DIPHTHONG DYNAMICS: PRODUCTION AND PERCEPTION IN SOUTHERN BRITISH ENGLISH

William J.M. Peeters*, William J. Barry**

*Research Institute for Language and Speech (OTS), University of Utrecht.
**Department of Phonetics and Linguistics, University College London.

Abstract

The preferred temporal structure of the isolated diphthongs /ai/ and /au/ in Southern English is examined by a means of paired comparison test using a comprehensive range of dynamic trajectory variants. This is compared with the temporal structure of those diphthongs produced by 10 speakers of Southern English. Results indicate a clear language-specific trajectory, differing from those found in British English and Dutch listeners. The English production data are more complex in their manifestations, but many tokens approximate the same language-specific dynamic structure.

Introduction

Until recently structural research into the parameters which constitute the diphthong as a perceptual rather than a functional unit was not undertaken systematically. Peeters (1987, 1989, forthcoming) shows that the inner temporal structure (the typical formant trajectory) of a diphthong probably cues its naturalness, and consequently its acceptability. However, little or nothing is known either about the language-specific temporal structure or about the relationship between the production and perception of dynamic sounds. This paper is a first study of these two aspects in Southern British English /ai/ and /au/. Results are presented of a perceptual experiment using a comprehensive series of synthetic diphthong glides, and of the spectral analysis of /ai/ and /au/ diphthongs produced by 10 speakers of Southern Standard English.

Perception Experiment

Identification tasks with temporally manipulated diphthongs have proved to be too easy a task to provide any indication of correct diphthongal structure. Even severely manipulated diphthongs show very high recognition rates. To illuminate the language-specific features of temporal structure the perception test was constructed as a preference task within a paired comparison frame.

A stimulus continuum was produced (as part of the doctoral dissertation study of the first author) varying only in the temporal structure of the diphthong formants, namely in the duration relations of 1) onset steady-state to 2) glide portion to 3) offset steady-state (see figure 1 for a representative selection of the /ai/ stimuli). Onset and offset frequencies were all constant, and F0, F4, F5 trajectories were fixed, as were the B4 and B5 (bandwidths). In addition, F3 and B3 trajectories were fixed for /au/ to avoid /l/-like percepts in some of the stimuli.

Table 1 Formant frequencies, formant bandwidths, amplitude temporal structures, and fundamental frequencies of the /ai/ and /au/ diphthongs.

Fig. 1. Selection from the /ai/-stimuli continuum. The first three formants are shown. The inner temporal division is made according to the pattern STEADY-STATE/GLIDE/STEADY-STATE; the figures given are in milliseconds.

The overall duration was set at 240ms. Table 1 gives details:

<table>
<thead>
<tr>
<th>/ai/</th>
<th>/au/</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>onset</td>
</tr>
<tr>
<td>F1</td>
<td>720</td>
</tr>
<tr>
<td>F2</td>
<td>1350</td>
</tr>
<tr>
<td>F3</td>
<td>2370</td>
</tr>
<tr>
<td>F4</td>
<td>3396</td>
</tr>
<tr>
<td>F5</td>
<td>4197</td>
</tr>
<tr>
<td>F6</td>
<td>5597</td>
</tr>
</tbody>
</table>

offset

| F1   | 380   | B1 60  |
| F2   | 2150  | B2 130 |
| F3   | 2598  | B3 450 |
| F4   | 3493  | B4 400 |
| F5   | 4145  | B5 340 |
| F6   | 5597  | B6 300 |

Ampl.

rise-time: 40ms
steady part: 70ms
decay-time: 130ms
F0: 125–85Hz

Table 1 Formant frequencies, formant bandwidths, amplitude temporal structures, and fundamental frequencies of the /ai/ and /au/ diphthongs.

EUROspeech '89, Paris, France, September 1989 1055
The stimuli paired for comparison always differed in + or - 40ms steps; either the onset portion (and with it the offset portion) or the glide portion differed within the pair. Taking figure 1 as a reference, stimuli were selected along the vertical axis in the former case, and along the horizontal in the latter.

47 untrained persons from the National Physical Laboratory, from the National Hospitals College of Speech Science, and from University College, London, took part in the test, listening to the stimuli in groups of maximally 7 over headphones (Zennoheiser HD 222 with closed dynamic system) in a quiet room. Listeners' preferences were recorded on a prepared answer sheet. The test of the /ai/ and /au/ diphthongs took 45 minutes overall, including pauses. The test was considered by subjects to be difficult but manageable; they were remunerated for their services.

Stimulus Production

The stimuli were generated at the Institute for Perception Research (IPQ) at Eindhoven, Netherlands, using LPC (LMS-equation) synthesis. Spectro-temporal waveform patterns were generated in steps of 20ms along the stimulus continuum plane of figure 1, which resulted in 39 temporal structures for /ai/ and 42 for /au/. Listings of all stimuli paired with next-to-nearest-neighbour (+ or - 40ms), were prepared to contain two of each possible pairing (2 x A-B and 2 x B-A, yielding 176 stimuli pairs for /ai/ and 200 for /au/), randomized and synthesized.

Synthesis is performed pitch-synchronously, i.e., the F1-F5 and B1-B5 values together with FO are calculated at the beginning of each new pitch period, synchronously with the excitation pulse, as a linear interpolation of the parameter values contained in consecutive 10ms frames. The number of filter coefficients is 10 with a further constant overall de-emphasis filter. The signal is LP-filtered at 5 KHz. The D-A conversion has a 12-bit resolution.

The analog signal was PCM-recorded onto digital tape using a Panasonic VHS (PAL)-video cassette recorder (NV-100-8) and a Technics Digital Audio Processor (SV-100). This processor has an in-built amplifier to which a maximum of 7 headphones was connected.

Perception Results

Significant preferences for the inner temporal structures of steady-state (SSI)/glide (GL)/steady-state (SS2) patterns were found to have the following structures.

In Dutch: /ai/ (which occurs in loanwords only) 060/140/040ms /au/ 080/140/020ms and 100/120/020ms.

In English: /ai/ 100/120/020ms, 120/120/000ms and 140/100/000ms; /au/ 100/140/000ms, 120/120/000ms and 140/100/000ms.

In German: /ai/ 060/100/080ms and 060/120/060ms; /au/ 040/120/080ms and 060/100/080ms.

(Underlined patterns were most strongly preferred). Comparing these results with those of Peters (1987) it is clear that there are definite language-specific trajectories. Compared to Dutch /au/, it is apparent that the preferred structure for English listeners requires a longer onset steady-state, and compared to German /ai/ and /au/ it appears that the preferred English structure requires a very short offset steady-state or no offset steady-state at all (off-glide only).

An interesting side-result of the experiment, which coincides with observations made by listeners with other language backgrounds, is that some stimuli in the /ai/ series were perceived as /ai/ by up to 70% of the subjects, despite the fact that the formant onset frequencies were kept constant. Although the stimuli causing these comments cannot be located definitively, after a first check the authors attributed this impression of /ai/ to those stimuli with no, or very short onset steady-states. This supports the assumption that timbre perception is the result of an integrating process over a considerable time window. Indications for such a process with speech and non-speech stimuli have been found in Brady et al. (1961), Mermelstein (1975), Huang (1986), D Benedetto (1987), and Stevens (1987).

Production

/haid/ and /haud/ utterances from 10 Southern English speakers (SF, SM) were selected from /haid/ words recorded as part of a normative study within Alvey project MMI 132 (Speech Technology Assessment, STA). Three tokens per diphthong and per person were produced, nominally with level, rising, and falling pitch, but varying somewhat with the speakers' ability to control their intonation. In most cases the realisations were a) shallow fall, b) rise, c) steep fall.

Spectral analysis was performed on the vocalic nuclei after excision from the word by using a waveform editor. One analysis process consisted of 12 pole LPC analysis (covariance method) on 20ms frames with 10ms overlap, followed by root-solving and formant-labelling using dynamic programming. Hard copies of the formant plots were made and the three phases of the diphthongs (onset steady-state, glide, offset steady-state) were identified where possible and measured manually. A second analysis was carried out using a filter-based spectrogram displayed on a graphics terminal. The durations and frequencies of the formants could be tracked with a cursor. The three phases were identified and measured a second time, independently of the first measurements. The two sets of measurements were compared, and where possible discrepancies were removed after checking the two displays.

Problems of several kinds occurred during measurement. In contrast to parameter definition for synthesis, discrete changes do not exist. The estimation of boundaries between steady-state and glide is therefore ultimately arbitrary. Clearly constant formant values are rare; therefore minimally rising or falling formants were taken as a steady-state if they preceded or followed portions of clearly increased rate of change. The point of gradient change in either F1 or F2 was taken to be the boundary. This meant, in fact, that the final steady-state portion had to include the clearly constant formant values are. Although the stimuli causing these comments could not be located definitively, after a first check the authors attributed this impression of /ai/ to those stimuli with no, or very short onset steady-states. This supports the assumption that timbre perception is the result of an integrating process over a considerable time window. Indications for such a process with speech and non-speech stimuli have been found in Brady et al. (1961), Mermelstein (1975), Huang (1986), D Benedetto (1987), and Stevens (1987).

Production

/haid/ and /haud/ utterances from 10 Southern English speakers (SF, SM) were selected from /haid/ words recorded as part of a normative study within Alvey project MMI 132 (Speech Technology Assessment, STA). Three tokens per diphthong and per person were produced, nominally with level, rising, and falling pitch, but varying somewhat with the speakers' ability to control their intonation. In most cases the realisations were a) shallow fall, b) rise, c) steep fall.

Spectral analysis was performed on the vocalic nuclei after excision from the word by using a waveform editor. One analysis process consisted of 12 pole LPC analysis (covariance method) on 20ms frames with 10ms overlap, followed by root-solving and formant-labelling using dynamic programming. Hard copies of the formant plots were made and the three phases of the diphthongs (onset steady-state, glide, offset steady-state) were identified where possible and measured manually. A second analysis was carried out using a filter-based spectrogram displayed on a graphics terminal. The durations and frequencies of the formants could be tracked with a cursor. The three phases were identified and measured a second time, independently of the first measurements. The two sets of measurements were compared, and where possible discrepancies were removed after checking the two displays.
Production results

Table 2 gives the measurements for the 10 speakers for all the tokens of /ai/ and /au/ for which all or some of the phases could be identified. A missing entry for one of the steady-states indicates that the diphthong started/ended with the glide portion. Missing entries for a whole token indicate the impossibility of identifying a structure relating to the stimulus structure used in the perception test, for reasons already given.

Table 2 gives the measurements for the 10 speakers for all the tokens of /ai/ and /au/ for which all or some of the phases could be identified. A missing entry for one of the steady-states indicates that the diphthong started/ended with the glide portion. Missing entries for a whole token indicate the impossibility of identifying a structure relating to the stimulus structure used in the perception test, for reasons already given.

Table 3 Normalised duration values for onset steady-state (SS1), glide (GL), and offset steady-state (SS2) for /ai/ and /au/.

It is apparent that the offset steady-state proportion of the diphthong is on average greater in production than is preferred in perception. However, it should be remembered that the perception stimuli were those of isolated diphthongs, whereas the production diphthongs were excised from a /h-d/ context. At least part of the offset steady-state is attributable to the transition into the final /d/. The values are still clearly different from the proportions found for German (SS1 25%, GL 42%, SS2 33%).
Conclusions

The perceptual data presented here show that there are definitely preferred structures for the diphthongs /ai/ and /au/ despite the recognisability of nearly all the stimuli as their functional equivalents. Furthermore, the preferred structure for English is clearly different from that found for German and Dutch. The production data showed a great variety of manifestations, and some were not of a comparable structure. Some, on the other hand, were very close in their temporal structure to the preferred stimuli in the perception experiment, though in general a greater proportion of the diphthong's duration came after the end of the glide. The average temporal structure remained, however, closer to the ideal English isolated diphthong than to the German diphthong structure.

Though these results represent a first step towards a specification of the diphthong, it also reveals points which need to be investigated in future research. What are the critical parameters of preference in natural diphthong production, and how do they interact and allow for articulatory compensation? For example, there was a great disparity between speakers in the dynamics of the amplitude envelope during the course of the diphthong, particularly with /au/, which possibly interacts with differences in the frequency dynamics in the formants. Also, with the present production data, there is no indication whether some tokens would be preferred to others. There is a strong need for acceptability comparison on natural diphthongs. The listeners' comments on the varying onset quality pose the question whether dialectal differences in speed of glide rather than systematically different onset configurations are responsible for the perceived onset differences.

Finally, it must be stressed that the emerging dynamic model of diphthong structure has wide-ranging implications for application as well as fundamental research. There is clearly a potential contribution from a dynamic diphthong model in several areas of speech technology, from improved synthesis to next-generation recognition systems.

References


Acknowledgments

Part of the research reported here was supported by Alvey grant MMI 132 and by the Netherlands Organization for Scientific Research (NWO), grant R 30-296.