ARTICULATORY SYNTHESIS OF FORMANT TARGETED SOUNDS WITH PARAMETERS DERIVED FROM THE INVERSE SOLUTION OF SPEECH PRODUCTION

Z.L.Yu

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
zlyu@ee.cuhk.edu.hk
Also: Hangzhou University, China
eed@whale.hzuniv.edu.cn

P.C.Ching

Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
pcching@ee.cuhk.edu.hk

ABSTRACT

A new approach to produce high fidelity speech sounds by applying both the inverse solution of speech production and the pitch-synchronous articulatory synthesis technique is presented. Given a formant trace target, the dynamic vocal-tract area function together with time variant VT length are estimated using an inverse solution of speech production. The improved Kelly-Lochbaum filter of the synthesizer, with multi-rate system sampling and dynamic scattering wave adjustment, is employed to deal with the variable VT length and acoustic continuity. The synthesizer is controlled by the estimated VT area function. A distinguished feature of this method is that artificially specified formant traces can be precisely obtained. Experimental results show that the formant targets can be precisely matched by the synthetic sounds. A potential application of this method for text-to-speech conversion is discussed.

1. INTRODUCTION

Synthesis of natural speech sounds with less control parameters has practical significance, especially for text-to-speech (TTS) systems. It is desired to produce sounds with artificially specified formant targets and scaled pitch parameters 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 necessitates derivation of too many input parameters and the naturalness of the synthetic sound is not satisfactory. The articulatory synthesizer which is mainly driven by vocal-tract (VT) area function has shown its promising advantage to produce natural sound with less control parameters. But the inverse problem creates some non trivial difficulties for widespread application of articulatory synthesis. Analysis-by-synthesis method has been proposed [1,2] to estimate articulators (VT shape with a fixed VT length) by matching the entire spectrogram of synthetic speech to the target one. However, if voice is to be synthesized to fit onto a particular formant trace, which may either be estimated from real speech or specified artificially, it is essential to investigate the inversion of formants to VT area function and then design the articulatory synthesizer accordingly.

The fundamentals for solving the problem of mapping formants to VT shapes have been established in [3,4]. Kelly and Lochbaum, on the other hand, have invented an area function controlled synthesis model (K-L model) [5]. This model is then developed to become a reflection type line analog model (RTLA) [6]. The authors of this paper have proposed a new approach to determine VT area function from formant target based on the perturbation theory and interpolation method. A codebook is generated to facilitate unique mapping between formants and VT parameters [7,8]. But how to incorporate the inverse solution into the synthesis model to produce sounds for specific targeted formant traces still remains to be solved.

In this paper, we will present the recent development of combining inverse solution and RTLA synthesizer for specified formant trace targets of vowel-to-vowel (VV) transition. The reason of using RTLA synthesizer is that it has good simplicity and 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 will also be elaborated.

2. THE INVERSE SOLUTION : FROM FORMANTS TO VOCAL-TRACT SHAPES

The proposed inverse solution is achieved by making use of a VT area function modeling, an unique acoustic-geometric mapping codebook, zero frequencies and VT length interpolation and perturbation theory.

2.1. VT Area Function Modeling and Perturbation Formula

To apply the perturbation theory, the VT area function is represented by

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

where i indexes the concatenate section tubes from glottis ($i = 1$) to lip end ($i = M$), M is the number of sections, l_0 is the unique length of each section tube, L is the total VT length which is varying with time, A_0 is the area of the uniform VT tube, and $\{p(k), k = 1, \cdots, 2N\}$ are the coefficients of band-limited Fourier cosine expansion. According to [3,4], given the distortion of the resonance frequencies of the VT (poles) $\{F_p(k), k = 1, \cdots, N\}$ and that of the lip closed VT (zeros) $\{F_z(k), k = 1, \cdots, N\}$, the area perturbation can be uniquely determined if L is known a priori. Define

$$\mathcal{D}_{\mathcal{F}} = [\cdot\mathcal{F}_{\surd}(\infty)/\mathcal{F}_{\surd}(\infty), \cdots, \cdot\mathcal{F}_{\surd}(\mathcal{N})/\mathcal{F}_{\surd}(\mathcal{N}), \Delta F_z(1)/F_z(1), \cdots, \Delta F_z(N)/F_z(N)]^T \quad (2)$$

and

$$\mathcal{D}_{\mathcal{P}} = [\cdot_{\surd}(\infty), \cdots, \cdot_{\surd}(\in\mathcal{N} - \infty), \cdot_{\surd}(\in), \cdots, \Delta p(2N)]^T \quad (3)$$

where $\Delta p(k)$ is the variance of the k^{th} area perturbation, and $\Delta F_p(k)/F_p(k)$ and $\Delta F_z(k)/F_z(k)$ are the variation of the k^{th} pole and zero frequency, respectively. A cross sensitivity matrix $\mathcal{A}_{\in\mathcal{N}\times\in\mathcal{N}} = \{\alpha_{\rangle,|}\}$ is defined as

$$\mathcal{D}'_{\mathcal{F}} = \mathcal{A}_{\in\mathcal{N}\times\in\mathcal{N}} \times \mathcal{D}'_{\mathcal{P}} \quad (4)$$

where $\mathcal{D}'_{\mathcal{P}}$ is the testing variance of area perturbation and $\mathcal{D}'_{\mathcal{F}}$ is the corresponding variation. The element of $\mathcal{A}_{\in\mathcal{N}\times\in\mathcal{N}}, \alpha_{i,j}$, is the Jacobian which can be obtained by direct VT calculation. The following formula gives a guideline of inferring the trial increments $\{\Delta p(k)\}$ for the desired variation of $\{F_p(k)\}$ and $\{F_z(k)\}$

$$\mathcal{D}_{\mathcal{P}} = \mathcal{A}^{-\infty}_{\in\mathcal{N}\times\in\mathcal{N}} \times \mathcal{D}_{\mathcal{F}} \quad (5)$$

During the process of direct VT calculation, Newton-Raphson procedure is employed to minimize the error between the calculated $\{F_p(k), F_z(k)\}$ of the candidate $\{p(k)\}$ and that of the target. (See [7])

2.2. Dynamic Constraints on Virtual Target and Acoustic-geometric Mapping Codebook

To utilize the above perturbation method, additional acoustic target $\{F_z(k)\}$ and VT length L are interpolated between the endpoints, i.e. the starting and destination of the VV transition [7]. 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 endpoint parameters can be determined by an improved acoustic-geometric mapping codebook [8]. The codebook is generated in the following way. Initially, seven VT parameters, namely L, and $\{p(k), k = 1, \cdots, 6\}$, are quantized in suitable ranges. From these quantized vectors, VT area functions are calculated by (1). Geometrical constraints are applied to eliminate physically and anatomically unreasonable VT shapes. 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 to 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 composite geometric and acoustic criterion to ensure the uniqueness of the code vectors. Finally, all the surpassed code vectors are clustered to form an acoustic-geometric mapping codebook with a much smaller size by applying both acoustic and geometric optimization.

3. PITCH-SYNCHRONOUS RTLA SYNTHESIZER

The synthesizer is shown in Fig.1. The VT 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 [6]. To fit in with the variable VT length and under-sampling of VT area function, two aspects are considered. The first aspect is that the dynamic scattering which accounts for area change is divided into two separate stages, one to take care the area change and the other the static scattering. First, the waves are adjusted as

$$\begin{cases} r_{i,tb} = r_{i,t-} - (r_{i,t-} + s_{i,t-}) \cdot \Delta Z_i/2Z_{i+} \\ s_{i,tb} = s_{i,t-} - (r_{i,t-} + s_{i,t-}) \cdot \Delta Z_i/2Z_{i+} \end{cases} \quad (6)$$

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. $t-$ and $t+$ signify the successive sample intervals, and tb denotes the moment after area change and before scattering. Then, static scattering is executed

$$\begin{cases} r_{i+1,t+} = (1 + k_i) \cdot r_{i,b} + k_i \cdot s_{i+1,tb} \\ s_{i,t+} = -k_i \cdot r_{i,b} + (1 - k_i) \cdot s_{i+1,tb} \end{cases} \quad (7)$$

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

The second consideration is the variation of VT length. The inverse solution gives frame-by-frame and pitch-synchronously segment based VT parameters. Area function and VT length need to be interpolated in each frame with 8 times lower than the system sampling frequency. To keep a fixed section number of VT, the system sampling should vary as $T_i = 2 \cdot l_0/c$ (l_0 is time-variant now). The output signal $x(n)$ is converted into $y(m)$ with constant sampling T_o by an IIR filter [10],

$$y(m) = \frac{T_i}{2T_c} \cdot \sum_{n=N_1}^{N_2} x(n) \cdot w(mT_o - nT_i) \cdot \frac{sin[2\pi(mT_o - nT_i)/T_c]}{2\pi(mT_o - nT_i)/T_c} \quad (8)$$

where $T_c > T_i/2$ and $T_c > T_o/2$ is a temporal sampling rate. The rectangular window $w(\cdot)$ has a size of $4 \sim 5$ points and 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 (7) to attenuate partial waves in each sectional tube. The RTLA model is excited by the Rosenberg's glottal waveform with pitch-synchronous pulse shape [9]. The time varying pitch duration and the gain of the excitation can either be estimated from real speech or artificially specified as desired on the base of estimation. Because of the variant VT length and pitch duration, the sample number in each individual pitch period is changeable. A tailor made procedure to compute the exact time of samples

is designed. The detail, however, will not be elaborated due to page length limitation.

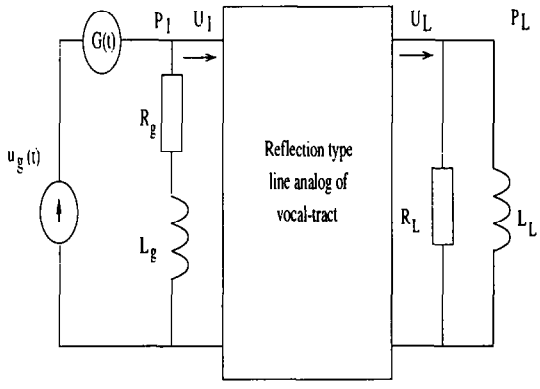

Figure 1: Structure of the articulatory synthesizer

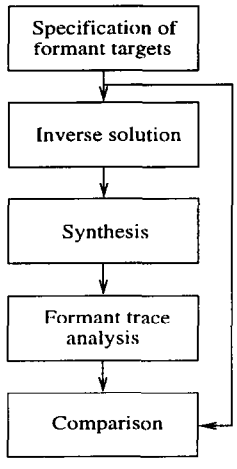

Figure 2: Flowchart of experimental test

4. EXPERIMENTAL TESTS

Experimental tests have been conducted. The procedures of these tests are shown in Fig.2. The formant trace target, pitch and gain parameters are either artificially specified or estimated from uttered vowel-to-vowel sounds. In the case of real speech, the parameters are estimated by the Entropic ESPS tools [11]. The inverse solution and synthesis are performed with the method described above. Quantized comparison of the formant trace between the target and the one 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

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

and

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

The RMSRE and RMSE indicate the average level of the relative error and the absolute error, respectively, between the formants of the synthetic speech and the target ones along the time domain. Fig.3 and Fig.4 show the spectrograms of the synthetic sounds targeted 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 the comparison of the formant trace of synthetic sound solid lines) to the target trace (dashed lines). The numerical data reveal that the SRMRE and SRME of the estimated target equal to 0.056 and 35.5Hz respectively, while SRMRE and SRME of the artificial target equal to 0.027 and 33.0Hz respectively. These data indicate good matching of the formants of the synthetic sounds to the targets. Besides, the sounds produced are considered to be perceptually good enough through informal listening test.

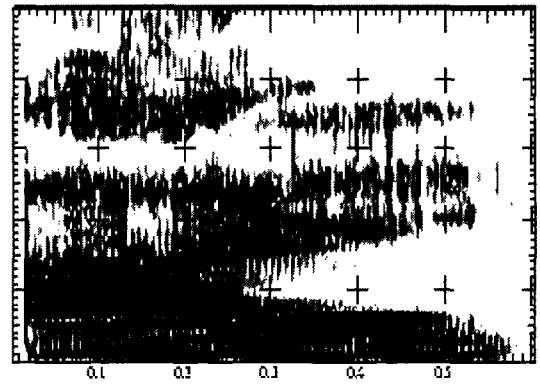

Figure 3: Spectrogram of /ae/ for estimated target

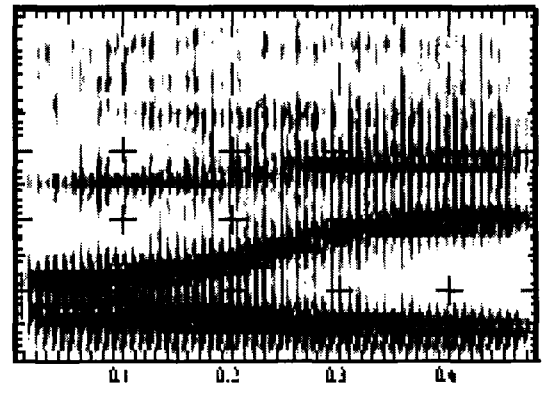

Figure 4: Spectrogram of /ae/ for artificial target

5. DISCUSSION AND CONCLUSION

There are several special features of the articulatory synthesis technique presented here. Firstly, a formant trace targeted synthesizer that is controlled by VT area function and modeled by variable VT

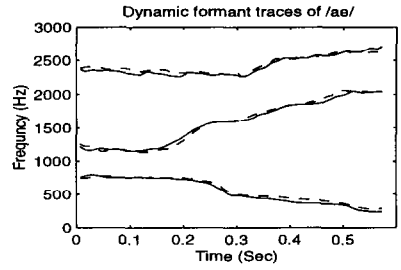

Figure 5: Formant trace comparing to estimated target

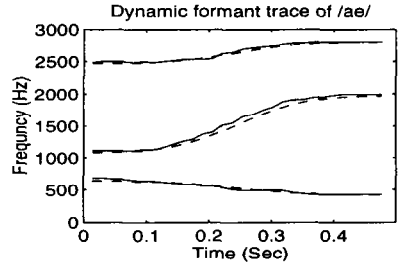

Figure 6: Formant trace comparing to artificial target

length is constructed. The elaborately designed optimized codebook combined with the time variant sampling rate conversion overcomes the problem of non-uniqueness of the inverse mapping and multi-rate sampling that often occur in speech synthesis due to variable VT length. Second, this approach only needs the first three formant trace 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 data base from which acoustic signal is synthesized is necessary. If the database is created according to spectrogram matching for a particular speaker, it is difficult to modify the sound timbre other than the training speaker. Whilst with our method, the formants trace and pitch period can be artificially and arbitrarily specified independently on an individual training speaker. Furthermore, unlike the formant synthesizer, the bandwidths are not necessary to be targeted because the RTLA model synthesizer inherently yields this characteristics.

Of course, it should be pointed out that this approach can not be used to synthesize consonants because it is mainly based on the perturbation method which only attacks the resonance frequencies of the VT as the acoustic target of the inverse problem. However, it is possible to deal with consonants by other simple ways such as the overlap-add technique being employed by PSOLA [12]. The general idea is, the vowel part can be resolved by the present approach while the consonant part can be simply copied from real speech waveform and the overall speech can be obtained from the two parts by overlap-add technique. Further investigation that addresses this topic is currently undergoing.

6. REFERENCES

[1] S.K.Gupta & J.Schroeter, "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," *J. Acoust. Soc. Am.*,vol.94, no.5, pp.2517-2530, 1993.

[2] J.Schroeter & M.M.Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech & Audio Processing*,vol.2, no.1(II), pp.133-150, 1994.

[3] M.R.Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*,vol.41, no.4, pp.1002-1010, 1967.

[4] P.Mermelstein, "Determination of vocal tract shapes from measured formant frequencies," *J. Acoust. Soc .Am.*, vol.41, no.5, pp.1283-1294, 1967.

[5] J.L.Kelly and C.C.Lochbaum, "Speech synthesis," *Proc. 4th Int. Congress on Acoustics*, Copenhagen, PaperG-42, pp.1-4, 1962.

[6] J.Liljencrants, *Reflection-type line analog synthesis*, Ph.D. Thesis, Royal Institution of Technology, Stockholm 1985.

[7] Z.Yu & P.C.Ching, "Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method," *Proc. ICASSP'96*, vol.1, pp.369-372, 1996.

[8] Z.L.Yu & P.C.Ching, "Geometrically and acoustically optimized codebook for unique mapping from formants to vocal-tract shape," *Proc. EUROSPEECH'97*, Rhodes, Greece, Sept., 1997.

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

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

[11] Entropic Research Lab., *ESPS programs Version 5.0*, 1993.

[12] W.B.Kleijn and K.K.Paliwal, *Speech coding and synthesis*,, Elsevier, Amsterdam, 1995.