CHARACTERISATION OF THE DIPHTHONGAL SOUND BEYOND THE F_1-F_2 PLANE

Dr Frantz CLERMONT
Department of Computer Science,
University College,
The University of New South Wales,
Australian Defence Force Academy,
Canberra, ACT 2601, Australia

December 1992
Paper published in the Proceedings of the IVth Australian International Conference on Speech Science and Technology, Brisbane, Australia

pp. 298-303

CHARACTERISATION OF THE DIPHTHONGAL SOUND BEYOND THE F_1-F_2 PLANE

Frantz CLERMONT

Department of Computer Science,
University College,
The University of New South Wales,
Australian Defence Force Academy,
Canberra, ACT 2601, Australia

ABSTRACT: A well entrenched and still dominating approach to characterise the diphthongal sound is based: (1) on the two lowest, vocal-tract resonance (or formant) frequencies (F_1 and F_2) considered individually and/or in a planar space; and (2) on a very sparse, temporal representation of these frequencies. While this time-honoured approach has been instrumental in deriving certain important properties of diphthongs, our basic knowledge appears to have advanced little beyond the F_1-F_2 plane as an acoustic phonetic framework for describing the dynamics and the complex vocalic nature of these speech sounds. In contrast, a new perspective on the formant space of the diphthong is offered here by studying the detailed time-varying behaviour of the individual formants; and by unveiling transition characteristics of particularly the F_3-contour, which have hitherto been either unacknowledged or severely attenuated by sparse time-sampling.

INTRODUCTION

The general impression emergent from examination (Clermont, 1991) of the acoustic-phonetic literature on the diphthong (also known as a two-target vowel) is that of a relatively less mature and less cohesive body of knowledge in comparison with the literature on the monophthong (also known as a one-target or simply a vowel). While it can be argued that the significant developments concerning properties and models derived for monophthongs can be brought to bear on the description of the diphthong which is vocalic by definition, it seems evident that extension of those previous findings would be more easily achieved with a better understanding of the dynamics and of the dual vocalic nature of the diphthong. In this regard, a fundamental and apparently unresolved question is whether the vocalic duality in a diphthong arises simply from a static combination of two monophthongs as the bi-phonemic representation (/aɪ/ for example) tends to suggest, and how the duality is manifest in the formant-contours of the diphthong.

The acoustic-phonetic description of diphthongs has been based primarily on the two lowest formant frequencies (F_1 and F_2), the co-variation of which lends itself to well established phonetic and articulatory interpretations. However, a rather consistent trend has been to represent very sparsely (two or four samples) the time course of these frequencies. While sparse time-sampling can be justified from the point of view of an economical description of the diphthong's formant-contours, it is flawed with a lack of explicit recognition of the gradual, rather than step-like, transition from the onset to the offset vowels. One of the aims in this paper is therefore to examine a detailed, temporal representation of diphthong formants, with the expectation (Clermont, 1991) that a dense sampling approach will enhance and/or render more accurate transition characteristics than those derived from sparsely-sampled formant-contours.

The significance of the F_1-F_2 planar space for describing and differentiating acoustic-phonetic features of individual vowels is well understood from the seminal studies by Potter and Peterson (1948), Potter and Steinberg (1950), Peterson and Barney (1952), and the comprehensive study by Pols *et al.* (1969). In this regard, analogous interpretations (Holbrook and Fairbanks, 1961; Bernard, 1970) have been achieved in the F_1-F_2 plane of diphthongs. However, a peculiarity of the diphthong's formant-contours, subject to dense time-sampling, is that the transition from the onset to the offset vowels is manifest not only along the F_2-contour as consistently portrayed in previous work, but also along the F_3-contour of the back-to-front diphthongs in particular (such as /aɪ/ and /ɔɪ/). The further aim of this paper is therefore to illustrate the importance of the F_3-contour and hence of the F_2-F_3 planar space (Clermont, 1991) for describing the transition characteristics of a diphthongal sound.

DIPHTHONG MATERIAL and FORMANT-CONTOUR PARAMETERISATION

A methodological trend in previous acoustic-phonetic studies of diphthongs is to consider these speech sounds in consonantal contexts. While it is understood that certain consonants exert little influence on vocalic nuclei, discussions of the procedures used or of the difficulties encountered in rendering a two- or four-point description of formant-contours of diphthongs coarticulated with influential consonants are either scant or not offered in previous work. The consistent trend concerning the use of syllable contexts is particularly misleading, as it conveys the impression that the acoustic phonetic properties of diphthongs produced by themselves are well understood. Our analysis (Clermont, 1991) of the literature indicated, however, that further research on characterising nominal diphthongs is warranted.

Thus, in contrast to most previous studies, our data set consists of English diphthongs produced in /hVV/ context (as in HIGH, HOY, HOW, HOE, HERE and HAIR), in which the initial voiceless, fricative /h/ is expected to have minimal coarticulatory effect on the following diphthong nucleus. The /hVV/ syllables were spoken in citation form, five times in a random order, by three adult male, native speakers of Australian English.

The syllable-detection module of an acoustic-prosodic system (Clermont and Butler, 1988; Clermont, 1991) was used to identify the nucleus boundaries of every diphthongal /hVV/ syllable. The segmentation algorithm is based on a multi-level peak search of the envelope contour of the speech waveform. A new formant-tracking method (Clermont, 1992a; 1991) based on temporally-constrained spectral matching was then applied to every diphthong nucleus, with a frame advance of 10msec, the result of which captures the detailed time-varying behaviour of the F_1, F_2 and F_3 through the nucleus duration. Recall that an *F*-pattern is generally used to designate F_1, F_2 and F_3 measured at a single time-frame. By extension, we refer to an F_n-pattern for describing a time-varying *F*-pattern, where n is a time index.

TIME-ALIGNMENT and REDUCTION to PROTOTYPE FORMANT-CONTOURS

Since there are five repetitions of every diphthong, it was determined to be more appropriate and more economical to study *prototype* formant-contours which embody features typical of repetition contours for a given diphthong, rather than get enmeshed in describing the details of all the repetition contours. The time-alignment method used is described in the next paragraphs.

A prototype-contour or prototype in short is defined as the best average of a number of contours obtained from repetitions of the same utterance. The method of contour alignment applied to individual formants consists of: (1) selecting the longest repetition-contour as a reference contour; (2) time-shifting the remaining repetition-contours (also referred to as test contours) with respect to the reference, and finding the best alignment in a least-mean-square sense; and then (3) averaging all repetition-contours over the points available for each frame to produce the prototype-contour. Every test contour is allowed to slide beyond the onset and the offset points of the reference contour by a fixed number of frames (l_w) referred to as the relaxation interval. The optimum alignment is the one which minimises the Root-Mean-Square (RMS) difference between the segments of the reference and the test contours over the alignment interval. The spread about a prototype (defined as inter-repetition (IR) dispersion) is then computed as the square root of the unbiased estimate of the local variance over the prototype duration.

If the method outlined above, however, is applied to every formant individually, there is no guarantee that the resulting prototypes of the first three formants will be in synchrony with respect to one another. In order to preserve temporal cohesion within every diphthongal nucleus, the method was then extended to constrain the prototype-contours of two formants to be in synchrony with the third one. By analogy with the notion of reference contour in the case of single formants, the notion of a *base formant* is advanced in the case of alignment of an F_n-pattern. Since there are three possible base formants, there are also three measures of inter-repetition dispersions which are averaged to yield a global measure of dispersion about the prototype F_n-pattern. Furthermore, since the size of the optimum relaxation interval is not known *a priori*, the extended method just described was carried out for a number of relaxation intervals of increasing size.

ILLUSTRATION of TIME-ALIGNMENT RESULTS

A prominent feature of the left-hand side graph of Figure 1 is that the global dispersion measured about F_3-based prototypes is relatively large (ranging from 40 to 70 Hz), and may therefore be interpreted in terms of the inappropriateness of F_3 as a base formant for synchronous alignment of the repetition-contours of the F_1, F_2 and F_3 of the diphthong /aɪ/ produced by one of our speakers. By contrast, the

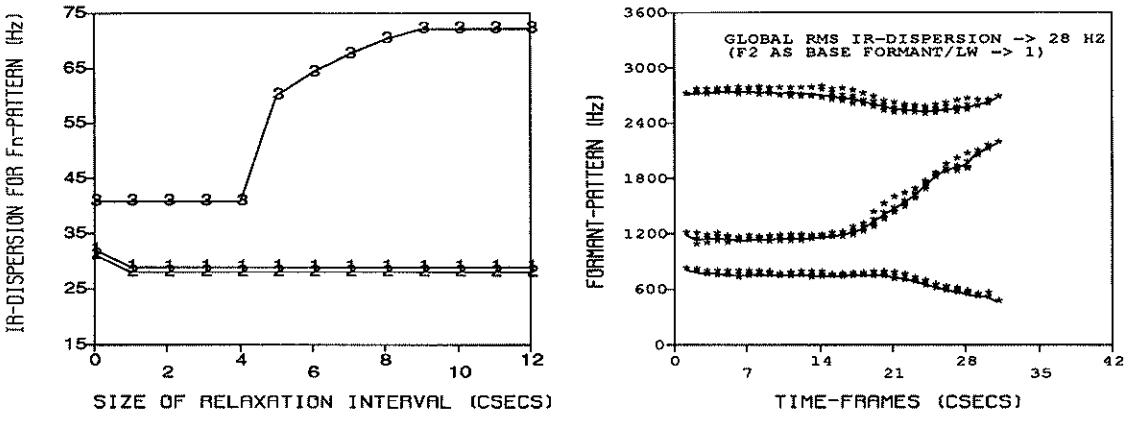

Figure 1: *Left graph*: Global IR-dispersions (Hz) as functions of relaxation intervals l_w. An l_w = 0 or > 0 implies a fixed or relaxed alignment-interval, respectively. The solid lines drawn through the 1's, 2's and 3's symbolise global IR-dispersions for F_1-, F_2- and F_3-based synchronous prototypes, respectively. *Right graph*: Prototype F_n-pattern of /aɪ/ produced by one of our speakers and obtained with the optimum alignment parameters (F_2 as base formant and l_w = 1) indicated by the minimum of the three curves. The prototype-contours of the F_1, F_2 and F_3 are shown (as solid lines) amidst clusters of points (shown as stars) which make up the repetition-contours.

left graph of Figure 1 also indicates that a better synchronous alignment (with a global RMS dispersion of 28 Hz) of the F_n-pattern can be expected if a relaxation interval (l_w) of 1 centisecond is allowed and if the base formant is selected to be F_2 (or even F_1). The resulting prototype F_1-, F_2- and F_3 contours (shown on the right-hand side graph of Figure 1) are seen to pass through tight clusters of the points which make up the repetition contours.

The curves of global dispersions yielded by our time-alignment method have been found (Clermont, 1991) to aptly indicate: (1) the base formant(s) best suited to synchronous alignment of multi-repetition contours of a diphthongal F_n-pattern; and (2) the size of the optimum relaxation interval.

TRANSITION CHARACTERISTICS OF DIPHTHONGAL F_n-PATTERNS

The description of diphthongal F_n-patterns has been predominantly based on very sparsely-sampled contours of the F_1 and F_2, while very little seems to be known about the F_3, let alone a densely-sampled F_3-contour. Some intriguing questions arise from this perspective (Clermont, 1991) on previous work: (1) Are transition characteristics of the diphthong sufficiently captured by means of two or four points selected near the initial and the final vowel regions attained or approached? and (2) Is the F_3-contour of the diphthong so unimportant?

The top left graph of Figure 2 illustrates (Bernard, 1970) a common description of diphthong formant-contours offered in previous studies. It consists of sampling the contours at four points, the result of which is a segmentation in three parts referred to as onglide, glide and offglide, or as onset steady-state, glide and offset steady-state, or as initial steady-state, glide and final steady-state. The middle left graph of Figure 2 illustrates a more objective and slightly more elaborate segmentation used by Holbrook and Fairbanks (1962), who selected three (equidistant between onset and offset points) rather than two interior points along F_1-, F_2- and F_3-contours, thus effectively segmenting each of these three contours in four parts. The description based on two-point sampling, which also appears in a number of studies of diphthongs (Gay, 1968 (American English); Burgess, 1969 (Australian English); Jha, 1985 (Maithili, a modern Indo-Aryan language); Toledo and Antoñanzas-Barroso, 1987 (Argentinian Spanish)), is intuitively the least satisfying since it cannot capture the time-varying behaviour expected of diphthong formants.

In sharp contrast, the bottom left graph of Figure 2 shows our densely-sampled F_n-pattern of the diphthong /ɔɪ/, the details of which strongly suggest that our knowledge of the contour movements of diphthong formants has been incomplete. Although the trend in previous studies is to very sparsely describe mostly the F_1 and F_2-contours of the diphthong, the five-point sampling used by Holbrook and Fairbanks raises the hope of re-interpreting their data in light of our own findings.

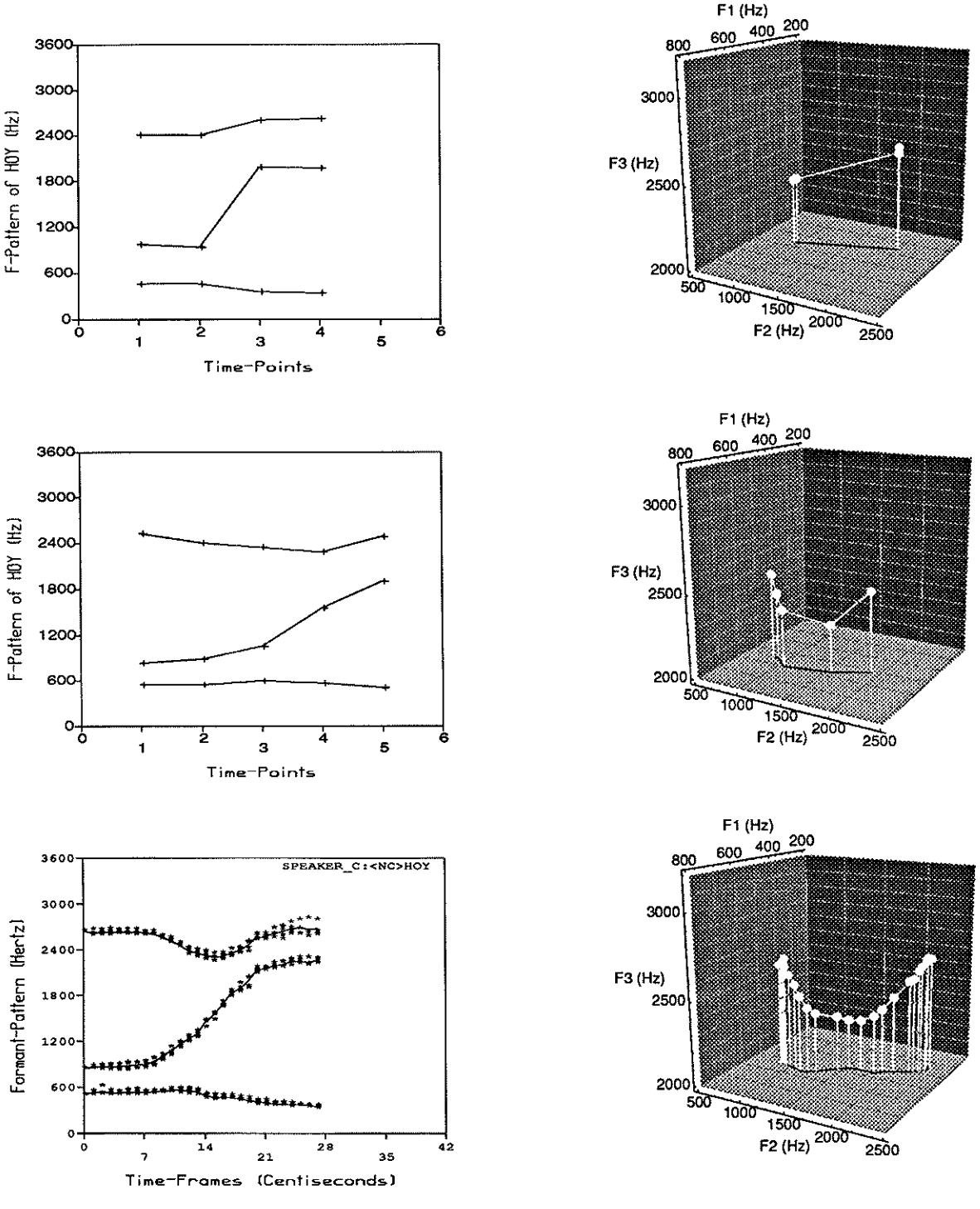

Figure 2: *Top, middle and bottom graphs* concern the diphthong /ɔɪ/ produced, respectively, in Australian English (Bernard, 1970), American English (Holbrook and Fairbanks, 1962), and Australian English (Clermont, 1991). *Left graphs* illustrate F_n-patterns of /ɔɪ/. Bernard's *4-sample* F_n-pattern is an average obtained from /hVVd/ syllables spoken by 50 adult male speakers; Holbrook and Fairbanks' *5-sample* F_n-pattern is an average measured from /hVV/ syllables spoken by 20 adult male speakers; Clermont's *detailed* F_n-pattern is shown as a set of lines amidst points (shown as stars) of the repetition-contours. This detailed F_n-pattern is an average (over 5 repetitions spoken by one adult male speaker) obtained by the contour-alignment method described herein. *Right graphs* are 3-D representations of the F_n-patterns. The white balls standing on the white stalks are individual points of the 1-D graphs. The projection of the balls onto the familiar F_1-F_2 plane appears as a dark interconnecting trace. The F_2-F_3 plane of interest here is the darkest side of the F_1-F_2-F_3 space.

One can observe in the bottom left graph of Figure 2 that curvilinearity is most pronounced in the F_2-contour of /ɔɪ/, not at all negligible in the F_3-contour, and relatively small in the F_1-contour. The same observation can be made for the F_n-pattern of /aɪ/ shown in the right-hand side graph of Figure 1. The notable curvilinearity of the F_3-contours of the back-to-front diphthongs (/aɪ/ and /ɔɪ/) raises the question of the universality of terms such as "onglide, glide and offglide", which are traditionally used to describe successive segments of diphthong formant-contours.

While the term "glide" adequately describes the intermediate segment in F_2 (and in F_1 to some degree) of /ɔɪ/ and /aɪ/ between initial and final steady-states, it misrepresents the F_3-movement. The common usage of the term "glide" to describe the primary feature of diphthongal F_n-patterns is therefore further evidence of a consistent neglect of F_3-contours in previous acoustic-phonetic studies of diphthongs. While the first three formant-contours of the diphthong /ɔɪ/, for example, have in common initial and final segments which may be called steady-states, the transition segment is manifest differently in the F_3-contour. The transition segments of the F_1 and F_2-contours resemble, in their most schematic form, ramp functions with endpoints curving in and out of steady-states. In contrast, the *V-shaped* transition segment of F_3 is characterised by a slow fall and rise into and out of a wide minimum, which points down toward the inflection point (near 1500 Hz) of the F_2-contour.

It is clear from the detailed F_n-pattern shown in the bottom left graph of Figure 2 that a straight-line representation between onset and offset points of the contours as sometimes used in previous work would yield not only a crude, but also an inaccurate description of the formant movements of diphthongs. Further, while a four-point sampling (two points near onset and offset steady-state segments, respectively) such as that used in most previous acoustic-phonetic studies (Lehiste and Peterson, 1961; Bernard, 1970; Gottfried, 1989) renders a better approximation of the F_1 and F_2-contours, it still is an inaccurate representation of the F_3-contours of the diphthongs /aɪ/ and /ɔɪ/, which would be better described with at least an additional point near the middle of the transition segment. In this sense, a five-point sampling similar to Holbrook and Fairbanks' is at least required in order to approximate all first three formant-contours of diphthongs. However, even with a slightly less sparse sampling, Holbrook and Fairbanks describe F_3-contours as "staying relatively constant" or "tending to follow" F_2-contours. Clearly, this is not the case for the diphthongs /aɪ/ and /ɔɪ/, the densely-sampled F_3-contours of which are observed to be not constant and to be notably different from the F_2-contours.

It emerges from our detailed, temporal representation of the F_n-patterns of /ɔɪ/ and /aɪ/ that transition cues are to be sought not only along the F_2-contour as established in previous work, but also along the F_3-contour, which is observed to pass through a minimum occurring near the point of maximum change of the F_2-contour. It is this finding which prompted us to examine the F_2-F_3 plane in F_1-F_2-F_3 space, with a view to gaining more insights into the nature of the transition from the initial to the final vowels of a diphthong.

The top right (3-D) graph of Figure 2 clearly brings into evidence the serious consequences which can be expected from sparse time-sampling of diphthongal F_n-patterns, while the transition characteristics are observed to be enhanced (bottom right graph) by a detailed temporal representation. Note further that our 3-D re-interpretation (middle right graph) of Holbrook and Fairbanks' data lends support to these authors' attempt to acknowledge the gradual transition from the onset to the offset vowels. However, the *V-shaped* movement between two distinct areas of the formant space is more clearly defined in the F_2-F_3 plane of our densely-sampled 3-D space, and is seen to convey more information about the (vowel-to-vowel) transition involved than the quasi-linear projection onto the familiar F_1-F_2 plane. One can indeed observe that the overall transition in the F_2-F_3 plane of the 3-D formant space of the diphthong /ɔɪ/ is characterised by a non-linear movement, with a change of direction taking place near a point corresponding to the minimum of the F_3-contour and to $F_2 \simeq 1500$ Hz (i.e., the point of maximum change of this formant).

CONCLUDING DISCUSSION

We have first illustrated the revealing power of dense time-sampling, when it is applied to the F_n-patterns of a subset of the Australian English diphthongs (/aɪ/ and /ɔɪ/). While sparse time-sampling can be justified as a means to provide an economical description of diphthong F_n-patterns, consequences of this persistent approach have been more serious than realised in previous work. A very good example of this is the V-shaped curvilinearity of the F_3-contours of /aɪ/ and /ɔɪ/, which can be irreversibly attenuated by sparse sampling and which has therefore escaped researchers' attention to date.

We have also provided evidence that the F_2- and F_3-contours of at least the back-to-front diphthongs embody more prominent cues of the vowel-to-vowel transition involved. Thus, in contrast to the F_1–F_2 plane which has long been used to study the end-vowels attained or approached during diphthong production, the new perspective offered here suggests that the F_2–F_3 plane may be expected to play a significant role in characterising the gradual transition between the initial and final parts of a diphthongal sound.

ACKNOWLEDGEMENTS

I wish to thank Dr Iain Macleod, Dr Bruce Millar, Professor Richard Brent and Dr Michael Wagner for their encouragement and support. I also thank Dr Andrew James for his help in generating the 3-D graphs.

REFERENCES

Bernard, J.R.L. (1970), "Toward the acoustic specification of Australian English", *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 23*: 113-128.

Burgess, N. (1969), "A spectrographic investigation of some diphthongal phonemes in Australian English", Language and Speech *12*: 238-246.

Clermont, F. (1991), "Formant-Contour Models of Diphthongs: A Study in Acoustic Phonetics and Computer Modelling of Speech", *Doctoral Thesis*, The Australian National University: Research School of Physical Sciences and Engineering, Computer Sciences Laboratory.

Clermont, F. (1992a), "Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching", Proceedings of the IVth Australian International Conference on Speech Science and Technology.

Clermont, F. and Butler, S.J. (1988), "Prosodically guided methods for nearest neighbour classification of syllables", Proceedings of the IInd Australian International Conference on Speech Science and Technology: 216-221.

Gay, T. (1968) "Effect of speaking rate on diphthong formant movements", Journal of the Acoustical Society of America *44(6)*: 1570-1573.

Gottfried, M. (1989), "Some acoustical properties of diphthongs", Journal of the Acoustical Society of America *86*: S123.

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

Jha, S.K. (1985), "Acoustic analysis of the Maithili diphthongs", Journal of Phonetics *13*: 107-115.

Lehiste, I. and Peterson, G.E. (1961), "Transitions, glides, and diphthongs", Journal of the Acoustical Society of America *33(3)*: 268-277.

Peterson, G.E. and Barney, H.L. (1952), "Control methods in a study of the vowels", Journal of the Acoustical Society of America *24*: 175-184.

Pols, L.C.W., van der Kamp, L.J.T. and Plomp, R. (1969), "Perceptual and physical space of vowel sounds", Journal of the Acoustical Society of America *46*: 458-467.

Potter, R.K. and Peterson, G.E. (1948), "The representation of vowels and their movements", Journal of the Acoustical Society of America *20*: 528-535.

Potter, R.K. and Steinberg, J.C. (1950), "Toward the specification of speech", Journal of the Acoustical Society of America *22*: 807-820.

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

Toledo, G.A. and Antoñanzas-Barroso, N. (1987), "Influence of speaking rate in Spanish diphthongs", Proceedings of the XIth International Congress of Phonetic Sciences: 125-128.