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-varying 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 frequencies 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 syntactic parameters.

FORMULAS OF THE EXPONENTIAL DYNAMIC MODEL

The observed time-varying 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 overall tendency 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 exponential functions. We thus hypothesize that a formant trajectory of a given diphthong can be approximated by the following formulas (Fig.1).

\[
F(t) = F_c + 0.55 \cdot F_d \cdot [1 - \exp(-d(t-t_0))]
\]

Where,

\[
F_c = 0.5(F_b + F_e)
\]

\[
F_d = F_e - F_b
\]

\[
S = \begin{cases} 1 & (t-t_0) > 0 \\ -1 & (t-t_0) < 0 \end{cases}
\]

Here,

- \(F_b\) is the beginning target value;
- \(F_e\) is the ending target value;
- \(t\) is normalized time;
- \(t_0\) is the time of division; and,
- \(d\) is the factor of transitional rate.

Fig.1 Schematic Dynamic Model for Diphthongs
Fig. 2a and 2b show respectively the dynamic trajectories of formant frequencies when \( T_0 \) and \( d \) are altered. It is quite clear that in this model, \( F_b \) and \( F_e \) can only approach the two target values rather than actually reaching them. The closer the division point is to the beginning point; the harder it is for the formant frequencies to reach the target value of \( F_b \); and the easier it is for the ending formant frequencies to reach the target value of \( F_e \); and the greater the \( d \); i.e., the factor of transitional rate, the easier it is for the formant frequencies of path of the extremities to reach their target values.

![Fig. 2a and 2b](image)

Fig. 2 Variation of the formant trajectories with a) \( T_0 \) and b) \( d \).

In this way, the dynamic aspect of a given formant in a triphthong can be defined by the three target values \( F_0 \); \( F_e \) and \( F_t \) and the two division times \( T_01 \) and \( T_02 \); and the two factors of transitional rate \( d_1 \) and \( d_2 \). 7 parameters in all.

In general, the dynamic behavior of a given formant can be approximated with the following formula:

\[
F(t) = \sum_{i=1}^{n} F_i(t) - \sum_{i=1}^{n} F_{i+1}(t) \quad (\text{Eq. 2})
\]

For a given formant, the division points are defined by the two targets values \( F_0 \) and \( F_t \) and the two division times \( T_01 \) and \( T_02 \). The division points are defined by the two targets values \( F_0 \) and \( F_t \) and the two division times \( T_01 \) and \( T_02 \).

In this way, the dynamic aspect of a given formant in a triphthong can be defined by the three target values \( F_0 \); \( F_e \) and \( F_t \) and the two division times \( T_01 \) and \( T_02 \); and the two factors of transitional rate \( d_1 \) and \( d_2 \). 7 parameters in all.

In general, the dynamic behavior of a given formant can be approximated with the following formula:

\[
F_i(t) = F_{i+1}(t) - \frac{F_i(t)}{d_i(t-T_01)} \quad (\text{Eq. 2})
\]

\[
F_{i+1}(t) = F_{i+2}(t) - \frac{F_{i+1}(t)}{d_{i+1}(t-T_02)} \quad (\text{Eq. 2})
\]

The parameters \( F_0 \); \( F_e \); \( d_1 \) and \( d_2 \) in the model can be determined through analysis-synthesis. In Fig. 3, the small circles represent the observed values of the first formant in /ai/ and /au/; while the thick solid line is the trajectory calculated with formulas [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 \( T_0 \) and \( d \) for \( F_1 \) and \( F_2 \) of the nine diphthongs are listed in Table 1.

**Table 1 The fitting values of \( T_0 \) and \( d \) for the 9 diphthongs in Standard Chinese**

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>( T_0 )</th>
<th>( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ai/</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>/au/</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>/ei/</td>
<td>0.45</td>
<td>0.25</td>
</tr>
<tr>
<td>/ui/</td>
<td>0.55</td>
<td>0.25</td>
</tr>
<tr>
<td>/iu/</td>
<td>0.65</td>
<td>0.35</td>
</tr>
<tr>
<td>/iu/</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>/iu/</td>
<td>0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>/iu/</td>
<td>0.80</td>
<td>0.50</td>
</tr>
</tbody>
</table>

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

\[
F(t) = F_{i+1}(t) - F_i(t) \quad (\text{Eq. 2})
\]

For the meaning of the symbols here please refer to Eq. 6.

In this way, the dynamic aspect of a given formant in a triphthong can be defined by the three target values \( F_0 \); \( F_e \) and \( F_t \) and the two division times \( T_01 \) and \( T_02 \); and the two factors of transitional rate \( d_1 \) and \( d_2 \). 7 parameters in all.

In general, the dynamic behavior of a given formant can be approximated with the following formulas:

\[
F(t) = F_{i+1}(t) - F_i(t) \quad (\text{Eq. 2})
\]

\[
F_{i+1}(t) = F_{i+2}(t) - \frac{F_{i+1}(t)}{d_{i+1}(t-T_02)} \quad (\text{Eq. 2})
\]

The Chinese is a tone language, and the FO-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 spectrograms of four syllables: both natural and synthesized.

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

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>( F_1 )</th>
<th>( F_2 )</th>
<th>( F_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ai/</td>
<td>3500</td>
<td>272</td>
<td>2600</td>
</tr>
<tr>
<td>/au/</td>
<td>3600</td>
<td>280</td>
<td>2720</td>
</tr>
<tr>
<td>/ei/</td>
<td>3700</td>
<td>290</td>
<td>2820</td>
</tr>
<tr>
<td>/ui/</td>
<td>3800</td>
<td>300</td>
<td>2920</td>
</tr>
<tr>
<td>/iu/</td>
<td>3900</td>
<td>310</td>
<td>3020</td>
</tr>
<tr>
<td>/iu/</td>
<td>4000</td>
<td>320</td>
<td>3120</td>
</tr>
<tr>
<td>/iu/</td>
<td>4100</td>
<td>330</td>
<td>3220</td>
</tr>
<tr>
<td>/iu/</td>
<td>4200</td>
<td>340</td>
<td>3320</td>
</tr>
<tr>
<td>/iu/</td>
<td>4300</td>
<td>350</td>
<td>3420</td>
</tr>
</tbody>
</table>

**Fig. 5 Spectrograms of four syllables containing diphthongs and triphthongs. The upper part for the natural ones and the lower part for the synthesized ones.**
end-synthesized; each containing a diphthong or a triphthong. It can be seen that the prominent 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, relatively 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 [ia] can have differences that should not be ignored. It means then, for synthesizing the 9 diphthongs and 6 triphthongs that are close to the natural ones, we will need 9*2+6*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


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