AN ARTICULATORY DYNAMIC MODEL FOR DIPHTHONGS AND TRIPHTHONGS IN CHINESE

Yang Shun-an

Institute of Linguistics, Chinese Academy of Social Sciences, No.5 Jianguomennei Dajie, Beijing, China

ABSTRACT

The present paper describes the Exponential Dynamic Model for compound vowels such as diphthongs and triphthongs. With this Model, actual formant frequencies of all the allophones occurred in different phonological and phonetic contexts can be generated. The 9 diphthongs and 4 triphthongs in Standard Chinese constituted by 30 allophones can thus be generated with the target values of 6 phonemes.This Model is applicable to speech synthesis, so that data memory size can be decreased, and both intelligibility and naturalness of the synthesized diphthongs or triphthongs can be improved.

INTRODUCTION

The changing of sound color in compound vowels like diphthongs and triphthongs is mainly produced by the continuous movement of the speech articulators, i.e. by the continuous movement of the vocal tract. According to the acoustic theory of speech production, a given set of formant frequencies correspond to a given shape of the vocal tract. Therefore,the time-varing characteristics of formants can reflect the dynamic features of the compound vowels. Because of the practical need in speech synthesis and automatic speech recognition, it is necessary to formulate a functional model for describing the time-variation of the formant frequencies in dynamic vowels. And only after the formulation of such a model can we discuss the process of transformation between the discrete speech code and the continuous speech sound waves.

This paper proposes an Exponential Dynamic Model based on the analysis of the formant frequency data of the 9 diphthongs and 4 triphthongs in Standard Chinese. Parameters for the Model were obtained through analysis-by-synthesis, and the dynamic trajectories of formant frequences are in close approximation with the observed data. The utilization of this Model in the Synthetic System for Standard Chinese has both improved the quality of the synthetic sound and reduced the memory size for the syntnetic parameters.

FORMULAS OF THE EXPONENTIAL DYNAMIC MODEL

The observed time-varing trajectories of the formant frequencies indicated that the formant frequencies of a diphthong are constantly changing from one set of target values to another set, and the overal tencency of such dynamic trajectories is to have relatively stable parts at the beginning and the end of the vowel and to change rather abruptly at the transitional part. And, compared with the typical formant values of the phonemes composing a given diphthong, the starting and ending frequencies of the formants are only approaching the target values rather than actually reaching them. This condition is very like a curve obtained by joining two reverse e exponential functions. We thus hypothesise that a formant trajectory of a given diphthong can be approximated by the following formulas (Fig.1).

$$\left.\begin{aligned} F(t)&=Fc+0.5S*Fd[1-EXP[-\alpha(t-t0)S]] \\ Fc&=0.5(Fb+Fe) \\ Fd&=Fe-Fb \\ S&=1 \quad (t-t0>0) \\ S&=-1 \quad (t-t0<0) \end{aligned}\right\} \quad (1)$$

Here,
Fb is the beginning target value;
Fe is the ending target value;
t is normalized time;
t0 is the time of division; and,
α is the factor of transitional rate.

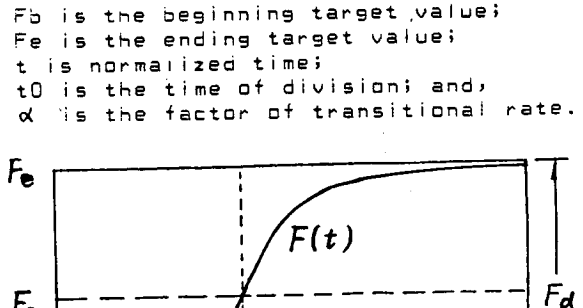

Fig.1 Schematic Dynamic Model for Diphthongs

Fig.2a and 2b show respectively the dynamic trajectories of formant frequences when t0 and α are altered. It is quite clear that in this model Fb and Fe can only approach the two target values rather than actually reaching them. The closer the division point is to the beginning point, the it is harder for the formant frequencies to reach the target value of Fb, and the easier it is for the ending formant frequencies to reach the target value of Fe; and, the greater the α, i.e. the factor of transitional rate, the easier it is for the formant frequencies of both of the extremities to reach their target values.

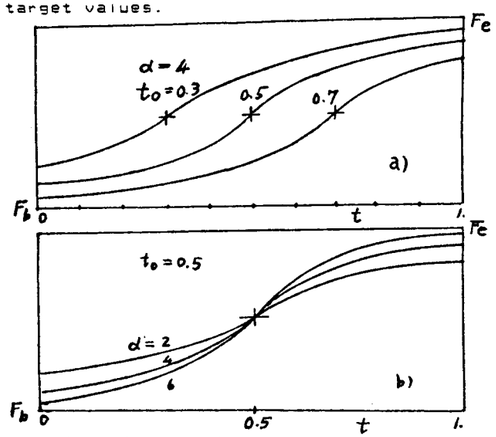

Fig.2 Variation of the formant trajectories with a) t0 and b) α

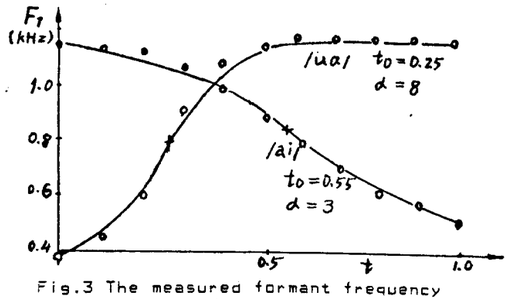

Fig.3 The measured formant frequency values and formant trajectories estimated with the formulas

The parameters Fb, Fe, t0 and α in the Model can be determined through analysis-by-synthesis. In Fig.3, the small circles represent the observed values of the first formant in /ai/ and /ua/, while the thin solid line is the trajectory calculated with formula (1) after the parameters for the Model had been determined. It can be seen that the two are in close approximation. As examples, the fitting values of t0 and α for F1 and F2 of the nine diphthongs are listed in Table 1.

Table 1 The fitting values of t0 and α for the 9 diphthongs in Standard Chinese

	/ai/		/ei/		/ao/	
	F1	F2	F1	F2	F1	F2
t0	0.55	0.43	0.27	0.19	0.52	0.49
α	3.0	4.0	3.1	4.2	1.9	2.1

	/ou/		/ia/		/ie/	
	F1	F2	F1	F2	F1	F2
t0	0.50	0.44	0.25	0.25	0.45	0.45
α	1.9	2.3	8.0	7.6	3.4	3.4

	/ua/		/uo/		/ye/	
	F1	F2	F1	F2	F1	F2
t0	0.20	0.25	0.35	0.42	0.35	0.23
α	7.1	7.8	3.8	3.8	3.5	3.5

Now, we can easily extend the Exponential Dynamic Model to include triphthongs. For triphthongs, considering the coarticulation effect between the three component phonemes, the dynamic trajectory of a given formant can be approximated by the following formula (Fig.4).

$$F(t)=Fo,m(t)+Fm,e(t)-Fm \qquad (2)$$

(for the meaning of the symbols here please refer to Fig.4)

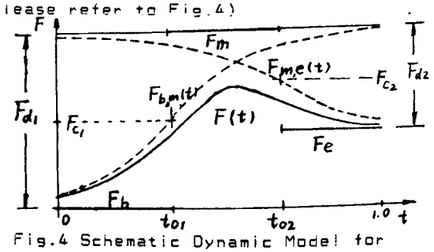

Fig.4 Schematic Dynamic Model for triphthongs

In this way, the dynamic aspect of a given formant in a triphthong can be defined by the tree target values Fb, Fm and Fe and the two division times t01 and t02 and the two factors of transitional rate α1 and α2, 7 parameters in all.

In overall generalization, for a given dynamic vowel that has n target values Fn, the cynamic trajectory of the frequency of a given formant can be approximated with the following formula:

$$\left.\begin{aligned} F(t)&=\sum_{i=1}^{n-1} Fi,i+1(t)-\sum_{i=2}^{n-2} Fi \qquad (n\geq 2)\\ Fi,i+1(t)&=Fci+0.5S*Fdi\{1-EXP[-di(t-t0i)S]\}\\ Fci&=0.5(Fi+F_{i+1})\\ Fdi&=F_{i+1}-Fi\\ S&=1 \quad (t-t0i\geq 0)\\ S&=-1 \quad (t-t0i<0) \end{aligned}\right\} \qquad (3)$$

SYNTHETIC PROOF AND APPLICATIONS

To verify the vailicity of this Exponential Dynamic Model, we had a syntnetic experiment with the Software System for Chinese Syllables [1, 2]. This system uses a cascade formant synthesizer; with 10 KHz of sampling frequency, and 12 bit of precision for D/A converter. The synthesis was operated on a BCM-3 microcomputer. the frequencies of the first three formants for the 6 target phonemes used for synthesizing the 9 diphthongs and 4 triphthongs in Standard Chinese are listed in Table 2. The F4 and F5 were fixed at 3500 Hz and 4500 Hz respectively.

Table 2 Frequency values of the first three formants for the 6 target phonemes used for syntnesizing the 9 diphthongs and 4 triphthongs in Standard Chinese

	/i/	/e/	/A/	/o/	/u/	/y/
F1(Hz)	270	520	1070	600	360	300
F2(Hz)	2350	2030	1200	1000	600	1890
F3(Hz)	3050	2720	2600	2500	2200	2250

Chinese is a tone language, and the F0-contour of each of the compound vowels were generated by a Tone Model [2].

All the syllables containing compound vowels in Standard Chinese were successfully synthesized. Fig.5 shows the spectragrams of four syllables, both natural

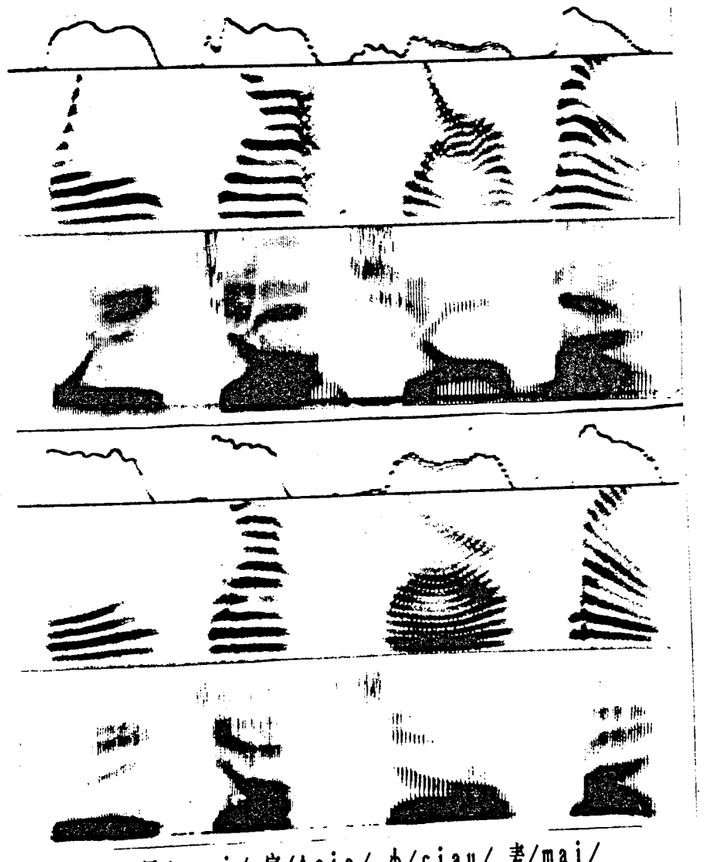

Fig.5 Spectragrams of four syllables containing diphthongs and triphthongs. The upper part for the natural ones and the lower part for the synthesized ones.

and synthesized, each containing a diphthong or a triphthong. It can be seen that the formant transition of those compound vowels are very smooth. Listening tests also indicated that both intelligibility and naturalness of the synthetic syllables were very close to those of the natural ones.

DISCUSSIONS

For synthetic application, there are two related features in this Exponential Dynamic Model: first, reltively few target values needed in input and storage, and second, better representation of the coarticulation effect. Speech analysis shows that one and the same phoneme in different compound vowels has different sound values. For example, the actual value of /ai/ and /ia/ are [ai] and [iA] respectively. Even two given vowels narrowly transcribed as the same sound in two different dynamic vowels, e.g. the [i] in [iA] and [iao] can have differences that should not be ignored. It means then, for synthesizing the 9 diphthongs and 4 triphthongs that are close to the natural ones, we will need 9*2+4*3=30 sets of target values. However, thanks to the ability of "approaching rather than actually reaching" the target values in the Exponential Dynamic Model, as few as 6 sets of target values listed in Table 2 are almost enough for this purpose. For instance, in synthesizing /ai/, [A] and [i] are used as target values; t0 is right in the middle and is relatively small. As a result, the beginning point is close to a open front vowel [a] rather than [A], and the ending point is a lower front vowel [I] rather than [i]. In synthesizing /ia/, [i] and [A] are also used as target values with t0 close to the beginning part and a relatively great , and the result is be that the two extremities are close to [i] and [A] respectively, and the /a/ part is relatively long and stable. In the acoustic vowel diagram in Fig. 6, the dynamic tracings are drawn for the synthetic /ai/, /ia/, /ao/, /ua/, /iao/ and /uai/ which use [i], [A] and [u] as the target values. The diagram shows that the beginning, middle and ending point of each of the compound vowels are just in their right places. In this sense, the synthesis of dynamic vowels with this Model is a synthesis with phonemic targets.

As a comparison, the trajectories generated by the exponential dynamic model reported in reference [3] and [4] always starts from the same first target value, disregarding the difference in factors like second target values and so on. The coarticulation effect is thus inadequately represented.

REFERENCE

[1] Yang Shun-an and Xu Yi (1987): A software system for synthesizing Chinese speech, Proc. 1987 Inter. Conf. on Chinese Information Processing, Aug. 4-6, Beijing, China.

[2] Yang Shun-an (1986): The effect of the dynamic characteristics of voice source upon the quality of synthesized speech, Zhongguo Yuwen, 1986 No.3, pp. 173-181 (in Chinese).

[3] Rabiner, L.R. (1968): Speech synthesis by rule: An acoustic domain approach, Bell System Tech. J., Vol.47,pp.17-37.

[4] Fujisaki, H. et al. (1973): Automatic recognition of connected vowels using a functional model of the coarticulatory process, J. Acoust. Soc. Japan, Vol. 29, pp. 636-637.

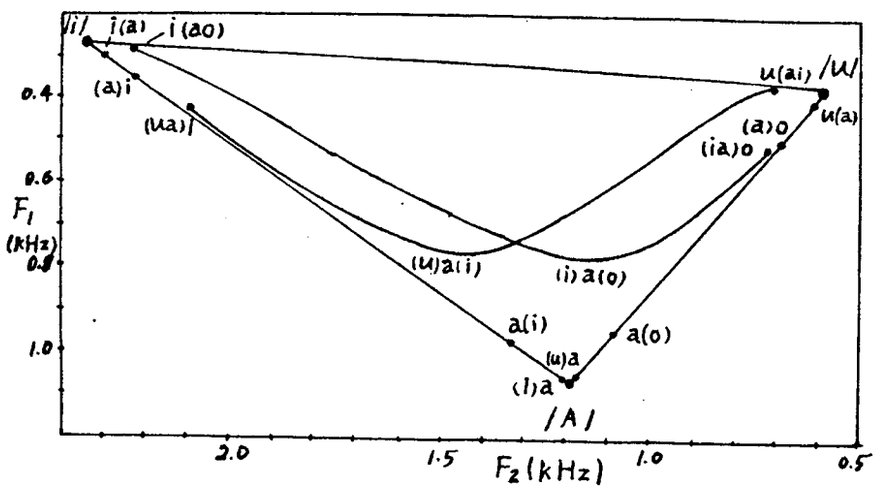

Fig.6 Acoustic vowel plot for the four diphthongs (/ai/, /ia/, /ao/, and /ua/) and the two triphthongs (/iao/ and /uai/).