

Articulatory synthesis evaluation of the performance of inverse speech solution for formant targeted vowel-to-vowel transition

YU Zhenli* CHING Pak-Chung** & CHEN Zhongbao***

* & **: Dept. of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China

* & ***: Dept. of Electronic Engineering, Zhejiang University, Hangzhou 310028, China
zlyu@ee.cuhk.edu.hk (also zlyu@public.hz.zj.cn); pcching@ee.cuhk.edu.hk

Abstract

An RTLA-typed articulatory synthesizer is constructed to evaluate the performance of the inverse solution of speech production based on perturbation theory. Vocal-tract area function is derived for a given formant trajectory target by applying the inverse solution. It is then used to control the RTLA synthesizer. Formant mimic synthesis and formant copy synthesis are implemented to validate the effectiveness of the method for both artificially specified formant trajectory targets and estimated formant traces of vowel-to-vowel transitions. The output quality of the synthetic sounds are found to be very good

1. Introduction

The authors have previously proposed an inverse solution of speech production based on perturbation theory that incorporates interpolation and codebook techniques [1-3]. The acoustic target information of the inverse problem is a set of poles, which are essentially formants, and zeros of the vocal-tract (VT), while the articulatory model is the band-limited Fourier cosine expansion of the VT area function. By applying the perturbation theory that relates pole/zero information to articulatory parameters, a close-loop optimization procedure is devised to determine the VT area function. An acoustically and geometrically optimized codebook that uniquely maps formants to zeros and VT length has been designed to enhance the robustness of the inverse solution. An interpolation method is used to impose constraints on the zeros and VT length for vowel-to-vowel transition so that VT shape can be derived for a given formant trajectory targets. The performance of the inverse solution is satisfactory with respect to the behavior of the naturalness and the dynamic smoothness of the area function to represent human vocal-tract shapes.

In this paper, we attempt to evaluate the performance of the inverse solution by using articulatory synthesis. The inverted VT area function is used to control a reflection-type line analog (RTLA) articulatory synthesizer [4]. Formant mimic synthesis as well as formant copy synthesis is employed to validate the

inverse solution, viz. the derived VT area function, in terms of its performance in producing good synthetic speech quality. This work also indicates a promising application of inverse solution of speech production based on perturbation theory.

2. Performance evaluation via synthesis

The evaluation system is shown in Fig. 1. The input parameters of the system are formant traces, pitch contour and amplitude. The system consists of two parts, the analysis component and the synthesis component. In the analysis stage, the inverse solution under evaluation is employed to obtain the VT area function for the specified input formant targets [1-3]. Time alignment for trajectories of area function, pitch and amplitude is necessary [5]. While in the synthesis stage, the vocal-tract area function, pitch and amplitude are used to control an articulatory synthesizer with RTLA model.

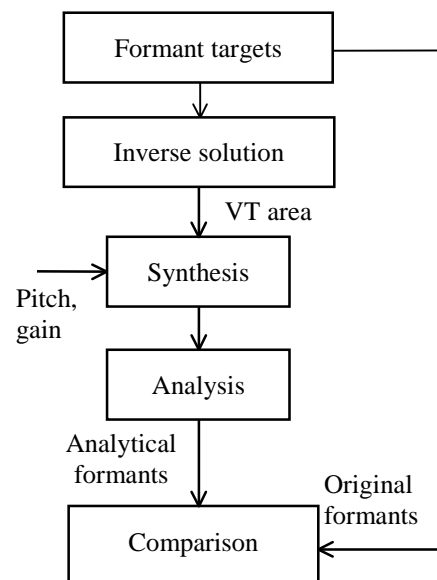


Fig.1 The flowchart of the synthesizer

2.1. RTLA synthesis model

The articulatory synthesizer is implemented according to Liljencrants' Reflection-type Line Analog model of vocal-tract [4]. The configuration of the RTLA synthesizer is depicted in Fig.2. The sound wave, i.e. pressure or volume velocity (P or U), is divided into two partial waves, that propagate inside and along the vocal-tract, by wave scattering principle. Different vocal-tract losses are taken into account. The glottal excitation is modeled by a pitch-synchronous pulse waveform [6]. A glottal impedance Z_g and a lip impedance Z_{lips} acts as glottal-to-vocal-tract coupling and lip radiation load, respectively. A time-variant sampling schedule is necessary to handle variable vocal-tract length. The details of the synthesis implementation can be found in [5][7].

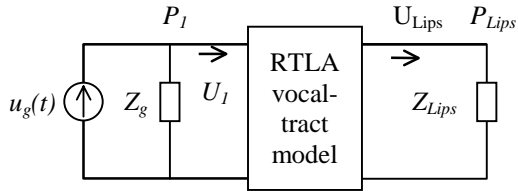


Fig.2 Configuration of RTLA synthesizer

2.2. Formant mimic synthesis

A formant mimic synthesizer is implemented to examine the inverse performance for artificially specified formant trajectory targets. The process starts with selecting endpoint formant parameters of desired vowel-to-vowel transition from an isolated vowel corpus. The transit formants are obtained by constraining the formants with sigmoid function along the pitch-synchronous segmental time coordinate. Vocal-tract area function is then determined by applying the inverse solution. The pitch contour as well as amplitude trace is generated according to the desired features of the target speech.

2.3. Formant copy synthesis

A formant copy synthesizer is also constructed to examine the performance of the speech production method, where the targets of formant traces, pitch and amplitude are copied from real estimated parameters obtained from uttered sounds. Speech sound is first recorded with a computer controlled interface device Dat-link. Formant, pitch and power are estimated from the stored speech data by the software ESPS (Entropic Lab, 1995). The estimation is performed using fixed analysis parameters, such as analytic window size,

overlap step, etc. To facilitate the pitch-synchronous RTLA synthesis implementation, a procedure to align the parameters into pitch-synchronous frame-by-frame based coordinate is necessary. The formant trace and the power parameters are interpolated with spline algorithm along the time axis. The square root of the power parameters is used as the gain control of the synthesis, while the formant trace is used as the acoustic target for the inverse solution of speech production that produces the vocal-tract area function sequence. The vocal-tract area function, pitch and gain are then used to control the RTLA synthesizer.

3. Validation

The synthetic sounds are analyzed by the same software ESPS for validation. The evaluation is focused on visual perception of spectrogram, analysis of formant trace, with comparison to the original targets, and listening tests.

Fig.3 gives an example of formant mimic synthesis for the vowel-to-vowel transition /a/-/o/. The vowel /a/ and /o/ are taken from Russian isolated vowels corpus [8]. On the other hand, Fig.4 and Fig.5 give an example of the result of formant copy synthesis for an original uttered sound /a/-/e/ by a male native mandarin speaker.

Fig.3(a) and Fig.4(a) are the spectrograms of the formant mimic synthetic /a/-/o/ and the formant copy synthetic sound /a/-/e/. As a comparison, Fig.4(b) gives the spectrogram of the original uttered /a/-/e/, from which we can observe a fair similarity between the synthetic spectrogram and the original one.

Fig.3(b) and Fig.5 illustrate the formant trace comparison between the synthetic sounds and the targets. From these figures we can see reasonable matching between these formant traces.

Furthermore, quantitative analysis is made to confirm the performance. A root mean square error and a root mean square relative error between the formant traces are calculated, respectively, from

$$E(k) = \left(\frac{1}{T} \sum_{t=1}^T [F^g(t, k) - F^s(t, k)]^2 \right)^{\frac{1}{2}}$$

and

$$E_R(k) = \left(\frac{1}{T} \sum_{t=1}^T \left[\frac{F^g(t, k) - F^s(t, k)}{F^g(t, k)} \right]^2 \right)^{\frac{1}{2}}$$

where $F^g(t, k)$ is the original formant trajectory and $F^s(t, k)$ represents the synthetic formant trajectory.

Table 1 gives the data of the errors for the two synthetic sounds. From the table, it can be seen that for formant mimic synthesis, the errors are relatively small where the largest $E(3)$ equals to 23.0Hz and the largest $E_R(1)$ equals to 0.029. The averaged $E(k)$ and $E_R(k)$ are 20.2Hz and 0.019, respectively. The results obtained by the formant copy synthesis are also satisfactory, where the largest $E(2)$ is 44.2Hz and the largest $E_R(1)$ is 0.092, and the averaged $E(k)$ and $E_R(k)$ are 38.6Hz and 0.045. It is intriguing to note that $E(k)$ and $E_R(k)$ obtained from formant copy synthesis are larger than that from formant mimic synthesis. This is because the formant trajectory target in copy synthesis is analyzed from the real uttered sounds through which estimation error may be induced, whilst the formant target in formant mimic synthesis is artificially formed by a sigmoid function which vary much smoothly.

In addition, the synthetic sounds are found to be perceptually good under informal listening tests.

4. Conclusion

In this paper, we present an articulatory synthesis evaluation of the performance of inverse solution of speech production based on perturbation theory. The Reflection-type Line Analog synthesis model is employed. Vocal-tract area function is derived for given formant trajectory targets by the inverse solution and then used to control the RTLA synthesizer. Formant mimic synthesis and formant copy synthesis are implemented to validate the inverse solution for both artificially specified formant trajectory targets and estimated formant traces of vowel-to-vowel transitions.

The evaluation is focused on visual perception of spectrogram, analysis of formant trace, with comparison to the original targets, and also through informal listening tests. The results illustrate that the proposed method of speech production can provide good synthetic speech.

It also clearly indicates that a practical application of the inverse solution is articulatory speech synthesis. Although articulatory synthesis techniques have been widely studied [7-9], the difficulty of estimating VT area function is still outstanding. The proposed solution for formant targeted inverse problem in this paper also provides a novel approach to realize formant mimic synthesis and formant copy synthesis. A variant of the formant copy synthesis in which the formant trace and pitch contour are modified separately to obtain colorful speech timbre has been examined. This added feature is

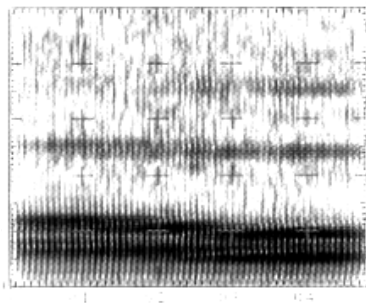
promising since it can be applied to text-to-speech synthesis.

5. Acknowledgement

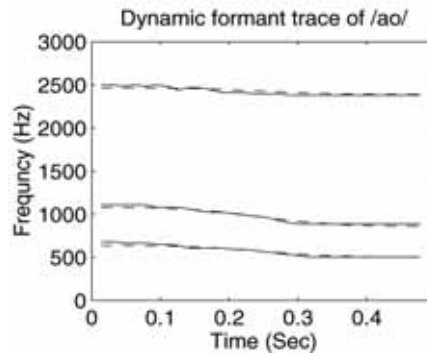
The research is partially supported by the Natural Science Foundation of Zhejiang Province.

6. References

- [1]. Yu, Z.L., "A method to determine the area function of speech based on perturbation theory," *STL-QPSR*, 4/1993, pp.77-95,1993
- [2]. Yu, Z.L, and Ching, P.C., "Determination of Vocal-Tract Shapes from Formant Frequencies Based on Perturbation Theory and Interpolation Method", *Proc. ICASSP'96*, pp.369-372, 1996
- [3]. Yu, Z.L, and Ching, P.C., "Geometrically and acoustically optimized codebook for unique mapping from formants to vocal-tract shapes," *Proc. EUROSPEECH'97* pp.2551-2554, 1997.
- [4]. Liljencrants, J., *Speech synthesis with a reflection-type line analog*, Ph.D. Thesis, Royal Institute of Technology (KTH), Stockholm, 1985
- [5]. Yu, Z.L, and Ching, P.C., "Complements to the articulatory synthesis of formant targeted sounds," *Proc. ICSP'98*, 1998
- [6]. 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
- [7]. Yu, Z. L, and Ching, P. C., "Articulatory synthesis of formant targeted sounds with the parameters derived from the inverse solution of speech production," *Proc. ICASSP98*, vol.2, pp.889-892, 1998
- [8]. Fant, G., *Acoustic theory of speech production*, the Hague: Mouton (2nd edition), 1970
- [9]. Flanagan, J.L., *Speech analysis synthesis and perception*, New York: Springer Verlag, 1972
- [10]. Sondhi, M.M., and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. ASSP*, vol.35, no.7, pp.955-966, 1987
- [11]. Gupta, S.K. and Schroeter, J., "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," *J.A.S.A.*, vol.94, no.5, pp.2517-2530, 1993



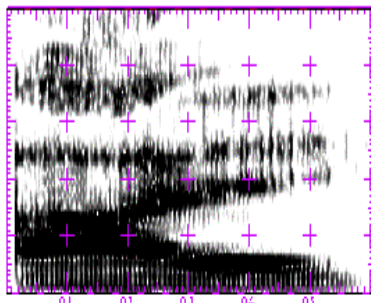
(a) Spectrogram



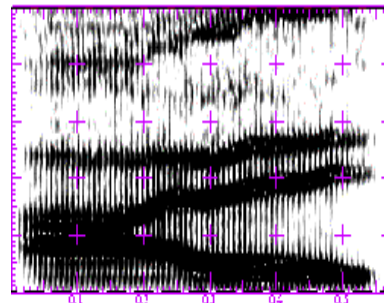
(b) Formant trace comparison

(solid: synthetic formants; dashed: targets)

Fig.3 Result of the formant mimic synthetic sound for /a/-/o/



(a) Original spectrogram



(b) Synthetic spectrogram

Fig.4 Spectrogram of the formant copy synthesis for /a/-/e/

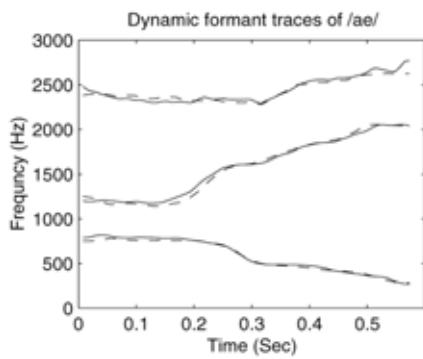


Fig.5 Formant trace comparison of the formant copy synthesis for /a/-/e/

Table 1 Errors analysis of the synthetic formants comparing with target formants

Type of synthesis		F1	F2	F3	Average
/a/-/o/ (formant mimic)	$E(k)$ (Hz)	17.8	19.8	23.0	20.2
	$E_R(k)$	0.029	0.020	0.009	0.019
/a/-/e/ (formant copy)	$E(k)$ (Hz)	34.1	44.2	37.5	38.6
	$E_R(k)$	0.092	0.028	0.016	0.045