A DUAL EXPONENTIAL MODEL
FOR FORMANT TRAJECTORIES OF DIPHTHONGS

by

Frantz CLERMONT

Published in the Proceedings of the
Second Australian Int. Conf. on Speech Science and Technology, pp. **146-151**,
November 1988.

A DUAL EXPONENTIAL MODEL
FOR FORMANT TRAJECTORIES OF DIPHTHONGS

Frantz CLERMONT

Computer Sciences Laboratory
Research School of Physical Sciences
The Australian National University
Canberra, ACT 2601, Australia

ABSTRACT : Australian English diphthongs are studied in terms of their second formant-frequency trajectories. These sigmoid-shaped trajectories may be decomposed, around a suitable breakpoint, as two exponential functions approaching two distinct vowel targets. In order to obtain this dual exponential representation, a set of candidate breakpoints defined along the inter-target transition are used to divide a given trajectory in two segments, thus simplifying the problem to that of fitting two single exponentials. A succession of such fits are performed, and the best pair of exponentials are determined in a root-mean-square sense. The method developed for constructing and evaluating the dual exponential model is described and illustrated. While the model fares well in a curve-fitting sense, its components do not always admit of a sensible phonetic interpretation in the case of an incomplete gesture towards the second vowel target.

INTRODUCTION

In contrast to monophthongs (or one-target vowels) which have been studied extensively, there are comparatively very few analytic descriptions of diphthongs (or two-target vowels). Early studies by Lehiste and Peterson (1961), Holbrook and Fairbanks (1962), Gay (1968), and Bernard (1970) have derived acoustic properties of diphthongs from their coarsely sampled trajectories, thus leaving the temporal dynamics of these trajectories not completely characterised.

Amongst the rare modelling studies (mostly aimed towards speech synthesis) devoted to formant trajectories of diphthongs in their densely sampled form, Rabiner (1968) and, later, Ren (1986) have proposed a critically damped solution completely specified by a single time constant. This approach suggests a continuous exponential behaviour from onset to offset targets. More recently, Yang (1987) has developed an articulatory model for synthesis of Chinese diphthongs, similar to the dual exponential model proposed by Clermont (1987). While both models are mathematically equivalent, it should be noted that generating formant contours in order to reproduce speech sounds contrasts significantly with modelling formant data of a realised speech sound. Consequently, the methods for constructing and evaluating these models do differ.

In addition, it appears that diphthongs have received considerable attention from a perception point of view. There exist a rather large number of studies which have been aimed towards determining portions of a diphthongal gesture essential to its correct identification. In an experiment on segmented diphthongs, for example, Gerber (1971) has reported that the cue to diphthong perception is the transition between the monophthongal components, and that a complete gesture towards the second vowel target is not necessary for the perceptual cue to persist.

Whether it be from an acoustic or perceptual point of view, it seems unquestionable that a diphthong always evokes some sort of dynamic behaviour to and from two distinct vowel regions. While there is no concensus on a diphthong's definition, there is agreement that onset targets are less variable than offset targets, and that the second vowel target may not be reached because of coarticulation or speaking rate variations.

This study is a first step towards extending Broad and Clermont's (1987) of monophthongal vowels to the more general case of diphthongs. As most English vowels are either diphthongs or tend to have diphthongised offsets, the intrinsic vowel dynamics of diphthongs form a good basis for developing a general model of syllable nuclei. Diphthongs are considered in null phonetic context and a parametric form of their formant-frequency trajectories is sought which characterises their dynamic behaviour.

EXPERIMENTAL METHODS

In order to derive the basic form of a descriptive model and to develop an evaluation methodology, we consider the simpler case of diphthongs in (nearly) null consonantal context (/hVV'/) and collect data from one speaker. Five repetitions of a subset of the Australian English diphthongs were presented, five times in a random order, to an adult male native speaker of Australian English. The same diphthongs were recorded both in citation form and in a carrier phrase in order to contrast production styles. Every diphthong, in both production styles, was presented visually at three-second intervals, thus allowing the speaker to take a fresh breath between presentations. In the case of a diphthong embedded in a phrase, the speaker was encouraged to utter the whole sentence as naturally as possible and to put the stress as he normally would. The carrier phrase ("Now, I see a [...] collection") was designed such that the target diphthong would be in a stressed position.

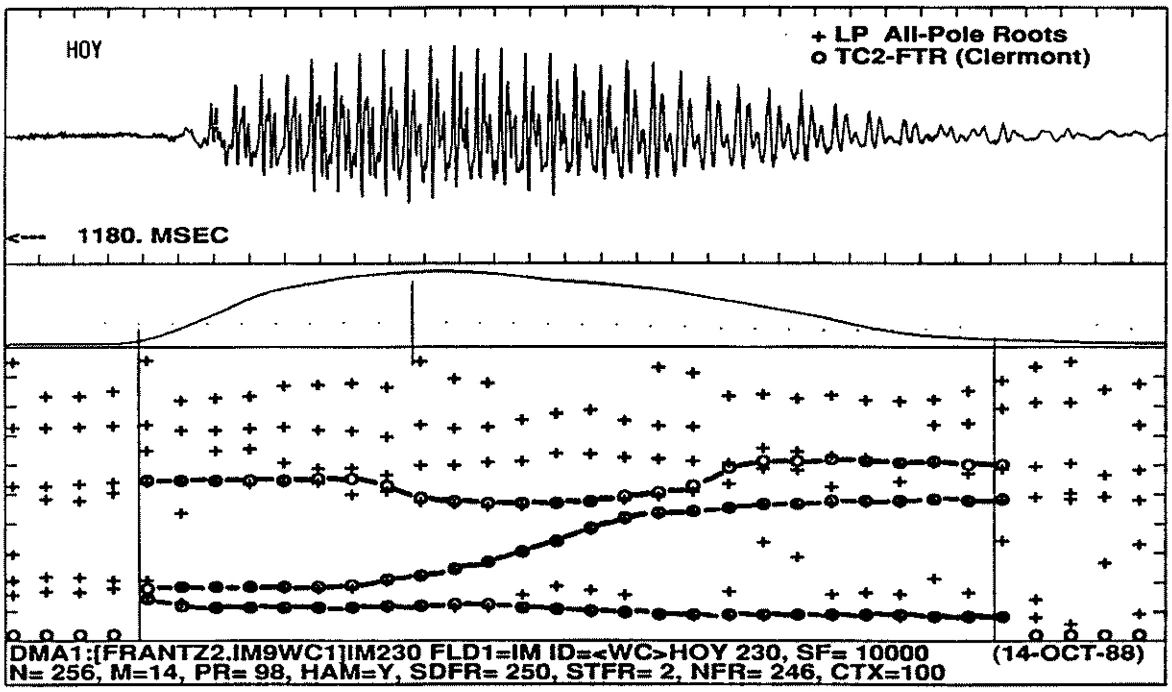

Figure 1: First Three Formant Trajectories of the Diphthong /hoi/

An acoustic prosodic analysis (Clermont, 1982) is used to obtain syllable boundary information. The moving-averaged envelope contour of the speech signal is extracted and partitioned by a number of thresholding levels determined between the overall maximum and minimum. The envelope contour is then searched at every level for left and right crossings between which are retained syllable peaks. Subsequently, a forward and backward search from every peak towards its successor and predecessor determines the limits of the interspersed intervals. The left and right boundaries of the syllable nuclei are then defined as the upper and lower limits of the preceding and following syllable, respectively. Furthermore, the "integral" under the curve of every syllable nucleus is used as a stress measure to automatically detect target diphthongs embedded in the carrier phrase.

Between the syllable boundaries just described (shown as vertical lines in Figure 1), formant-frequency contours are extracted by a method recently developed by Clermont (1988). The method is based on certain filtering and formant enhancing properties of the negative derivative of the linear prediction phase spectrum (Yegnanarayana and Reddy, 1979). Moreover, the formant tracking problem is recast in a dynamic programming framework in order to extract temporally constrained contours. Figure 1 shows, for the diphthong /hoi/, the first three formant-frequency trajectories superimposed on the peaks obtained from root extraction of the linear prediction polynomial (order 14).

DUAL EXPONENTIAL MODEL

Formulation

Modelling is herein restricted to the second formant trajectory (F_2) as it exhibits the most prominent structure. However, the methodology developed for this case does not preclude its applicability to the other formants, provided the structure of their trajectories is accounted for. The analytic description of F_2-trajectories of diphthongs is approached by seeking a convenient mathematical representation, which facilitates localisation of regions pertinent to the description of this complex type of vocalic nuclei.

It is first observed that the formant-frequency trajectories of diphthongal sounds are sigmoid-shaped functions gradually moving from and to more or less well-defined targets, with a transition particularly prominent for the second formants, and less so for the first and third ones, respectively. The first-formant movements are roughly mirror images of those of the second formants, with a much slower inter-target transition. Generally, there is also a noticeable time lag between first and second formants' onsets of inter-target transitions. In contrast, the third formant trajectories tend to remain fairly flat, except for a gentle minimum pointing down towards the inflection point of the second formants. In Figure 2 are shown a typical F_2-trajectory of a diphthong and its regions of interest. It is further observed that F_2 may be decomposed, around a breakpoint $[n_B, F(n_B)]$, as two exponentials which point towards distinct asymptotical values T_1 and T_2, and which may not necessarily have the same time constants. Rather than constrain both exponentials to form an odd function, the approach suggested here is to allow an unconstrained fit of two single exponentials which are juxtaposed near a breakpoint determined for a best fit to the data. For simplicity, no attempt is made here to impose continuity at the breakpoint.

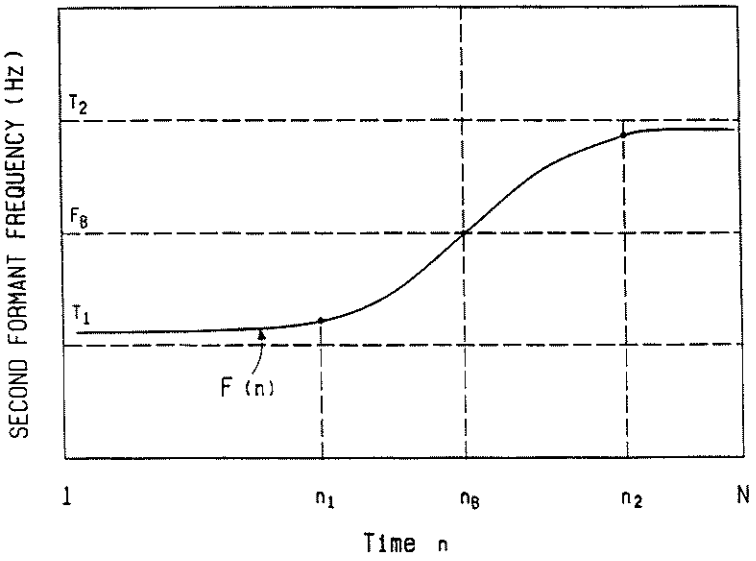

Figure 2: Schematic Representation of a Diphthong's F_2-trajectory

The proposed model (Clermont, 1987), henceforth referred to as the DEXP-model, is governed by the following equations :

$$F(n) = T_1 + [F(n_B) - T_1] \exp[\ k_1(n - n_B)] \qquad \text{for} \quad n \leq n_B \tag{1}$$

$$F(n) = T_2 + [F(n_B + 1) - T_2] \exp[-k_2(n - n_B - 1)] \quad \text{for} \quad n > n_B \tag{2}$$

Equations (1) and (2) describe the falling and rising exponentials to the left and to the right of the breakpoint, respectively. As $n \longrightarrow \mp\infty$, these equations yield the asymptote-defined target values T_1 and T_2. The abscissas n_1 and n_2 of the exponentials' knees are computed as twice the reciprocals of the constants in the exponential terms, or twice the time constants (2TC). The formulation of the DEXP-model differs slightly from Yang's (1987), which assumes a fixed breakpoint equal to the average of both target values. So, while Yang's model simplifies the problem by allowing a symmetrical partitioning of a given trajectory, the model presented here assumes no prior knowledge of the breakpoint and therefore allows more degrees of freedom in finding the best fit.

Model Construction

The mechanism for constructing the DEXP-model presents two aspects : (a) the generation of a prototype trajectory from a number of repetitions of the same utterance, which are expected to show temporal variations and (b) the curve-fitting algorithm.

(a) Generating Prototype Trajectory

- Select the longest of N_r repetition trajectories and call it the reference trajectory F_R. The $(N_r - 1)$ trajectories become test trajectories F_T.
- Slide every F_T along the F_R and compute an array of root-mean-square (RMS) errors. The time-window within which every F_T "best" matches the F_R is determined for a minimum RMS value of that array.
- Align each F_T within its respective "best" time-window of the F_R. Average all N_r trajectories over the points available for each frame. The resultant trajectory is the desired "prototype" and has the duration of the selected reference F_R.

(b) Two-Exponential Fitting Algorithm

- Select an initial breakpoint n_B as the abscissa of the mean frequency of a given trajectory.
- Fit falling and rising exponential functions by Prony's method (Markel and Gray, 1976). Compute constants $2/k_1$ and $2/k_2$, which define the limits of the knee-to-knee interval.
- Shift n_B within the knee-to-knee interval and iterate the fitting procedure until a minimum global RMS is obtained.

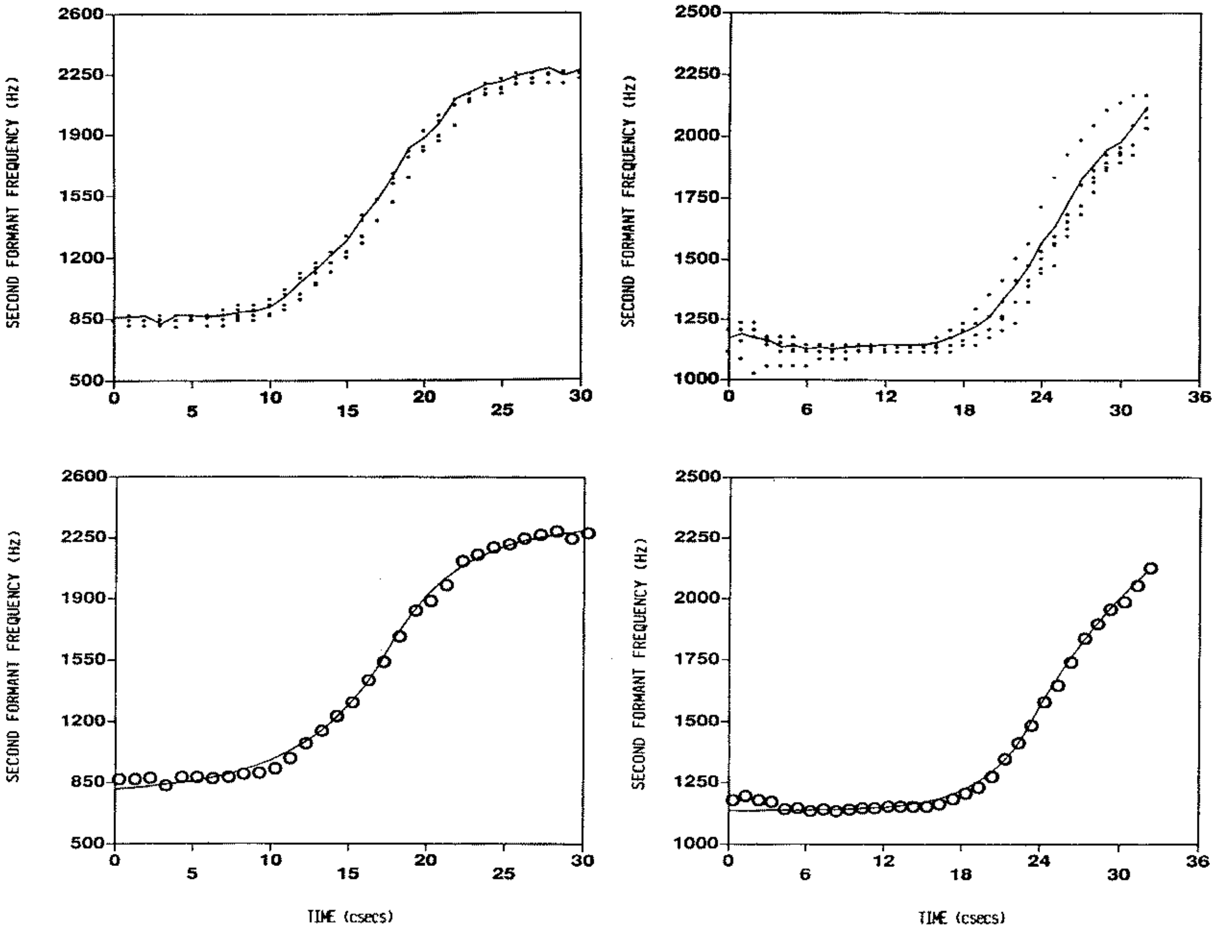

Figure 3: Prototype Trajectories (top) and DEXP fits (bottom) for Diphthongs /hoi/ (left) and (/hai/ (right).

The top graphs of Figure 3 illustrate, for the diphthongs /hoi/ and /hai/, prototype trajectories amidst a cloud of points which make up the trajectories of individual repetitions. In the case of /hoi/, the prototype trajectory is well behaved as it shows two stable target regions and a clearly defined inter-target transition. Also, there seems to be little inter-repetition variation. In contrast, the prototype trajectory for /hai/ is not so well behaved as the gesture towards the second vowel target is hardly realised. In addition, the inter-repetition variation appears to be larger. The bottom graphs of Figure 3 show the DEXP fits (solid line) superimposed on the prototype trajectories (circles). While, in the case of /hoi/, the curve-fitting algorithm succeeds in finding a well defined knee-to-knee interval, in the case of /hai/ the knee to the right of the trajectory exceeds the last sample point. This is explained by the absence of curvature of the rising part of the /hai/-trajectory, thus causing a fitted exponential to point towards a distant target. In this case, the curve-fitting algorithm is not allowed to iterate, and the breakpoint n_B is kept as the abscissa corresponding to the mean frequency of the whole trajectory.

Model Evaluation

We approach the evaluation of the DEXP-model in two ways. Firstly, we use the variation amongst the N_r available repetitions of our data set as a baseline to gauge the success of the model in a curve-fitting sense. The raw trajectories of individual repetitions are time-aligned with the prototype trajectory as described earlier, and the inter-repetition variation (RMS_{IR}) is computed as the unbiased estimate of the local variance over N_r.

/hVV/	RMS_{IR} (Hz)	RMS_P (Hz)	RMS_{r_1} (Hz)	RMS_{r_2} (Hz)	RMS_{r_3} (Hz)	RMS_{r_4} (Hz)	RMS_{r_5} (Hz)
/hoi/	44	23	35	27	52	27	32
/hai/	71	17	32	42	27	23	29

Table 1: RMS errors

The RMS_{IR} values in Table 1 confirm the observations made earlier about the diphthong /hai/ showing larger deviations from its prototype. In addition, the DEXP-fits to the prototype and to the individual repetition-trajectories divided around the prototype breakpoint yield the tabulated RMS_P and RMS_{r_i}. Generally, these values lie within the inter-repetition variation, and, in particular, the rms errors in fitting the prototype trajectories are much smaller than the RMS_{IR} (52% and 24% for /hoi/ and /hai/). The small RMS_P for the /hai/-prototype may again be explained by the short segment of the unrealised gesture towards the second target. An unconstrained exponential fit to this quasi-linear segment is likely to contribute little to the global rms error.

While the dual exponential model succeeds in a curve-fitting sense, the question of interpretability of its components is raised as part of the evaluation methodology. Tables 2 and 3 contain most components of the DEXP-model applied to the prototype trajectories of diphthongs /hoi/ and /hai/, respectively.

/hoi/	T_1 (Hz)	$F(n_1)$ (Hz)	$F(n_B)$ (Hz)	F_{avg} (Hz)	$F(n_2)$ (Hz)	T_2 (Hz)	$2TC_1$ (ms)	$2TC_2$ (ms)	T. rate;T. dur. (Hz/ms);(ms)
Citation	813	909	1425	1430	2165	2333	86	89	7.4;170
Carrier	813	889	1348	1352	2180	2373	62	66	10;130

Table 2: DEXP-Model components for /hoi/

For the diphthong /hoi/, both sets of vowel targets are consistent with one another, and fall within the expected range of their corresponding monophthongal vowel targets. In both production styles, the frequencies ($F(n_B)$) at the breakpoint determined by the DEXP-model are very close to the mean frequencies (F_{avg}) of the whole trajectories within 5 to 7 Hz. The knees (2TC) of both sets of exponentials are within 3 to 5 msecs from one another. Furthermore, the transition rate (T. rate), calculated as the rate of change in frequency within the knee-to-knee interval, is larger when the diphthong is uttered in a carrier phrase, thus implying an increase in speaking rate in a continuous speech (carrier phrase) environment. If this is so, then the duration of the transition (T. dur.) is expected to be shorter than it would be in the citation production style. The corresponding tabulated values verify this.

While the DEXP-model components for the diphthong /hoi/ lend themselves to meaningful phonetic intepretations, those derived for the diphthong /hai/ do not entirely.

Comparable observations can be made about the falling portion. However, because of the incomplete gesture towards the second vowel targets, the model yields, for the rising portion of the /hai/-prototype, components which are not readily interpretable.

/hai/	T_1 (Hz)	$F(n_1)$ (Hz)	$F(n_B)$ (Hz)	F_{avg} (Hz)	$F(n_2)$ (Hz)	T_2 (Hz)	$2TC_1$ (ms)	$2TC_2$ (ms)	T. rate;T. dur. (Hz/ms);(ms)
Citation	1141	1178	1394	1406	2112	2845	63	290	5.5;170
Carrier	1006	1079	1448	1472	2243	2468	89	129	6.1;190

Table 3: DEXP-Model components for /hai/

CONCLUSION

A dual exponential model for characterising F_2-trajectories of Australian English diphthongs has been described and illustrated. Two cases (/hoi/ and /hai/) have been examined which validate the modelling method in a curve-fitting sense, but which also contrast the success and limitations in interpreting the components of the proposed model. If the second vowel target is sufficiently realised, then the present method is validated. If the second vowel target is only slightly approached, then the method fails to yield components interpretable in phonetic terms. While a fully realised second target may not be necessary for a diphthong to be correctly perceived (Gerber, 1971), the acoustic manifestation of this phenomenon can not be ignored from a modelling point of view. Consequently, special forms of a dual exponential model will have to be developed in order to accomodate cases of unrealised gestures of diphthongs.

ACKNOWLEDGEMENT

I am grateful to Dr. David J. Broad for his invaluable intellectual and moral support.

REFERENCES

Bernard, J. (1970), "Towards the acoustic specification of Australian English", Zeitshrift für Phonetik, 23(2/3), 114-128.

Broad, D.J. and Clermont, F. (1987), "A methodology for modeling vowel formant contours in *CVC* Context", J. Acoust. Soc. Am, 81(1), January 1987, pp. 155-165.

Clermont, F. (1987), "A mathematical description of F_2-trajectories of diphthongs", PhD Thesis : Interim Progress Report", The Australian National University, Computer Sciences Laboratory, July 1987.

Clermont, F. (1988), "Formant contour extraction by a temporally-constrained search of the spectral resonance space", J. Acoust. Soc. Am., 84, S21-22, Fall 1988.

Clermont, F. (1982), "Syllabic epoch detection by multi-level search of the envelope contour", Unpublished research report.

Gerber, S. (1971) "Perception of segmented diphthongs", Proc. VIIth Int. Congr. Ph. Sci., 479-492, 22-28 August, Montréal, Canada.

Gay, T. (1968) "Effect of speaking rate on diphthong formant movements", J. Acoust. Soc. Am., 44(6), 1570-1573.

Holbrook, A. and Fairbanks, G. (1962), "Diphthong formants and their movements", J. Speech and Hearing Research, 5(1), 38-58.

Lehiste, I. and Peterson, G.E. (1961), "Transitions, glides, and diphthongs", J. Acoust. Soc. Am., 33(3), 268-277.

Markel, J.D. and Gray, A.H., Jr. (1976), *Linear Prediction of Speech*, Springer:Berlin.

Rabiner, L.R. (1968) "Speech synthesis by rule : An acoustic domain approach", The Bell Syst. Tech. Journal, 47(1), 17-37.

Ren, H. (1986) "On the acoustic structure of diphthongal syllables", PhD dissertation, University of California at Los Angeles.

Yang, S. (1987), "An articulatory dynamic model for diphthongs and triphthongs in Chinese", Proc. XIth Congr. Ph. Sci., 1, 1-7 August, Estonia, USSR.

Yegnanarayana, B. and Reddy, R. (1979), "A distance measure based on the first derivative of the linear prediction phase spectra", IEEE Int. Conf. on Acoust., Speech and Sig. Proc., Conf. Record, 744-747.