Dynamics in Diphthong Perception

Ewa Jacewicz, Osamu Fujimura and Robert A. Fox
Ohio State University, Columbus, OH, USA
E-mail: jacewicz.1@osu.edu, fujimura.1@osu.edu, fox.2@osu.edu

ABSTRACT

The acoustic and perceptual characteristics of F2 movement in the diphthong [aɪ] in a Midwestern variety of American English are examined. The acoustic pattern of preconsonantal lengthening in 'bide' is investigated applying an ERB-based segmentation technique, which utilizes a constant Δ*F*. Listeners' response to dynamic information in the diphthongal onglide and offglide is tested. Data indicate that the diphthong [aɪ] in this variety of American English has a bipartite structure, consisting of a target and a gliding part. The terminal frequency values of the diphthong are not essential characteristics, as evident from the acoustic lengthening pattern in 'bide' and from perceptual response to the formant frequency change in this frequency range.

1. INTRODUCTION

The view of a diphthong as a vocalic syllable nucleus containing two target positions and associated transitional formant movement has dominated past research on diphthong dynamics [3]. The position of the second target has been questioned in subsequent work, which showed that its formant frequencies vary across changes in duration due to speaking rate [1]. The fact that the second target may never reach its frequency values and listeners perceive an articulatory movement [αε] as [αi] served as an argument for the undershoot hypothesis [4], which emphasized the directionality of F2 transition rather than an attainment of the target. A cross-linguistic and cross-dialectal investigation showed however that language-specific temporal pattern is an important cue in diphthong perception [5]. Accordingly, the German diphthong [aɪ] conforms to the tripartite temporal pattern (target /glide/target) whereas the English [aɪ] has a bipartite structure (target/(off)glide).

This study investigates the acoustic and perceptual characteristics of F2 movement in the diphthong [aɪ] in a Midwestern variety of American English. A well-known pattern of preconsonantal lengthening in English predicts that the stressed vocalic nucleus in 'bide' is longer than that in 'bite'. The expected differences are in both duration of the diphthong [aɪ] and the terminal frequency of its offglide, which influence the rate of formant change [3]. We first examine the rate of formant change due to acoustic differences in the two realizations of the diphthong to verify its bipartite structure as a vocalic nucleus. Next, the perception of formant change in a given frequency range is tested to investigate listeners' response to dynamic information in the diphthongal onglide and offglide.

2. ACOUSTIC STUDY

2.1 F2 SEGMENTATION.

According to [3], the rate of formant change is "the frequency range in cycles per second through which the formant moves in a given time interval" (p. 273). Before we can assess the acoustic differences between both realizations of [aɪ] in 'bite' and 'bide', we must address the problem of how to represent the continuously changing pattern most effectively. To compare the differences in duration between 'bite' and 'bide', we use a nonlinear equivalent rectangular bandwidth (ERB) scale [2] instead of a linear scale and apply a segmentation technique based on a constant Δ*F* found in individual diphthong production. The entire F2 trajectory is segmented in such a way that the Δ*F*, defining the boundaries between the consecutive segments, remains constant.

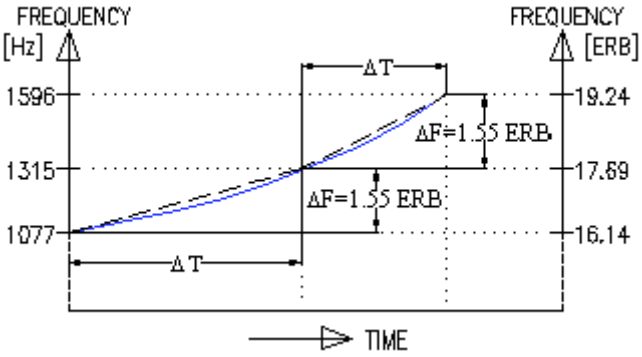

Figure 1. F2 segmentation technique applied to assess the acoustic differences between both realizations of [aɪ] in 'bite' and 'bide'.

Figure 1 illustrates the segmentation process. The first point of temporal measurement is located at the vowel onset, defined as the start of voicing following initial stop closure release. The second point is crucial to our analysis and is determined by a visual inspection of the relative positions of F1 and F2. The point at which F1 and F2 diverge from their parallel configuration is the second measurement point. The F2 frequency difference between the first and the second point in ERBu is our constant Δ*F*, which determines the duration (Δ*T*) of the first and each

ISBN 1-876346-48-5

subsequent segment. The subsequent measurement points are determined by applying the constant Δ*F* to the entire F2 trajectory.

2.2 METHODS AND RESULTS

Nine native speakers of American English spoken in the Midwestern state of Ohio (5 male and 4 female) recorded 6 repetitions of each 'bite' and 'bide' embedded in a carrier phrase. Recordings were digitized at 22.05-Hz sampling rate. The F2 segmentation technique described in 2.1 was used to measure the duration of each segment of [aɪ] in 'bite' in the production of each speaker. First, the constant Δ*F* for each instance of [aɪ] in 'bite' was obtained (segment 1). This Δ*F* was then applied to segment the entire F2 trajectory and to measure the duration (Δ*T*) of each consecutive segment. The last segmental boundary was located at the beginning of the stop closure. From six repetitions, an average Δ*F* for each speaker was next calculated and applied in a similar way to segment the diphthong [aɪ] in 'bide'.

Speaker/ Gender	Constant ΔF (ERBu)	F2 total ΔF (ERBu)	Duration (ms)
S1(M)	1.57 (0.07)	5.70 (0.23)	191 (6.15)
S2(M)	1.40 (0.09)	5.34 (0.16)	178 (16.02)
S3(M)	1.34 (0.13)	5.14 (0.22)	173 (9.63)
S4(M)	1.02 (0.11)	5.03 (0.22)	195 (6.13)
S5(M)	1.28 (0.13)	5.33 (0.30)	197 (22.96)
S6(F)	1.30 (0.05)	5.05 (0.39)	216 (16.41)
S7(F)	1.22 (0.05)	5.12 (0.23)	231 (9.95)
S8(F)	1.41 (0.04)	6.36 (0.34)	221 (17.77)
S9(F)	1.47 (0.14)	6.33 (0.65)	182 (18.00)

Table 1. Mean constant Δ*F*, total Δ*F*, and duration for [aɪ] in 'bite' for each speaker.

Speaker/ Gender	Constant ΔF (ERBu)	F2 total ΔF (ERBu)	Duration (ms)
S1(M)	1.57 (0.07)	5.43 (0.28)	315 (17.24)
S2(M)	1.40 (0.09)	5.05 (0.21)	272 (28.37)
S3(M)	1.34 (0.13)	4.98 (0.29)	249 (21.74)
S4(M)	1.02 (0.11)	3.51 (0.23)	251 (11.51)
S5(M)	1.28 (0.13)	3.81 (0.12)	306 (22.05)
S6(F)	1.30 (0.05)	4.24 (0.40)	281 (16.98)
S7(F)	1.22 (0.05)	4.20 (0.21)	374 (23.99)
S8(F)	1.41 (0.04)	5.49 (0.33)	453 (52.79)
S9(F)	1.47 (0.14)	5.39 (0.30)	258 (23.22)

Table 2. Mean constant Δ*F*, total Δ*F*, and duration for [aɪ] in 'bide' for each speaker.

Table 1 shows mean Δ*F* for each speaker, total duration of [aɪ] in 'bite', and a total Δ*F*, which is a difference between the lowest and the highest frequency value of the entire F2 trajectory. One standard deviation is given in parentheses. Similar measures for 'bide' are listed in Table 2.

As can be seen, the constant Δ*F* ranges from 1.02 - 1.57 ERBu for individual speakers. The total F2 Δ*F* for 'bite' reaches values from 5.03 - 6.36 ERBu. The total F2 Δ*F* for 'bide' has lower values, ranging from 3.51 – 5.49 ERBu. The drop in ERBu values for 'bide' corresponds to the increase in its total duration: mean duration for [aɪ] in 'bite' is 198 ms and in 'bide' is 307 ms.

Figure 2 presents a group plot of mean duration values for each segment (Δ*T*) of [aɪ] in 'bite' and 'bide' isolated by the segmentation technique described in 2.1. Each data point represents mean of 72 instances. There is a pattern of decreasing duration values with the second and third segment of F2 glide for 'bite', the first segment being the longest and the third segment the shortest. The final segment 5 has a slightly lower value for the diphthong in 'bide' which is clearly in contrast with the higher values for the other three segments. The increased duration of the first segment in 'bide' is of course the most striking result. Segment 4 reached exactly the same value of 34 ms for both realizations of the diphthong in the group data. It has to be mentioned that segments 4 and 5 belong to the same Δ*F*. The purpose of isolating segment 5 (i.e. the final steady-state of the offglide) is to show that its duration is even slightly lower in 'bide' than in 'bite'.

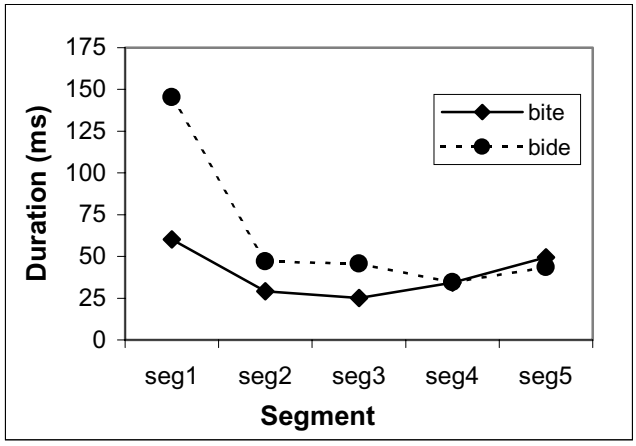

Figure 2. Group data from nine speakers showing duration of each segment (Δ*T*) of [aɪ] in 'bite' and 'bide'.

2.3 SUMMARY

The results have shown that there is a relationship between the overall duration of the diphthong and the rate of F2 change. The increased duration of [aɪ] in 'bide' affects primarily the duration of the first segment which corresponds to the first vowel target. The mean ratio for bide/bite duration is 2.3, which exceeds the ratios for the other segments. Longer durations can also be observed in the first part of the glide (segments 2 and 3 with the ratios

ISBN 1-876346-48-5

1.6 and 1.9, respectively), after which the "lengthening effect" in 'bide' disappears. The rate of change in 'bide' occurs over a reduced total F2 ΔF mostly due to lower terminal frequency values of the offglide. The unchanged duration of the offglide (segments 4 and 5) suggests a bipartite structure of [aɪ] in this variety of American English.

3. PERCEPTION STUDY

Perceptual characteristics of F2 movement can be studied in various ways. Our present approach focuses on the perception of formant change in a given frequency range which approximates the diphthongal onglide (i.e., initiation) and offglide (i.e., release). We examined F2 change in a fixed time interval of 70 ms to answer the question of whether listeners are equally sensitive to dynamic information in the onglide (variable terminal F2 values) and in the offglide (variable initial F2 values).

3.1 METHODS AND RESULTS

Three-formant synthetic stimuli were modeled after production of one male participant of the acoustic study. Two sets of tokens were generated. In the first set, the initial frequency value of F2 was fixed at 1200 Hz and the final frequency varied from 1287 to 1958 Hz in seven log steps. In the second set, the final F2 value was fixed at 2100 Hz and the initial frequency varied from 1287 to 1958 Hz in the same seven steps. The duration of each token was 70 ms and F0 was 120 Hz. Bandwidths were 80 Hz (B1 and B2), and 200 Hz (B3). Synthesis parameters are listed in Table 3.

Token	F2 initial (Hz)	F2 final (Hz)	F1 initial-final (Hz)	F3 initial-final (Hz)
fin1	1200	**1287**	660-680	2150
fin2		**1380**	660-670	2150
fin3		**1480**	660	2150
fin4		**1588**	660-620	2150-2130
fin5		**1702**	650-550	2150-2200
fin6		**1826**	660-530	2150-2250
fin7		**1958**	660-510	2150-2300
in1	**1287**	2100	660-400	2150-2500
in2	**1380**		660-400	2150-2500
in3	**1480**		650-400	2150-2500
in4	**1588**		590-400	2150-2500
in5	**1702**		550-400	2200-2500
in6	**1826**		530-400	2250-2500
in7	**1958**		500-400	2300-2500

Table 3. Synthesis parameters for two sets of stimuli. The final F2 values were varied in set 1 (fin#), and the initial final F2 values were varied in set 2 (in#).

The first set (fin#) approximated F2 change in the onglide to observe at which frequency value the listeners begin to perceive a diphthong and not a stationary [a]. The second set (in#) was an approximation of F2 change in the offglide and tested the perception of the transition from intermediate F2 frequencies to the final frequency value. F1 and F3 were appropriately modeled for each token based on the present acoustic measurements to create more natural conditions for the perception of F2 movement at a given F2 frequency.

The two sets of stimuli were presented randomly in one interval 3 AFC identification task with the choices [a], [aɪ], [ɛɪ], yielding 12 responses to each stimulus per listener. Four native speakers of the same variety of American English participated as listeners.

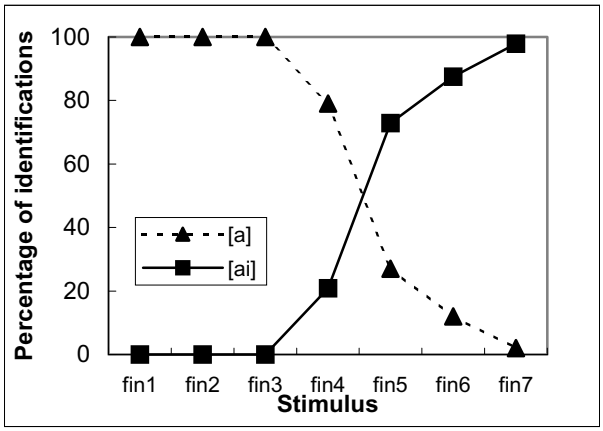

Figure 3. Group data from four listeners in response to the first stimulus set (fin#).

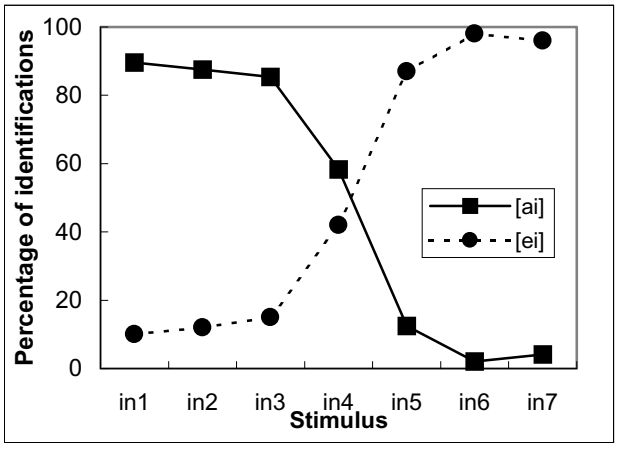

Figure 4. Group data from four listeners in response to the second stimulus set (in#).

Figures 3 and 4 display mean group responses to the first and second sets of stimuli, respectively. Each data point represents mean of 48 responses. The crossover point in Figure 3, corresponding to the change in percept from [a] to

ISBN 1-876346-48-5

[aɪ], falls between the stimuli fin4 and fin5, i.e., in F2 terminal frequency range from 1588 -1702 Hz. The crossover point in Figure 4 which signals the perceptual change from [aɪ] to [ɛɪ] is located at stimulus in4, whose beginning F2 frequency is 1588 Hz. The change in the percept from stationary [a] to diphthongal [aɪ] and from the qualitatively different [aɪ] to [ɛɪ] occurs at comparable F2 frequency values. This suggests the existence of a region in the spectrum where a more rapid change occurs in the perception of the formant movement.

It is noteworthy that the identification of [aɪ] reached 79% already at stimulus fin5, whose terminal F2 frequency value is 1702 Hz. Given that mean F2 terminal value in the production data from the present male participants is 2213 Hz for 'bite' and 1983 Hz for 'bide', we can conclude that the F2 information above certain value becomes less important in the identification of the diphthong [aɪ]. This observation supports earlier findings [1], [4], which emphasized the limited role of the terminal frequency values in acoustic and perceptual structure of diphthongs.

It is clear from Figure 4 that the diphthong [aɪ] sounds as [ɛɪ] with the initial F2 value of 1702 Hz at stimulus in5 and continues to be identified as [ɛɪ] through stimuli in6 and in7. This is exactly the region where the terminal frequency information for the identification of [aɪ] does not seem to supply essential information about the identity of the diphthong. Further investigation is necessary to provide a more conclusive statement about the way the dynamic information is coded and processed in the diphthongal offglide.

4. SUMMARY AND CONCLUSIONS

This study investigates the dynamics of the acoustic and perceptual characteristics of F2 in the diphthong [aɪ] in a Midwestern variety of American English. The acoustic pattern of preconsonantal lengthening in 'bide' in relation to 'bite' has been examined by applying a segmentation technique which compares the duration of consecutive segments throughout the entire F2 trajectories in both realizations of [aɪ]. The constant ΔF which defines the boundaries between the segments is determined for each speaker, addressing the differences between individual speakers' productions.

A comparison of segmental durations shows that the diphthong in 'bide' lengthens primarily during the first segment, i.e. the vowel target, and also during the first part of the onglide. The duration of the final segments remains the same for 'bite' and 'bide'. This pattern implies that the onglide contains the essential acoustic information which differentiates both realizations of [aɪ].

The perception of F2 change in a frequency range which approximates the diphthongal onglide indicates that listeners can identify the diphthong relatively early, considering the frequency change only. The ΔF at which the percept changes from [a] to [aɪ], i.e., the frequency difference between stimuli in1 and in5 in Figure 3 is 2.75 ERBu. This is clearly a reduction from ΔF values observed in individual productions of the diphthong in Tables 1 and 2 (compare F2 total ΔF values). Given the change in the percept from [aɪ] to [ɛɪ] at the same ΔF of 2.75 ERBu, the question arises as to the role of this auditory threshold in the processing of the dynamic information in F2 trajectory.

The present acoustic and perceptual data indicate that the diphthong [aɪ] in this variety of American English has a bipartite structure, consisting of a target and a gliding part. The information in the terminal frequency values is not an essential characteristic of the diphthong, as evident from the acoustic lengthening pattern in 'bide' and from perceptual response to the formant frequency change in this frequency range. This conclusion has been reached in light of the data on F2 movement only. Further investigation of F1 and F3 patterns is necessary, including the time dimension in perceptual testing as an essential variable.

ACKNOWLEDGMENTS

This work was partially supported by the NIDCD (R03 DC005560 to the first author).

REFERENCES

[1] T. Gay, "Effect of speaking rate on diphthong formant movements", *Journal of the Acoustical Society of America*, vol. 44, pp. 1570-1573, 1968.

[2] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, vol. 47, pp. 103-138, 1990.

[3] I. Lehiste and G. Peterson, "Transitions, glides, and diphthongs", *Journal of the Acoustical Society of America*, vol. 33, pp. 268-277, 1961.

[4] B. Lindblom and M. Studdert-Kennedy, "On the role of formant transitions in vowel recognition", *Journal of the Acoustical Society of America*, vol. 42, pp. 830-843, 1967.

[5] W. Peeters, *Diphthong Dynamics: A Cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German*, Kampen, NL: Mondiss, 1991.