Characterizing the formant movements of German diphthongs in spontaneous speech

Adrian P. Simpson

Abstract

Die Charakterisierung der Formantbewegungen von Diphthongen in Spontansprache wirft eine Reihe von Problemen auf. Im Gegensatz zu Lesesprache kann die Elizitation von Spontansprache einen verhältnismäßig geringen Einfluß auf die zeitliche Umgebung und die Häufigkeitsverteilung eines bestimmten Phänomens nehmen.


The characterization of formant movements in spontaneous speech raises a number of problems. In the elicitation of spontaneous speech – in contrast to read speech – the analyst has relatively little control over the temporal context and frequency of occurrence of the phenomenon under investigation.

A method of measuring diphthongal formant movements is described which takes into account large durational differences between individual vocalic portions and at the same time allows a comparison of average formant movements for diphthongs of different durations to be made. The three ‘real’ German diphthongs aɪ, aʊ and ɔʏ are analysed using data from the *Kiel Corpus of Spontaneous Speech*. Finally, the results are compared with findings of other descriptive and perceptual studies.

1 Introduction

Providing adequate acoustic characterizations of diphthongs is a problematic enterprise. Whereas the monophthongs of a language can often be described by averaging over the spectral analyses made at the mid-points of a number of vocalic portions, there is still no agreement as to what it is that characterizes a diphthong. The conclusions reached in Gay’s (1968, 1970) descriptive and perceptual studies stand in apparent conflict with Bladon’s (1985) perceptual findings. The problem of finding an appropri-
ate acoustic characterization for diphthongs is reflected in the many methods of measuring formant movement. The simplest method (e.g. Lehiste and Peterson 1961) identifies formant frequencies at vowel onset and offset and most closely reflects the general phonetic and phonemic representation of diphthongs in terms of two vowel symbols representing initial and final auditory vowel qualities. In their acoustic and electromyographic study of Dutch diphthongs Collier, Bell-Berti, and Raphael (1982) identify formant frequencies at onset, offset and at vowel midpoint. Both Holbrook and Fairbanks (1962) and Narahara, Shimoda, and Okamoto (1977) use five measurement points: initial, final and three equidistant points inbetween. Gay (1968) employs the most theory-driven method of all, establishing onset and offset frequencies and rate of movement in the glide phase as well as duration of any initial and final steady state phases.

The object of this study is to present a description of the formant movements in the three ‘real’ German diphthongs ai, au, oy as they are produced in spontaneous speech. The use of spontaneous speech produces a range of further problems. Under the controlled situation of laboratory speech we can control the linguistic and rhythmic context as well as the quantity and distribution of the phenomenon being investigated. In spontaneous speech we can only control partly for these aspects by imposing strict limitations on scenario being used to elicit the spontaneous speech (e.g. Swerts and Collier 1992). However, these problems are far outweighed by the need to examine the acoustic and articulatory patterns of spontaneous speech.

2 Data

The data for this study are drawn from the Kiel Corpus of Spontaneous Speech (IPDS 1995-97). This corpus is made up of dialogues in which speakers carry out an appointment-making task (Pätzold et al. 1995; Kohler et al. 1995). Each speaker pair completed seven short dialogues arranging appointments in each on the basis of academic time-tables and diaries covering a two-month period. The Kiel Corpus of Spontaneous Speech contains 16 complete dialogue settings as well as single dialogues from 5 further settings. Only the data from the 16 complete dialogue settings are used here. The 32 speakers divide into 14 female and 18 male speakers.

Even though the dialogues are confined to the relatively narrow semantic field of appointment-making the frequency of distribution of vowels and
consonants is comparable to that documented for poetry and prose (Meier 1967). The implications of this for diphthongs are that out of 5058 diphthong tokens, ai has 3704, au 1094 and ey only 260. When tokens are grouped into 20 ms wide bands (see Figure 1) the most frequent durational group for both ai and au is 80–100 ms, and 120–140 ms for ey. The low frequency of occurrence of ey and the higher distributional peak can both be attributed to the absence of ey in function words in the corpus. ey only occurs in the informal personal plural pronoun euch and other related forms (euer, eure), none of which feature in the dialogues. However, even if function word diphthongs are excluded the peak distribution for ai and au only shifts up to the 100–120 ms band.

3 Method

Measurements made in diphthong portions should as far as possible only reflect the formant movements which we would want to associate with the
correlates of the diphthong and not other surrounding material. Furthermore, the uncontrollable environment of spontaneous speech demands a method of measurement which is temporally flexible, but at the same time enables measurements from a number of diphthong portions to be grouped together and averaged over.

The following method extends and modifies Holbrook and Fairbanks (1962) five-point method. Any diphthong portion less than 60 ms in length is treated technically as a monophthong and is characterized by a single set of formant measurements taken at the mid-point of the vowel. Any diphthong portion with a duration ≥ 60 ms is measured at two or more points. The first and last measurements are taken at points 20 ms after the beginning and 20 ms before the end of the vocalic portion. This is designed to avoid the most immediate effects of transitions into and out of adjacent consonant and vowel articulations which are included in the segmented vocalic portions of the Kiel Corpus. The remaining measurement points between the first and last are then calculated by dividing the duration of the measurement stretch by the whole number of notional 20 ms portions. Each diphthong portion within a 20 ms wide band is then automatically assigned to a durational group characterized by a fixed number of measurement points. A set of formant movements for each group can then be simply calculated by averaging over formant measurements at the same points in each group. Figure 2 shows the five points at which formant measurements are taken in an \textit{a} token from the word \textit{Zeit} ("time"). \textbf{Beg} and \textbf{Ende} mark the start and end of the vocalic portion as segmented in the data base. Points \textbf{A} and \textbf{E} are the first and last measurement points taken 20 ms from the start and the end of the vocalic portion. Points \textbf{B–D} represent the remaining equidistant points between \textbf{A} and \textbf{E}.

Values for F1 and F2 were determined automatically using analysis and sorting similar to that described in Scheffers and Simpson (1995). At each measurement point LPC coefficients were used to provide two estimates of formant frequencies (Vogten 1983; Willems 1987). Reference frequencies of female and male neutral vocal tract configurations together with bandwidth information were then used to establish which frequency estimates were the most likely candidates as values for F1 and F2.

4 Results

Figure 3 contains average formant tracks for F1 and F2 for a selection of the duration groups for female (left) and male (right) diphthongs. The
Figure 2: Measurement points for an *ai* token with a duration of 137 ms from the word *Zeit* (“time”). **Beg** and **Ende** mark the vocalic portion boundaries; the lines **A**–**E** mark the points at which formant measurements are taken (see text).

course of F1 and F2 for each durational group are plotted against the lower time boundaries for that group. So, for instance, the duration group 140–160 ms is represented by 6 measurement points plotted at 20 ms intervals from 20 ms to 120 ms, with the last point representing the last measurement point for a token with a duration of 140 ms. The last group to be plotted in each case usually only has around 10 tokens and average courses of any groups beyond this become unreliable.

For all three diphthongs we can identify a number of patterns. All duration groups for all diphthongs contain a glide. In the shorter duration groups this is the only element present. As duration increases the extreme positions reached by each formant become increasingly further apart suggesting articulatory movements which reach more open positions in the initial phase and end closer finally. With increasing duration we can also see a slowing down of formant movement towards the beginning and for *ai* and *au* a slowing down towards the end, which can be equated with a steady state period.
Figure 3: Average formant tracks for F1 and F2 for female and male diphthongs. Each course is plotted against the lower boundary of the durational group (see text).
The longer duration groups (from 180 ms upwards) for both female ar as well as female and male au display further movements of F2 towards a more centralized value. For both ar and au this movement of F2 can be caused by a movement of the tongue towards a more central position, however in the case of au it may accompanied by or caused solely by lip-spreading. The reasons for centralization/lip-spreading can also have two different origins. Phrase-final centralization of syllable-final ar would seem to be the most likely cause of F2 movement for long female ar tokens. Of 170 female ar tokens with a duration greater than 220 ms more than a third occurred in phrase- and syllable-final context (frei, 27; Mai, 17; vorbei, 12; Julei, 7; drei, 6). A further possibility is the earlier start of formant transition into adjacent consonant and vowel articulations, which then become visible despite the 20 ms margin. This would seem to provide a more likely explanation for the final F2 movement in longer au tokens.

5 Discussion

The results paint a complex yet systematic picture of formant movements in the three German diphthongs ar, au and ey. The glide element is present throughout, the size of its excursion increasing with duration. Steady state periods of slower formant movement appear with increasing duration. These patterns match up well both with Gay’s (1968) observations of American diphthongs spoken at different rates as well as his perceptual findings (Gay 1970). While the initial and final steady state periods were sacrificed at faster rates, a glide element exhibiting approximately the same rate of movement was maintained in tokens of different duration.

It is interesting to consider our findings in the light of perceptual experiments described in Bladon (1985), which in turn set out to refute the claims made by Gay that the glide phase is the essential element of a diphthong. Bladon claims on the basis of three experiments that the qualities at the beginning and end of the diphthongal movement are the most important elements and the glide between is of only secondary importance. In the first two experiments phonetically trained subjects were asked to transcribe as narrowly as possible isolated diphthongal stimuli created by successively cutting away portions at the beginning (exp. 1) and end (exp. 2) of a closing diphthong. In a third experiment subjects were asked to recognize diphthongal stimuli embedded in a sentence frame. The stimuli were of two types. In the first the glide phase had been excised leaving only a hard transition from initial to final quality; in the second steady state portions
at the beginning and end were removed leaving only the glide. Results of all three experiments appeared to indicate the importance of initial and final quality in a diphthong. In the first two experiments the subjects identified initial qualities becoming closer as more of the diphthong was removed and identified final qualities becoming increasingly more open as the end was removed. In other words the slope of the glide itself which remained unchanged was insufficient to maintain a constant diphthongal quality. In the third experiment subjects were more successful at correctly identifying the glideless stimuli, even finding them more natural despite their artificial form.

These results seem to contradict our findings as well as Gay’s descriptive and perceptual results. A closing diphthong such as air can be nothing more than a glide and does begin and end at different acoustic and articulatory locations, but is still successfully recognized as being an air token. A closer look at the linguistic status of Bladon’s stimuli as well as the temporal contexts which the stimuli were placed in, provides a plausible explanation of the apparent contradiction. Stimuli from the first two experiments were judged on phonetic and not linguistic terms, i.e. phonetically trained subjects were required to provide general phonetic classifications of sounds out of the context of any language. In the third experiment the stimuli were indeed given linguistic context and subjects were required to classify them linguistically, however in this case the glide-only stimuli were temporally inappropriate, i.e. they had the wrong shape for the temporal context in which they were situated. The glideless stimuli on the other hand provided the necessary steady state portions and hearers were perceptually in a position to reconstruct the glide between the steady states both in terms of acoustic content and temporal structure. Indeed, had the subjects in Bladon’s first two experiments been required to classify isolated stimuli in terms of the phonological categories of English they would probably not have done better, because again the shape of the formant movements were inappropriate to their temporal context.

The descriptive results of this study show that any perceptual experiment which manipulates the form of a diphthong must pay careful attention to the temporal context that the stimuli are placed in. Combining our findings with a reinterpretation of Bladon’s results we can predict that hearers would give a different set of responses for the same diphthongal stimulus in different temporal environments.
References


