique and the comparative training of vocal tract shape previously developed are applied for the effective educational method of speech training for the hard of hearing person.

CONCEPT

Ordinarily, the tone of adults' voice is lower than the children's voice. This reason is considered as follows: difference of vibration of glottis, and difference of the length of vocal tract (Figure 1): the length of vocal tract decides the placement of formants (If the length is long, the frequency of all formants become low frequency band). Based on this concept, we try to convert the formant frequency into low frequency area, and explain the theoretical method for getting the natural synthesis sound (the voice synthesis algorithm is shown in Figure 2).

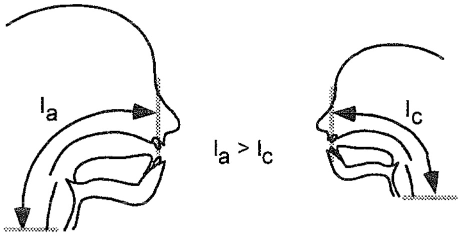

Figure 1. Difference of the length of vocal tract

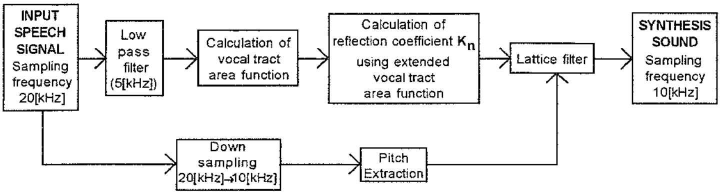
INPUT SPEECH SIGNAL Sampling frequency 20[kHz]
Low pass filter (5[kHz])
Calculation of vocal tract area function
Calculation of reflection coefficient K_n using extended vocal tract area function
Lattice filter
SYNTHESIS SOUND Sampling frequency 10[kHz]
Down sampling 20[kHz]→10[kHz]
Pitch Extraction

Figure 2. Voice synthesis algorithm

THE ESTIMATION OF VOCAL TRACT AREA FUNCTION

The estimation method of vocal tract area function is shown in Figure 3 (H.KAMATA, Y.ISHIDA & Y.OGAWA (1980)). $\mathbf{A}_n$ is vocal tract area function and **M** is the analysis order.
The estimation method of vocal tract area function is based on Levinson-Durbin algorithm. As a pre-processing, the adaptive inverse filter is necessity for excluding radiation and glottis characteristics. We use the adaptive inverse filter for the auto-correlation function although the conventional adaptive inverse filter deals with time series signal. In this method, the auto-correlation function **R(i)** is given from time series signal and is renewed by (1). Factor ε_n in (1) is given by (2).

$$R_{new}(i)=(1+\varepsilon_n^2)R_{old}(i)-\varepsilon_n\{R_{old}(i+r)+R_{old}(|i-r|)\} \qquad (1)$$

$$\varepsilon_n=\frac{R_{old}(r)}{R_{old}(0)} \qquad (2)$$

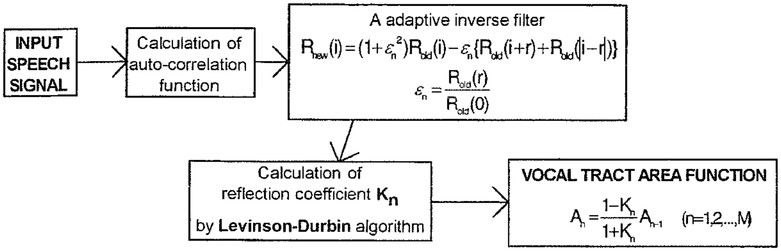
INPUT SPEECH SIGNAL
Calculation of auto-correlation function
A adaptive inverse filter
$R_{new}(i)=(1+\varepsilon_n^2)R_{old}(i)-\varepsilon_n\{R_{old}(i+r)+R_{old}(|i-r|)\}$
$\varepsilon_n=\frac{R_{old}(r)}{R_{old}(0)}$
Calculation of reflection coefficient K_n by Levinson-Durbin algorithm
VOCAL TRACT AREA FUNCTION
$A_n=\frac{1-K_n}{1+K_n}A_{n-1} \quad (n=1,2,...,M)$

Figure 3. Estimation method of vocal tract area function

THE LENGTH OF VOCAL TRACT

The result of vocal tract area function estimation is shown in Figure 4. In Figure 4, distance **l** between points of vocal tract area function is calculated by (3).

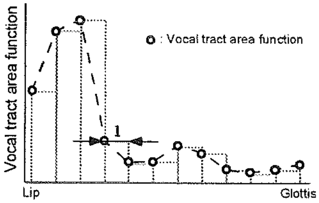

Figure 4. Vocal tract area function estimation

$$l = \frac{c}{2f_s} \qquad (3)$$

C :Speed of sound ($\cong 340$[m / sec])

f_s :Sampling frequency

The length of vocal tract of adult is known as about 17[cm], and the first and second formants which can be considered as phonemic information in voices generally located below about 5[kHz]. Therefore, we have researched the estimation of vocal tract area function using the signal which is sampled 10[kHz] and the 12th analysis order. Under this condition, distance **l** in figure 3 is calculated about 1.7[cm].

While, in order to the concept that the length of vocal tract affects the placement of all formants is adopted, the length of vocal tract is extended when the voice is synthesized.

EXTENSION OF THE LENGTH OF VOCAL TRACT

In this paper, we adopt the lattice filter, because the parameters k_n of this filter which are very important are decided by the vocal tract area function estimation. Therefore, when a voice is synthesized by using the vocal tract area function which is analyzed with such lower order (12th), the synthesized speech sounds unnatural. The one of the reason why it sounds unnatural, is that the amount of information containing in vocal tract area function is restricted. Accordingly, the positions in which the vocal tract area function can be obtained are fixed (like in Figure 5-a). Even if the vocal tract area function is interpolated by some kinds of nonlinear functions (e.g. spline, lagrange), the quality of the synthesized speech can not be much improved, because the interpolated value of vocal tract area function is not always correct value. In order to obtain nearly twice bits of information than usual, we change the sampling frequency from 10[kHz] to 20[kHz]. Therefore, distance **l** is shorten from about 1.7[cm] to about 0.85[cm] and the analysis order is extended from 12th to 23rd. It means that the number of positions mentioned above can be increased twice. (like in Figure 5-b).

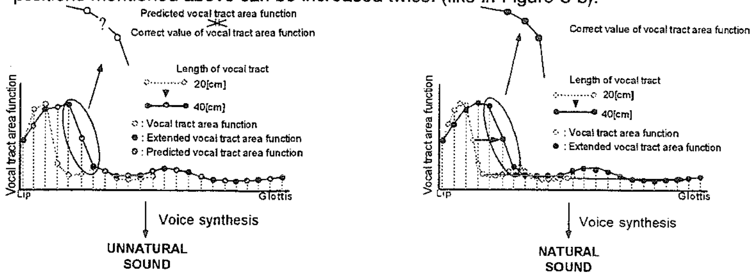
Predicted vocal tract area function
?
Correct value of vocal tract area function
Length of vocal tract
20[cm]
40[cm]
: Vocal tract area function
: Extended vocal tract area function
: Predicted vocal tract area function
Vocal tract area function
Lip
Glottis
Voice synthesis
UNNATURAL SOUND
Correct value of vocal tract area function
Length of vocal tract
20[cm]
40[cm]
: Vocal tract area function
: Extended vocal tract area function
Vocal tract area function
Lip
Glottis
Voice synthesis
NATURAL SOUND

Figure 5-a. Sampling frequency 10[kHz] Figure 5-b. Sampling frequency 20[kHz]

Figure 5. Extension of length of vocal tract

PITCH EXTRACTION

Pitch extraction algorithm is shown in Figure 6. In this paper, sampling frequency of the synthesis sound is 10[kHz]. Therefore, when the pitch of input speech signal is extracted, as the pre-processing, the method of down sampling from 20[kHz] to 10[kHz] is needed. This algorithm is not to obtain the phase equalized residual signal, but to find out the excitation points. This is achieved by using points are defined as the positions where the auto-correlation of the residual signal becomes maximum. The excitation points obtained by this method are equivalent to the local peak positions of the phase equalized residual signal.

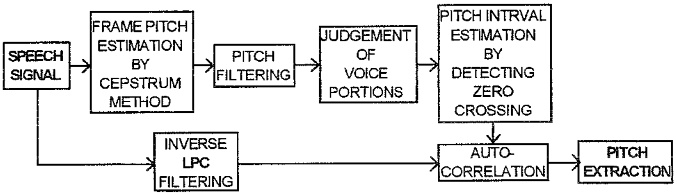
SPEECH SIGNAL
FRAME PITCH ESTIMATION BY CEPSTRUM METHOD
PITCH FILTERING
JUDGEMENT OF VOICE PORTIONS
PITCH INTRVAL ESTIMATION BY DETECTING ZERO CROSSING
INVERSE LPC FILTERING
AUTO-CORRELATION
PITCH EXTRACTION

Figure 6. The algorithm of pitch extraction

SPEECH SYNTHESIS USING LATTICE FILTER

In this paper, we use the lattice filter (J.D.Markel, A.H.Gray, Jr & H.Suzuki 1980) with the result of vocal tract area function estimation for voice synthesis. The detailed structure of the lattice filter is shown in Figure 7.

It is known that the parameter $\mathbf{k}_n$ in Figure 7 is equal the reflection coefficient $\mathbf{K}_n$ in Figure 2.

Therefore, in the voice synthesis, factor $\mathbf{k}_n$ is calculated using the extended vocal tract area function defined in (4). In the each stage, input $\mathbf{E}_n$ is renewed by (5) and (6).

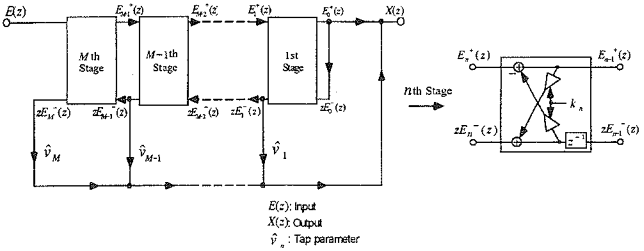

Figure 7. Detailed structure of lattice filter

$$k_n = \frac{A'_{n-1} - A'_n}{A'_{n-1} + A'_n} \quad (n = 1,2,\ldots,M') \ , \ \ A'_0 = 1 \qquad (4)$$

A'_n : Extended vocal tract area function

M' : Extended analysis order

$$E_{n-1}^{+}(z) = E_n^{+}(z) - k_n E_{n-1}^{-}(z) \qquad (5)$$

$$zE_n^{-}(z) = k_n E_{n-1}^{+}(z) + E_{n-1}^{-}(z) \qquad (6)$$

In Figure 7, the factor $\hat{v}_n$ called tap parameter is used to produce the high quality synthesis sound . And, when the transfer function of vocal tract is stable, the absolute value of reflection coefficient $\mathbf{K}_n$ is smaller than one. Therefore, a DSP with fixed point calculation capability is enough to perform this filter.

RESULT

The example of voice synthesis using the proposed method (Japanese vowel /a/) is shown in Figure 8. In this paper, the length of vocal tract is extended from about 20[cm] to 40[cm], thus the stage of lattice filter is increased from twelve to twenty-three.

As the Figure 8 shows, all formants are shifted into low frequency band. Actually, when the synthesis sound is heard, it is understood that the tone quality changed (convert the natural speech into low voice). We plan to experiment with some people who are hard to hear. We hope they can hear such voices more comprehensive than normal voices and the speech training becomes effective further more for the hard of hearing person.

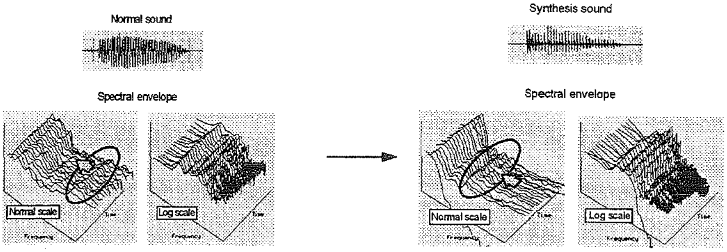

Figure 8. Example of voice synthesis
(Japanese vowel /a/)

CONCLUSIONS

In this paper, we propose the new educational system of speech training for the hard of hearing person (Figure 9). We believe that this system encourage a lot of the hearing impaired people. Furthermore, we have a project to achieve the real-time processing with DSP.

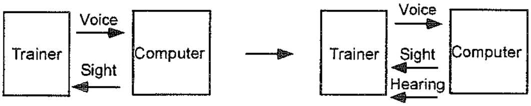

Figure 9. Educational speech training system

REFERENCES

H.Kamata, Y.Ishida & Y.Ogawa "The Fast Estimating of the Vocal Tract Area Using DSP" IEEJ, Vol. 110-7,pp.773-780 (1980)

J.D.Markel, A.H.Gray,Jr & H,Suzuki "Liner Prediction of Speech" CORONA PUBLISHING CO., LTD (1980).

T.Kitayama, F,Sugano, H.Kamata & Y.Ishida "Consonant Training System for Hearing-Impaired Children Using Vocal Tract Area Function" ISCPAT'95, Vol.1, pp.253-257 (1995)