6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Articulatory Synthesis Using a Vocal-Tract Model of Variable Length

Zhenli Yu[1,2], Shangcui Zeng[1]

1. Dept. of Information and Electronic Engineering, Zhejiang University
Tian Mu Shan Rd. 34, Hangzhou 310028, China; zlyu@mail.hz.zj.cn

2. Motorola China Research Centre
3/F, 16 Henan Rd. S., Shanghai 200002, China; a16300@email.mot.com

ABSTRACT

A method of articulatory synthesis using a vocal-tract model with variable length is proposed. The vocal-tract length is derived prior to the unique determination of vocal-tract area parameters incorporated with a codebook that maps formants to vocal-tract length is used. A two-dimensional interpolation function for irregularly spaced data is conducted to confine vocal-tract length in the first and second formant plain to generate the codebook. The reflection-type line analog (RTLA) model of articulatory synthesis is employed to generate speech sound. Variable vocal-tract length yields the issue of discrete-time implementation with multi-rate sampling of the RTLA model. Multi-rate sampling conversion is conducted.

1. INTRODUCTION

The method of articulatory synthesis using a vocal-tract model "Band-limited Fourier Cosine Expansion of the logarithmic area function" (abbreviated LogBLFCE) has been previously proposed [1, 2, 3], In that proposal, vocal-tract length is derived associated with a root-cell codebook technique. The root-cell codebook is that the vocal-tract is constrained against $F_1 - F_2$ (first and second formant) plain by 2-order curve fitting. One limitation of the curve fitting is that the difficulty of solving the equations for determining the curve parameters will increase as the number of vowels in a corpus increase.

The entire profile of the work is shown in Figure 1. Formant frequencies are used as the acoustic target of synthesis. Time-variant vocal-tract length and area function are derived from formant trajectories incorporated with the codebook looking up and vocal-tract inverse procedure. In the codebook generation, vocal-tract length is confined by the method presented in this paper. Sound wave output at lip end is generated with the reflection-type line analog (RTLA) model [4] and pitch-synchronous pulse waveform excitation [5]. Synthetic speech signal is finally obtained after multi-rate sampling conversion.

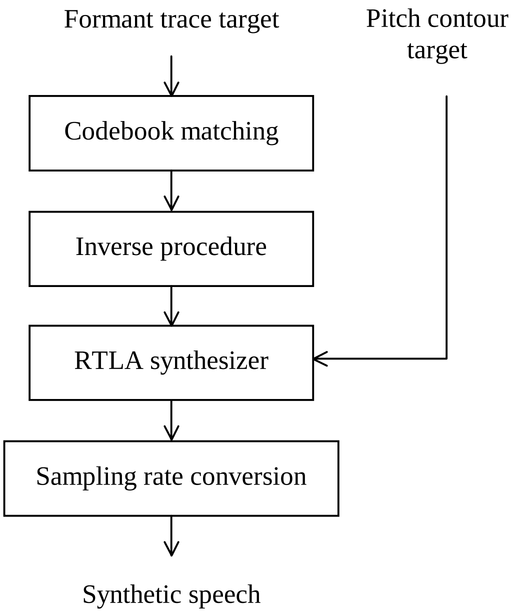

Figure 1 Profile of articulatory synthesizer

Vocal-tract is considered with piecewise constant area and variable length. The area function is represented by

$$Log[A(i)] = Log[A_0] + \sum_{k=1}^{2N} a_k \cdot \cos(k\pi \frac{i \cdot x_0}{L}), \quad i = 1, \ldots M \tag{1}$$

Where L is the total length of the vocal-tract, x_0 is the length of section tube that has piecewise constant area. M is the number of section tubes. As L is time-variant and M should be a constant for the convenience of digital implementation, $x_0 = L(t)/M$ is time-variant, too. A_0 is the uniform area of neutral vocal-tract. $\{a_k\}$ is the set of coefficients of Fourier cosine expansion of logarithm area. Since L is time variable

[1] Granted by NSF of China (Grant Code: 69972046)

and M is fixed, x_0 is changed against time. Thus, both $\{a_k\}$ and L are the parameters representing vocal-tract area function. To determine the vocal-tract area, an inverse issue that maps formant information of speech signal to vocal-tract area parameters is essential. Given formant frequencies of the voiced part of speech signal, $\{F_k\}$, area parameters is determined through a perturbation procedure [6].

In case of time variable vocal-tract model is adopted, the perturbation procedure needs input of vocal-tract length L determined a prior. Speech signal, particularly formant traces, however, cannot provide straightforward information of vocal-tract length. A 2-dorder curve surface fitting of vocal-tract length along F_1 and F_2 is used to generate a codebook that constraints the mapping of formants to vocal-tract length. However, that method is limited in a special formant corpus of the six Russia vowels from Fant [7].

This paper presents a method of interpolation of vocal-tract length against first and second formants with the two-dimensional interpolation function for irregularly spaced data of corpus with arbitrary amount of vowels. The vocal-tract length constraint is detailed in section 2. RTLA synthesizer will be described in section 3. Discussion and conclusion will be given in section 4.

2. VOCAL-TRACT LENGTH CONSTRAINTS VS FORMANTS

The method of two-dimensional interpolation function for irregularly spaced data [8] is adopted to constrain vocal-tract length as function on $F_1 - F_2$ plane, where F_1 and F_2 are the first and second formant frequencies.

We have finite irregularly spaced data points $\{ D_i, i = 1 \sim N_r \}$ from which a 'smooth' (continuous and once differentiable) function is to be interpolated. Each data point D_i is represented by a triplet (x_i, y_i, z_i), where x_i , y_i are the locational coordinates, and z_i is the value of D_i . An interpolation function $z = f(x, y)$ to assign a value to any location $P(x, y)$ in the plane is sought. The interpolation function is:

$$f(P) = \begin{cases} \left(\sum_{i=1}^{N_r} w_i \cdot z_i\right) \Big/ \left(\sum_{i=1}^{N_r} w_i\right) & if \;\; d_i \neq 0 \;\; for \; all \; D_i \\ z_i & if \;\; d_i = 0 \;\; for \; some \; D_i \end{cases} \tag{2}$$

where

$$w_i = (s_i)^2 \times (1 + t_i) \tag{3}$$

$$(s_i)^2 = \frac{1}{(x - x_i)^2 + (y - y_i)^2} \tag{4}$$

$$t_i = \{ \sum_{j=1}^{N_r} s_j \cdot [1 - \cos(D_i P D_j)] \} \Big/ (\sum_{j=1}^{N_r} s_j) \tag{5}$$

$$\cos(D_i P D_j) = \frac{(x - x_i) \cdot (x - x_j) + (y - y_i) \cdot (y - y_j)}{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{6}$$

Let (x_i, y_i, z_i) express the first two formants and vocal-tract length, viz., ($F_{1,i}, F_{2,i}, L_i$), of a reference vowel, the above formulae give an interpolation function of vocal-tract length against $F_1 - F_2$ plane. This interpolation function $L = f(F_1, F_2)$ can be used as vocal-tract length constraint. Figure 1 depicts an example of the vocal-tract length constraint on the $F_1 - F_2$ plane for the nine English vowels. Since the above method is not in forms of analytic functional formula, a table lookup approach is used in practice. Thus, a codebook constraining formant frequency and vocal-tract length is needed. The principle to use the vocal-tract length constraint of the codebook is that if the variables L , $f_p(1)$ and $f_p(2)$ of a vector is required to satisfy

$$\left| Z(f_p(1), f_p(2)) - L \right| \leq \sigma = 0.5 \cdot \Delta L_0 \tag{7}$$

The value of the tolerance σ depends on the quantization of the vocal tract length, and it is made to satisfy the following condition. Where ΔL_0 is the quantization level of the vocal-tract length in the codebook.

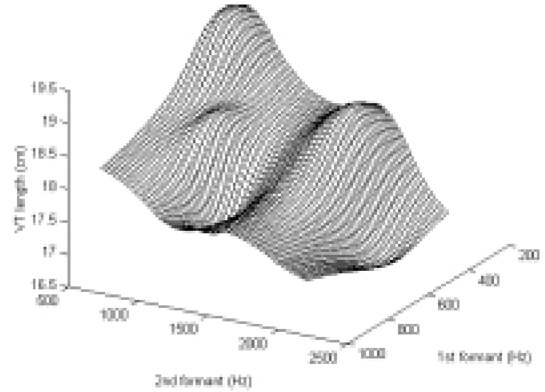

Figure 2 Vocal-tract length constraints against formants for six Russian vowels corpus

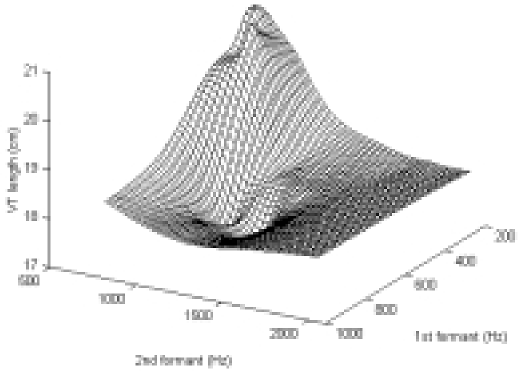

Figure 3 Vocal-tract length constraints against formants for nine English vowels corpus

3. SYNTHESIZER WITH VARIABLE VOCAL-TRACT LENGTH

RTLA model is a simple and efficient realization of articulatory synthesizer. In this model, speech generation is modeled as partial waves scattering inside the vocal-tract. The continuities of pressure and velocity of the underling waves are employed for digital implementation of synthesis [4]. Various vocal-tract losses are accounted in the model. Figure 4 shows the picture of partial wave scattering at the joint of two successive section tubes as an example.

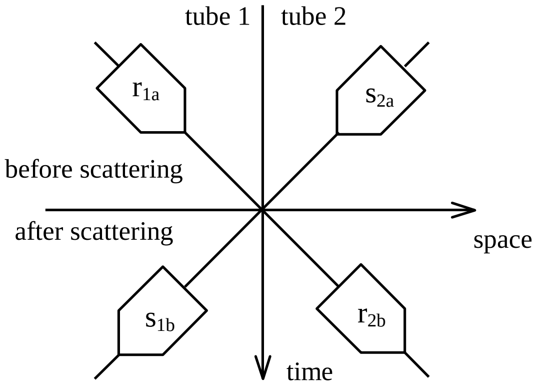

Figure 4 Partial wave scattering at the joint of successive sections

In Figure 4, r_{1a} and s_{2a} is the forward and backward partial waves reaching the joint of two tube before scattering happens. The scattering results in the forward partial wave r_{2b} and backward partial wave s_{1b}. The reflection coefficients of the scattering, k_{12}, is determined by the cross-sectional area functions, A_1 and A_2, of the successive section tubes,

$$k_{12} = (A_2 - A_1)/(A_2 + A_1) \tag{8}$$

The losses of vocal-tract is accounted for with a factor in forms of

$$\gamma_1 = 1 - 0.006 \cdot x_0 / \sqrt{A_1} \tag{9}$$

The resultant partial waves are derived as

$$\begin{cases} r_{2b} = \gamma_1 \cdot (1 + k_{12}) \cdot r_{1a} + k_{12} \cdot s_{2a} \\ s_{1b} = \gamma_1 \cdot [-k_{12} \cdot \gamma_1 \cdot r_{1a} + (1 - k_{12}) \cdot s_{2a}] \end{cases} \tag{10}$$

Figure 5 is the digital implementation of the scattering.

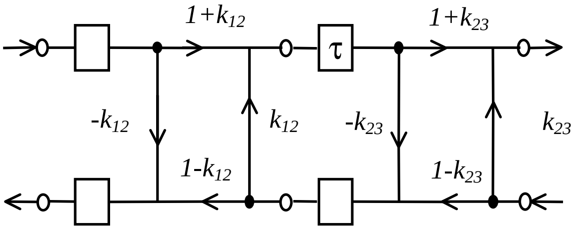

Figure 5 Discrete-time implementation of RTLA

A special issue to be considered for the realization of the RTLA synthesizer with variable vocal-tract length is the multi-rate sampling of synthesizer. According to the theorem of RTLA model, the sample interval of the discrete-time implementation should be $T_s = 2 \cdot x_0 / c$. Since the length of each tube x_0 varies against time, the model has to work with a varying T_s, which means a varying system sampling frequency $F_i = 1/T_s$. To obtain output signals with a constant sampling rate F_o, it is necessary to convert the multi-sampling rate signal at the terminal of the system into a constant sample rate. Also, since VT length varies continuously, the frequency conversion factor $F_i / F_o = c/(2 \cdot x_0 \cdot F_o)$ is not a constant rational number. Hence, the sampling conversion system must be time-variant. A time-variant filter [9] is employed to deal with this issue. Denoting F_i the time-varying sampling frequency of the input signal $x(n)$, F_o the sampling frequency of the output signal $y(m)$, and F_c a temporal sampling frequency that satisfies $F_c < F_i / 2$ and $F_c < F_o / 2$, the formula of the filter in discrete-time domain is

$$y(m) = \frac{F_c}{2F_i} \cdot \sum_{n=N_1}^{N_2} x(n) \cdot w(mT_o - nT_i) \cdot \frac{\sin[2\pi F_c (mT_o - nT_i)]}{2\pi F_c (mT_o - nT_i)} \tag{8}$$

where T_o is the reciprocal of F_o. $w(mT_o - nT_i)$ is a windowing function of the time variant filter. The window is symmetric around the original point of $x(n)$, and several window types can be used, for instance, "Rectangular" or "Hamming", etc. N_1 and N_2 determine the size of the window. Wu, Badin *et al* [10] have showed that a rectangular window with 4-5 points is optimal for the sampling conversion. Figure 6 illustrates how to determine the values of N_1 and N_2 . The values of F_o and F_c are set to be 10 kHz and 8 kHz, respectively.

Limited experiments have shown that the proposal is promising in formant copy synthesis and formant mimic synthesis.

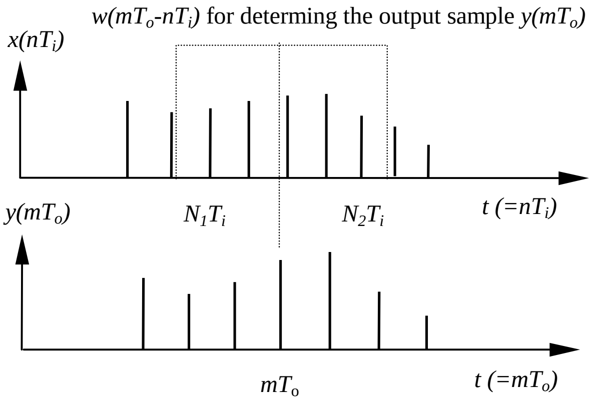

Figure 6 Determination of the window size of IIR filter

4. DISCUSSION AND CONCLUSION

A method of articulatory synthesis using a vocal-tract model with variable length is proposed. The vocal-tract length is derived prior to the unique determination of vocal-tract parameters from formant frequencies. The two-dimensional interpolation function for irregularly spaced data is conducted to confine vocal-tract length in the first and second formant plain. The reflection-type line analog (RTLA) model of articulatory synthesis is employed. Variable vocal-tract length yields the issue of discrete-time implementation with multi-rate sampling of the RTLA model. Multi-rate sampling conversion is conducted.

The proposal has features of producing natural synthetic speech output and enabling voice modification in speech synthesis of text-to-speech technology. Firstly, variable vocal-tract length model simulates the speech production mechanism better than fixed length approach since in real world vocal-tract length varies dependent on what is said and who says. Secondly, the method of two-dimensional interpolation function for irregularly spaced data can be adopted for corpus with arbitrary amount of vowels. Thirdly, RTLA synthesis model is of simple discrete-time implementation and gives natural output speech. Envisioning future embedded TTS technology, speech production based synthesizer has potential applications because it uses little amount of parameters to represent speech information. Also, as formant and pitch contour are used to represent acoustic parameter of speech target voice timbre can be modified through articulatory synthesizer.

REFERENCES

[1]. Yu Z.L. and Ching P.C., “Acoustically and geometrically optimised codebook for unique mapping from formants to VT shape,” *EUROSPEECH’97*, Vol.1, pp.235-238, 1997

[2]. Yu Z.L. and Ching P.C., “Articulatory synthesis of formant targeted sounds with parameters derived from the inverse solution of speech production,” *IEEE ICASSP’98*, vol. II, pp.889-892, 1998

[3]. Yu Z.L. and Ching P.C., “A synthesis method based on speech production and articulatory model,” *Chinese Journal of Acoustics (in English)*, Vol.19, No.2, pp.128-141, 2000

[4]. Liljencrants, J., *Speech synthesis with a reflection-type line analog*, Ph.D. Thesis, Royal Institute of Technology (KTH), Stockholm

[5]. Rosenberg, A. E., "Effect of pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, Vol.49, No.2, pp.583-591, 1971

[6]. Yu Z.L. and Ching P.C., “Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method,” *IEEE ICASSP’96*, Vol.1, pp.786-789, 1996

[7]. Fant, G., *Acoustic theory of speech production*, the Huge: Mouton (2nd edition), 1970

[8]. Shepard, D., "A two-dimensional interpolation function for irregularly-spaced data," *Proc. of 1968 ACM National Conference*, pp.517-524, 1968

[9]. Crochiere, R. E., and Rabiner, L. R., *Multirate digital signal processing*, New Jersey: Prentice-Hall, 1983

[10]. Wu, H. Y., Badin, P., Cheng, Y. M., and Guerin, B., “Vocal tract simulation: implementation of continuous variations of the length in a Kelly-Lochbaum model, effects of area function spatial sampling,” *IEEE ICASSP’87*, pp.9-12, 1987