ISSN 1009 - 3095 *Journal of Zhejiang University(SCIENCE)* V.1, No.3, P.388 - 393, July - Sep., 2000
http://www.chinainfo.gov.cn/periodical; http://www.zju.edu.cn;
http://lib.zju.edu.cn/English; http://www.zjupress.com

PITHC-SYNCHRONOUS ARTICULATORY SYNTHESIS INCORPORATED WITH THE INVERSE SOLUTION OF SPEECH PRODUCTION*

YU Zhen-li(俞振利)[1], CHING Pak-chung(程伯中)[2]

([1]*Dept. of Information & Electronic Engineering, Zhejiang University, Hangzhou 310028, China*)
zlyu @mail.hz.zj.cn
([2]*Dept. of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong*
pcching@ee.cuhk.edu.hk

Received Sep.12, 1999; revision accepted May.4, 2000

Abstract: This paper presents a new proposal to synthesize natural sounds with less control parameters by combining the inverse speech production and pitch-synchronous articulatory synthesis. The pitch-synchronous excited Reflection-Type Line Analog (RTLA) model is employed as the synthesis filter. Multi-rate system sampling and dynamic scattering wave adjustment are used to handle the variable VT length and the acoustic continuity. The synthesizer is controlled by vocal-tract (VT) area functions. Given the targets of formant trajectories, the dynamic VT area function which is modeled by time variant VT length is derived using an inverse solution of speech production. A distinguishing feature of this method is that artificially specified formant trace can be precisely aimed in the synthetical sounds. Experimental results show that the formant target can be well matched by the synthetic sounds. Potential application to text-to-speech conversion of this method is discussed.

Key words: speech synthesis, articulatory model, speech production, speech processing
Document code: A **CLC number:** TN912.326, O432

INTRODUCTION

Synthesis of natural speech sounds with less control parameters has practical significance, especially in text-to-speech (TTS) for producing sounds with artificially specified formant targets and scaled pitch parameter so that the timbre of speech output can be arbitrarily toned. Klatt's formant synthesizer and articulatory synthesizer are traditional parameter controlled synthesizers. However, the formant synthesizer needs too many parameters and the natural quality of the sound is not satisfactory. The articulatory synthesizer which is mainly driven by vocal-tract (VT) area function has shown its promising advantage in producing natural sound with less control parameters. But the inverse problem stands in the way of the application of articulatory synthesis. Gupta et al. (1993) and Schroeter et al. (1994) proposed an analysis-by-synthesis method to estimate the VT shape (with a fixed VT length) by matching the entire spectrogram of synthetic speech to the target spectrogram. However, if we wish to synthesize sounds of targeted formants that are either real estimated or specified artificially, it is essential to invert formants to VT area function and co-design the articulatory synthesizer accordingly.

Schroeder (1967) and Mermelstein (1967) established the fundamental theorem for resolving the problem from formants to VT shapes. Kelly and Loch-baum invented an area function controlled synthesis model (K-L model) (Kelly et al., 1962). Later, this model was developed as the Reflection-type Line Analog model (RTLA) (Liljencrants, 1985), which is simple and has good flexibility to take into account the effect of dynamic VT change and to insert various types of losses into the sound production simulation. The authors of this paper have proposed an improved method to determine VT area function from the formant target based on perturbation theory and interpolation method incorporated with codebook techniques (Yu et al., 1996; 1997). But how to combine the inverse solution and the synthesis model to produce sounds for targetted formant traces still remains to be resolved.

This paper presents recent developments in

* Project supported by NSFC (69972046), and Zhenjiang Provincial Natural Science Foundation of China (698076).

combining the RTLA synthesizer and the inverse solution for specified formant trace targets of vowel-to-vowel (VV) transition. The reason for using the RTLA synthesizer is that it is simple and has good flexibility to take into account the effect of dynamic VT change and to insert various types of losses into the simulation procedure of speech production. The VT model with variable VT length for obtaining more reasonable and smoother dynamic behavior is featured in this method.

PITCH-SYNCHRONOUS RTLA SYNTHESIZER

The entire system to implement the synthesis consists of two mainparts: the synthesis model, and the inverse solution of speech production.

The synthesis model is shown in Fig. 1. The vocal-tract is simulated by a reflection type line analog model with which the forward and backward partial waves of either air pressure or flow are handled by the scattering principle (Liljencrants, 1985). The RTLA model is excited by the Rosenberg's glottal wave with pitch-synchronous pulse shape (Rosenberg, 1971). The time varying pitch duration and the gain of the excitation can either be estimated from real speech, for voice copy synthesis, or artificially specified as desired on the base of estimation, for voice mimic synthesis. To fit in with the variable VT length and under-sampling of VT area function, two special aspects are emphasized. The first aspect is associated with the dynamic scattering that elaborates the area change during wave scattering and the second aspect is related to the variation of VT length.

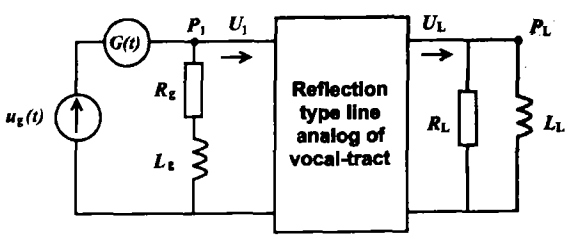

Fig. 1 Structure of the articulatory synthesizer

1. Dynamic wave scattering

Usually, wave scattering occurs simultaneously with VT area function changes. Dynamic scattering is employed to account for area change for the purpose of better smoothness of wave scattering. The scattering is divided into two separate stages, a stage of area change and a stage of static scattering. First, the waves are adjusted as expressed below.

$$\begin{cases} r_{i,tb} = r_{i,t-} - (r_{i,t-} + s_{i,t-}) \cdot \Delta Z_i / 2Z_{i+} \\ s_{i,tb} = s_{i,t-} - (r_{i,t-} + s_{i,t-}) \cdot \Delta Z_i / 2Z_{i+} \end{cases} \quad (1)$$

where r_i and s_i are the forward and backward partial waves of air flow, $Z_i = \rho \cdot c / A_i$ is the acoustic impedance of the i-th section tube with cross-sectional area A_i. $t-$ and $t+$ are the successive sample intervals, and tb denotes the moment after area change and before scattering. Then, static scattering is done as expressed below.

$$\begin{cases} r_{i+1,t+} = (1 + k_i) \cdot r_{i,b} + k_i \cdot s_{i+1,tb} \\ s_{i,t+} = -k_i \cdot r_{i,b} + (1 - k_i) \cdot s_{i+1,tb} \end{cases} \quad (2)$$

where k_i is the reflection coefficients of waves at the joint of two successive tubes and $k_i = (A_{i+1} - A_i)/(A_i + A_{i+1}) = (Z_i - Z_{i+1})/(Z_i + Z_{i+1})$.

2. Sampling rate conversion

The second aspect is the variation of VT length. The inverse solution gives frame-by-frame and pitch-synchronously segment-based VT parameters. For obtaining good smoothness, area functions and VT lengths is interpolated in each frame with $\frac{1}{8}$ at the system sampling frequency. To have a fixed section number of VT, the system sampling should vary as by the equation $T_i = 2 \cdot l_0 / c$ (l_0 is time-variant now). The output signal $x(n)$ is converted into $y(m)$ with constant sampling of T_o by an interpolation filter (Wu et al., 1987),

$$y(m) = \frac{T_i}{2T_c} \cdot \sum_{n=N_1}^{N_2} x(n) \cdot (mT_o - nT_i) \cdot \frac{\sin[2\pi(mT_o - nT_i)/T_c]}{2\pi(mT_o - nT_i)/T_c} \quad (3)$$

where $T_c > T_i/2$ and $T_c > T_o/2$ is temporal sampling rate. The 4 – 5 points rectangular window $w(\cdot)$ is symmetric around the point of the input signal.

To account for the VT losses, a loss factor $\gamma_i = 1 - 0.006 \cdot l_0 / \sqrt{A_i}$ is inserted into (2) to

attenuate partial waves in each section tube.

THE INVERSE SOLUTION

The inverse solution of speech production yields the dynamic VT area function for given targets of formant trajectories. The proposed inverse solution is achieved by using a VT area function modeling, an unique acoustic-geometric mapping codebook, zero frequencies and VT length interpolation, and the perturbation theory.

1. VT area function modeling

In the application of the perturbation theory, the VT area function is represented by

$$\log[A(i)] = \log[A_0] + \sum_{k=1}^{2N}[p(k) \cdot \cos(k\pi \cdot \frac{i \cdot l_0}{L})], \quad i = 1, \cdots, M \tag{4}$$

where i indicates the concatenate section tubes from glottis ($i = 1$) to lip end ($i = M$, the number of sections), l_0 is the length of each section tube, L is the total VT length which varies with time, A_0 is the area of the uniform VT tube, and $p(k), k = 1, \cdots, 2N$ are the coefficients of the terms of band-limited Fourier cosine expansion.

2. The perturbation theorem

According to Schroeder (1967) and Mermelstein (1967), given the distortion of the resonance frequencies of the VT (poles, identical to the formants) $F_p(k), k = 1, \cdots, N$ and that of the lip closed VT (zeros) $F_z(k), k = 1, \cdots, N$, the area perturbation can be uniquely determined, assuming that L is known in advance. Define

$$D_F = [\Delta F_p(1)/F_p(1), \cdots, \Delta F_p(N)/F_p(N), \Delta F_z(1)/F_z(1), \cdots, \Delta F_z(N)/F_z(N)]^T \tag{5}$$

and

$$D_P = [\Delta p(1) \cdots, \Delta p(2N-1), \Delta p(2), \cdots, \Delta p(2N)]^T \tag{6}$$

where $\Delta p(k)$ is the variance of the kth area perturbation, and $\Delta F_p(k)/F_p(k)$ or $\Delta F_z(k)/F_z(k)$ is the variation of the kth pole or zero frequency. A cross sensitivity matrix $A_{2N\times 2N} = \{a_{i,j}\}$ is defined by

$$D_F^0 = A_{2N\times 2N} \times D_P^0 \tag{7}$$

where D_P^0 is the testing variance of area perturbation and D_F^0 is the corresponding variation. The element of $A_{2N\times 2N}$, $\alpha_{i,j}$, is the Jacobian which can be obtained by direct VT calculation. The following formula can serves as guideline for inferring the trial increments $\{\Delta p(k)\}$ for desired variation of $\{F_p(k)\}$ and $\{F_z(k)\}$

$$D_P = A_{2N\times 2N}^{-1} \times D_F \tag{8}$$

Incorporated with the direct VT calculation, the Newton-Raphson procedure is employed to minimize the error between the calculated $\{F_p(k), F_z(k)\}$ of the candidate $\{p(k)\}$ and the target (Yu et al., 1996).

3. Dynamic acoustic and geometric constraints

Usually, only the poles (formants) can be estimated from the normal speech signal, while the zeros cannot or are very difficult, if not impossible, to estimate. To utilize the above perturbation method given formant targets only, additional acoustic targets, viz. $\{F_z(k)\}$, and VT length L are interpolated between the endpoints, i.e. the starting and destination of the VV transition (Yu et al., 1996). These $\{F_z(k)\}$ and L are then merged with the given formant target $\{F_p(k)\}$ to become the virtual target of the inverse process. The endpoints parameters can be determined by an improved acoustic-geometric mapping codebook (Yu et al., 1997). The codebook is generated in the following way. Initially, seven VT parameters, namely L, $\{p(k), k = 1 \cdots, 2N\}$, are quantized in suitable ranges. From these quantitative vectors, VT area functions are calculated by (4). Geometrical constraints are applied to avoid unreasonable VT shape. Acoustic constraints that vowels normally reside in a confined boundary in each of the $F_1 - F_2$, $F_1 - F_3$ and $F_2 - F_3$ subspaces are also applied on the calculated acoustic vectors. In addition, a distributed VT length in $F_1 - F_2$ subspace which is well defined based on measured data is used as a combined geometric and acoustic criterion to ensure the uniqueness of the code vectors. Finally, all the surpassed code vectors are clustered by acoustic and geometric optimization into an unique acoustic-geometric

mapping codebook with a much smaller size.

EXPERIMENTAL EVALUATION TESTS

Experimental tests were conducted for evaluating the proposed method. The procedure of the test is shown in Fig.2.

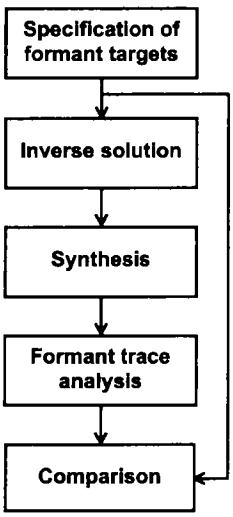

Fig.2 Flowchart of experimental test

The formant trace target, pitch and gain parameters are either artificially specified or estimated from real uttered vowel-to-vowel sounds. In the case of real sounds, the parameters are estimated by the "Xwaves" of Entropic ESPS tools (Entropic, 1993). The inverse solution and synthesis are performed with the method described above. Quantitive analysis with numerical comparison between the formant trace of the target and that of the synthetic sounds is carried out. A root mean square relative error (RMSRE) and a root mean square error (RMSE) between the formant trace of the synthetic sound, $F^s(t,k)$, and the target, $F^o(t,k)$, are defined as

$$E_{fR}=\sqrt{\frac{\sum_t\sum_k\{[F^o(t,k)-F^s(t,k)]/F^o(t,k)^2\}}{T\cdot K}} \tag{9}$$

and

$$E_f=\sqrt{\frac{\sum_t\sum_k[F^o(t,k)-F^s(t,k)]^2}{T\cdot K}} \tag{10}$$

The RMSRE and RMSE indicate the average level of the relative error and the absolute error, respectively, between the formants of the synthetic and the targets againt time and all the three formants. Fig.3 and Fig.4 show the spectrograms of the synthetical sounds targetted to an estimated formant trace of real sound /ae/ and to an artificially specified formant trace of /ae/, respectively. Fig.5 and Fig.6 show comparisons of the formant trace of synthetic sound (solid lines) to the target trace (dashed lines). The numerical data revealed that the SRMRE and SRME of the estimated target are 0.056 and 35.5 Hz respectively, and SRMRE and SRME of the artificial target are 0.027 and 33.0 Hz respectively. These data indicate that there is good matching of the formants of the synthetic sounds to the targets; and that the sounds are perceptually reasonable.

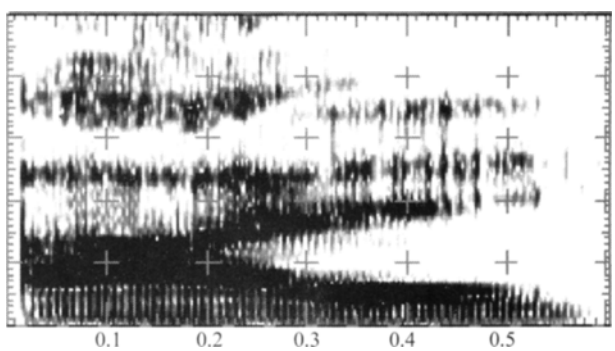

Fig.3 Spectrogram of /ae/ for estimated target

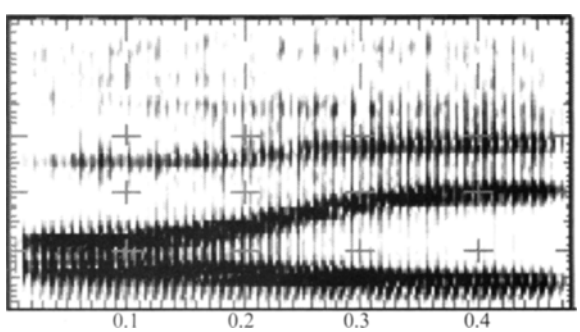

Fig.4 Spectrogram of /ae/ for artificial target

DISCUSSION AND CONCLUSION

There are several features of this present method. Firstly, this proposal implemented the

formant trace targetted synthesizer. VT area function that is modeled with variable VT length and derived from the format trace target through inverse solution, controls the synthesizer in the way of simulating the real world human speech production. Therefore, it provides natural quality and good dynamic behavior of the synthetic sounds. The elaborately designed optimized codebook combined with the multi-rate sampling conversion overcomes the increased non-uniqueness of the inverse mapping and multi-rate sampling in synthesis due to the variable VT length.

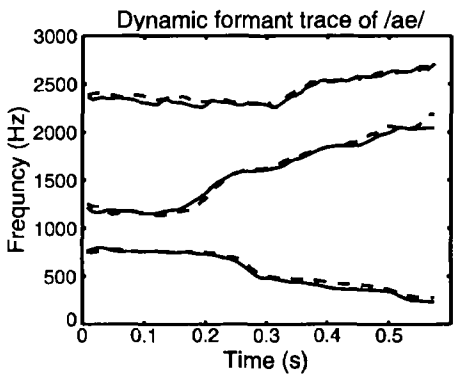

Fig.5 Formant trace compared to estimated target
—— synthetic formant; --- target formant

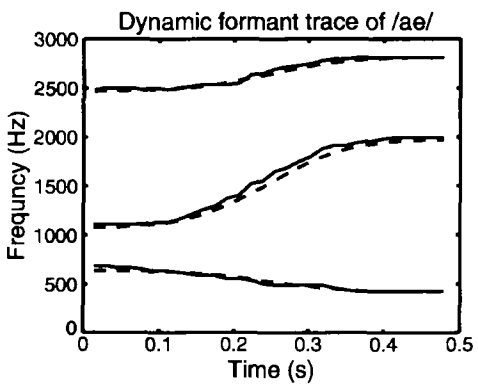

Fig.6 Formant trace compared to artificial target
—— synthetic formant; --- target formant

Secondly, this approach needs only the first three formant traces as the acoustic target to find the VT shapes. There is an advantage concerning its potential application to speech synthesis in TTS. In traditional TTS, a database from which the acoustic signal is synthesized is necessary. If the database is created according to spectrogram matching for an individual speaker, it is difficult to modify the sound timbre other than that of the training speaker. While with our method, the formants trace and pitch period can be artificially and arbitrarily specified independently of an individual training speaker. This property cannot be obtained by the method of Gupta et al. (1993). Furthermore, unlike the formant synthesizer, the bandwidths need not be targetted because the RTLA model synthesizer inherently yields these characteristics.

Thirdly, compared with the concatenate technique, typically the PSOLA approach, of speech synthesis, the present method can more feasibly control speech timbre. In PSOLA, only the pitch can be controlled. While in the present method, both pitch and formant traces can be controlled. Pitch control can be achieved just as same as in PSOLA. Formant trace control can be realized by deriving the area functions from formants in the inverse phase. The area functions, on the other hand, affect the formant traces through the RTLA module in the synthesis phase. Therefore, by using the present approach, pitch and formants as speech timbre can be controlled precisely and separately.

It should be pointed out that this approach cannot cover the consonants because it is mainly based on the perturbation method aimed only at the resonance frequencies of the VT as the acoustic target of the inverse problem. However, it is possible to deal with the consonants by other simple ways such as the overlap-add technique being employed. The general idea is, the vowel part can be handled by our present technique while the consonant part can be simply copied from the real speech waveform and the entire speech can then be reassembled by using the overlap-add technique, for instance PSOLA (Moulines, 1995). Another possible solution is that a consonant to vowel transition can be articulatorily synthesized with dynamic transitional VT shapes. The VT area function of the vowel part is resolved as described and that of the consonant part can be selected from a frozen set. The frozen VT shapes can be estimated by some other methods, for example, direct measurement or analysis-by-synthesis technique. Because there are only limited consonants and the sound timbre is mainly involved in the variation of vowels, it is reasonable to freeze the consonants' VT shapes while determining the underline vowels' VT area shapes. In this way, there is no neces-

sity of an on-line inverse process for consonants. Our ongoing investigation is addressing this subject.

References

Entropic Research Lab., 1993. Manmual of Xwaves, ESPS programs Version 5.0

Gupta, S.K. and Schroeter, J., 1993. Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *J. Acoust. Soc. Am.*, **94**(5): 2517-2530.

Kelly, J.L. and Lockbaum, C.C., 1962. Speech synthesis. Proc. 4th Int. Congress on Acoustics, Copenhagen, G(42):1-4.

Liljencrants, J., 1985. Reflection-type line analog synthesis. Ph. D. Thesis, Royal Institution of Technology (KTH), Stockholm, p.141.

Mermelstein, P., 1967. Determination of vocal tract shapes from measured formant frequencies. *J. Acoust. Soc. Am.*, **41**(5):1283-1294.

Moulines, E., 1995. TIme-domain and frequency-domain techniques for prosodic modification of speech. *In*: Speech Coding and Synthesis, Edited by Kleijn, W.B. and Paliwal, K.K., Elsevier, Amsterdam, p.519-555.

Rosenberg, A.E., 1971. Effect of pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, **49**(2): 583-591.

Schroeder, M.R., 1967. Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Am.*, **41**(4):1002-1010.

Schroeder, J. and Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech & Audio Processing*, **2**(1-II):133-150.

Wu, H.Y., Badin, P. and Cheng, Y.M., 1987. Vocal tract simulation: implementation of continuous variation of the length in Kelly-Lochbaum model, effects of area function spatial sampling. Proc. ICASSP'86, **1**:9-12.

Yu, Z.L. and Ching, P.C., 1996. Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method. Proc. ICASSP'96, Atlanta, USA, **1**:369-372.

Yu, Z.L. and Ching, P.C., 1997. Geometrically and acoustically optimized codebook for unique mapping from formants to vocal-tract shape. Proc. EUROSPEECH'97, Rhodes, Greece, **5**:2551-2554.