

Acoustic Structure and Perceptual Relevance of 'Steady States' and 'Glides' within Formant Trajectories of Diphthongs, Complex Vowels, and Vowel Clusters.

W.J.M. PEETERS

The traditional description of diphthongs can be given as follows, c.f. CATFORD (1977):

"(...) a diphthong may, (...) consist, of two distinct, discrete 'elements' with a relatively rapid transition between them. On the other hand, it may be more correctly characterized as a continuous, gliding movement from a starting point to a finishing point. There are thus two extreme types of diphthong, a 'sequential' type, (...), and a 'gliding' type, (...), with, of course, a continuum of possible gradations between these extremes. (...) A (...) more traditional (...) division of diphthongs is into falling and rising. It is important to note that these terms, contrary to expectations, do not refer to the direction of the transitional or gliding diphthongal movement. What they do refer to is the relation of the diphthong to the 'stress curve', or initiator-power pulse with which it is associated. A falling diphthong is one with what may be called 'decrescendo stress', (...)." (215-6).

At an early stage of phonetic research these findings were confirmed: theoretical and articulatory insights as well as experimental analysis yielded the following definitions: 'only glide' (Brücke 1856); 'full vowel plus glide' (Sweet 1877); 'full vowel plus glide plus full vowel' (Merkel 1866, theoretically), (Donders 1864, Martens 1889, acoustically), and (Wagner 1889, Rousselot 1901-08, physiologically); 'three different structures' within a broader range of complex vowels (Meyer 1903, physiologically); 'vowel directly followed by another vowel' (Menzerath 1941, perceptually); and again 'two vowels connected by a glide' (Potter e.a. 1947, Visible Speech).

The problem of describing diphthongs as language specific speech entities is still far from being solved.

It is still problematic to use physiological data for obtaining invariant diphthong models across speakers and languages. Experimental research has, however, made clear that a diphthong movement is not a ballistic but a controlled motor movement, c.f. MACNEILLAGE and SHOLES (1964), SOVIJÄRVI (1969).

Acoustic data are available in huge quantities but do not produce satisfying solutions. Analytical methods such as 'Visible Speech' (spectrography) have given a better insight into the structure of diphthongs (wide closing types), simple vs. complex vowels (narrow closing diphthongs), not, however, into vowel clusters; for detailed criticisms on this matter see: PIKE (1947), JOOS (1948).

But analysis has remained tricky: the frequency values of the first two (sometimes three) formants are described in terms of their onsets and offsets, which are then related to formants of comparable vocoids from one's own language or other languages. But researchers are in most cases unaware of the fact that they are strongly influenced by orthography and other cognitive factors. Consequently, identical or nearly identical -cross linguistic- acoustical patterns are related to their own specific language systems and interpreted differently. The observed 'undershoot' of diphthong offset targets can be explained in this way, c.f. LINDBLOM and STUDDERT-KENNEDY (1967).

Assistant Professor at the German Department, University of Utrecht (NL).

A complicating acoustical factor is perhaps the limited spectral resolving power for F1 and F2 in traditional spectrography: these formants are depicted only as gross shapes, bearing nevertheless the invariant character of diphthongs, c.f. BLADON (1982); another factor which has to be stressed is the lacking of investigations into parameters such as Fo, overall amplitude, and formant bandwidths, c.f. PETERS and WATKINS (1984).

Perceptual research into diphthongs has always consisted of optical, electrical and digital gating techniques on real, resynthesized or, synthetic speech samples, using various kinds of manipulations such as deleting the 'glide' between the 'discrete' diphthong elements and truncation of the diphthong signal, starting at various points in a given diphthong. MENZERATH (1941) was the first to undertake perception tests by manipulating, in a very sophisticated manner, the optical track of sound films; truncation tests were done, as well as glide deletion in order to synthesize a diphthong with two steady-states only. The set-up of the perception experiment was, however, very poor; the results were defended with vehemence and aggression causing distress among phoneticians and linguists, insofar, in the circumstances of World War II, the work was accessible to them. Menzerath defined a diphthong as 'two discrete elements without a gliding movement.' The interpretation is questionable, particularly with respect to the techniques he used.

GAY (1967) used synthesized speech (Pattern-Playback, Haskins Labs.) to analyze the structure of American English diphthongs. It is a pity that Gay used only continuously gliding synthetic stimuli. He denies the existence of other diphthong types both on acoustical and on perceptual grounds; in contrast to Menzerath he stipulates that '(...) a gliding movement alone, exclusive of steady state targets, is sufficient for providing diphthongal quality.' In his experiments the only possible outcome could be: 'that the formant rate of change of the gliding movements (...) is a fixed feature', which means that diphthongs are 'each characterized primarily by an invariant speed of articulatory movement.'

GERBER (1972) oscillates between these standpoints, depending on the type of experiment! BOND (1982) combines these standpoints: '(...) in order to identify vocalic nuclei as diphthongs, listeners need two targets; either a glide beginning and ending at those targets or, alternatively, the targets without a detectable glide prove to be sufficient.'

BLADON (1985) used, in his experiments, 'curtailed stimuli' which he presented, unlike Gerber, without their original durations; since the Germanic diphthong type is an entity (a view supported by the typical behaviour of Fo, amplitude, overall duration, effects of overall duration on all diphthong components, effects of context on all diphthong components, data on speech errors) these experiments cannot solve the problems of auditory and phonetic hearer strategies. This goes also for his experiments with 'transitionless diphthongs': these are not speech-like sounds anymore. That the 'transition-only diphthongs' were difficult to identify is due to the unrealistic temporal relationship between the glide and the remaining speech signal carrier. Another problem arises in separating the transition from the steady states.

Speech technology is an indispensable tool for achieving a better understanding of problems in foreign language learning, for example of the sub-phonemic variants of one sound type that are language specific, and for giving explanatory power to evolutionary sound processes, e.g. diphthongization. For these reasons experiments have to be carried out with stimuli containing the properties of real speech.

Therefore we designed a stimulus continuum in the years 1982-83 that happened to be, independently, almost identical with the one presented by COLLIER and 'T HART (1983). A slightly modified model of their design is shown below. The stimulus parameters, however, were determined differently. The relevant parameters (ampl., Fo, F1-B1, F2-B2, F3-B3, F4-B4, F5-B5) were LPC synthesized in order to obtain continua of diphthongs, simplex-to-complex vowels, and vowel clusters.

Considering the number of stimuli that would have been caused by a great variability of parameters we gave all stimuli a fixed duration of 240 ms. To test the hypothesized perceptually higher position in the hierarchy of the time-related division of the formant trajectory parts, the onset and offset frequency values were fixed. Only the inner structure of steady states and glides were varied in steps of 20 ms. This goes for /ai/, /au/, /e--e /, and /o--o /.

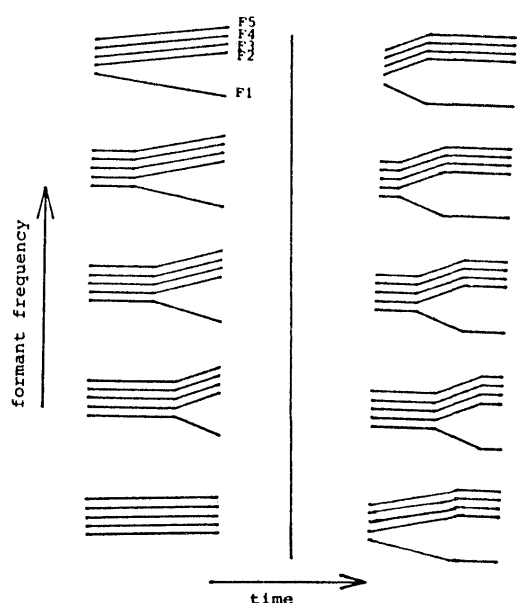


Fig. 1. Stimulus continuum. C.f. COLLIER and 'T HART (1983), p. 35. Example of an 'ai'-continuum.

It took quite a long time to synthesize stimuli of good quality. After the equipment of the Institute of Perception Research (IPO), Eindhoven NL, had been made accessible progress was much faster. The whole stimulus continuum was synthesized and each stimulus was compared in steps of +40 ms and -40 ms (as far as possible) with neighbouring stimuli. The steps were taken by keeping either the transition time or the onset steady state duration constant, or by varying both. The stimulus pairs were presented binaurally via headphones from a Revox A-77 tape recorder, and had, in randomized order, to be compared in A-B and B-A mode. They had to be evaluated in a forced choice procedure. 55 Stimuli were synthesized according to the continuum. For the 'au'-type diphthong 142 pairs could be generated. A first result with German natives follows here: there were 9 subjects, one, however, could not accomplish the task which was considered difficult by all

subjects. Nobody noticed that the stimuli were synthetic! The result was fairly remarkable: for 'au'-diphthongs the types with onset steady states of 60 and 80 ms were preferred; for 'ai' those of 80, 100 and 120 ms. 80% of the preferred 'au' types (5) had an offset steady state of 60 ms, whereas only 40% of the preferred 'ai' types (5) had this value, but never less than 40 ms. A few examples may illustrate the perception behaviour of these German subjects who performed their task very well:

an 'au' diphthong type with duration values for steady state/transition/
steady state of 80/100/60 ms compared to a 120/100/20 type: 100% vs. 0%;
in B-A mode: 0% vs. 87.5%,
12,5% could not make a decision;
compared to a 40/100/100 type: 61.5% vs. 37.5%,
in B-A mode: the result was symmetrical;
compared to a 80/140/20 type: 87.5% vs. 12.5%,
in B-A mode: 12.5% vs. 75%,
12.5% could not make a decision.

The parameter values of the 80/100/60 type are shown in Fig. 2.

For 'ai' the results were less clear. However, the 80/120/40 type compared to a 120/120/00 type scored 87.5% vs. 12.5%; B-A: a symmetrical result was obtained. An other pair did give only one clear result: a 80/120/40 type compared to a 80/160/00 type scored 75% vs. 25%, but B-A: 50% vs. 50%. Therefore it was concluded to throw out the pairs differing in both transition and steady state duration, and to double the remaining pairs. These new tests will be done with Dutch, English, and German speaking natives.

The results obtained with these continua that were presented to German subjects are completely different from the results of COLLIER and 'T HART (1983) with Dutch speaking subjects: '(...) there is no need for a steady state final part, this suggests that, perceptually, a diphthong is interpreted as a monophthongal vowel (short or long), followed by a glide.' It is needless to say that this statement cannot be THE definition of a diphthong. The perceptual results are not in accordance with those of GAY: '(...) There is no clear perceptually preferred rate of change for the formant trajectories in Dutch diphthongs. Admittedly, there is a minimum duration of the transition of about 100 ms.' The latter observation can be supported.

It is still too early to draw conclusions. Perhaps diphthong production is controlled by 'indirect auditory targeting', c.f. GAY, LINDBLOM and LUBKER (1981), based on '(...) auditory patterns in time (...) dependent on interactions obscured by the separation of component frequencies,' c.f. SCOTT (1976).

1	-77	50	700	200	1100	140	2070	120	3398	200	4348	600
2	-77	126	700	200	1100	140	2070	120	3398	200	4348	600
3	-78	315	700	200	1100	140	2070	120	3398	200	4348	600
4	-79	792	700	200	1100	140	2070	120	3398	200	4348	600
5	-79	1991	700	200	1100	140	2070	120	3398	200	4348	600
6	-80	1991	700	200	1100	140	2070	120	3398	200	4348	600
7	-80	1622	700	200	1100	140	2070	120	3398	200	4348	600
8	-81	1322	700	200	1100	140	2070	120	3398	230	4338	590
9	-82	1077	670	190	1063	134	2081	145	3398	260	4328	580
10	-82	878	640	180	1027	129	2093	170	3398	290	4319	570
11	-83	715	610	170	990	123	2105	196	3398	320	4308	560
12	-83	583	580	160	954	118	2117	221	3398	350	4298	550
13	-84	475	550	150	918	112	2129	247	3398	380	4289	540
14	-85	387	520	140	881	107	2140	272	3398	410	4278	530
15	-85	315	490	130	845	101	2152	298	3398	440	4268	520
16	-86	257	460	120	809	96	2164	323	3398	470	4258	510
17	-87	210	430	110	772	90	2176	349	3398	500	4248	500
18	-87	171	400	100	736	85	2188	374	3398	500	4248	500
19	-88	139	370	90	700	79	2200	400	3398	500	4248	500
20	-89	113	370	90	700	80	2200	400	3398	500	4248	500
21	-89	92	370	90	700	80	2200	400	3398	500	4248	500
22	-90	75	369	90	700	80	2200	400	3398	500	4248	500
23	-91	61	370	90	700	80	2200	400	3398	500	4248	500
24	-91	50	370	90	700	80	2200	400	3398	500	4248	500
1	2	3	4	5	6	7	8					

Fig. 2.

- | | |
|-------------------------------|-----------|
| 1: frame number (each 10 ms), | 5: F2-B2, |
| 2: $-10.000/F_0$, | 6: F3-B3, |
| 3: relative amplitude values, | 7: F4-B4, |
| 4: F1-B1, | 8: F5-B5. |

ACKNOWLEDGEMENTS: I am indebted to Dr. M.E.H. Schouten, Institute of Phonetics, Utrecht, and to Dr.Ir. L.L.M. Vogten, IPO, Eindhoven, for their inspiring ideas and support.