Linguistic Research 35(1), 179-201
DOI: 10.17250/khisli.35.1.201803.007

Spectral patterns of the American English diphthong /aɪ/ as a function of coda voicing produced by native Korean speakers[*]

Eunjin Oh
(Ewha Womans University)

Oh, Eunjin. 2018. Spectral patterns of the American English diphthong /aɪ/ as a function of coda voicing produced by native Korean speakers. *Linguistic Research* 35(1), 179-201. This study aimed to investigate how native speakers of Korean who learned English as a second language realize spectral differences in the American English diphthong /aɪ/ as a function of coda voicing. Ten Korean learners of English and eight native speakers of American English participated in a production experiment. The monosyllabic words "bite" (/baɪt/) and "bide" (/baɪd/) were read along with filler words in isolation and in a carrier sentence. The native group demonstrated significantly smaller F1 and larger F2 before /t/ than /d/ both in the nucleus /a/ and in the offglide /ɪ/ (Moreton 2004). The learner group did not show statistically significant spectral changes in the nucleus and the offglide. Also, the native group significantly reduced the temporal distance between the nucleus and the offglide, and showed spectral peripheralization in the offglide before /t/ than /d/ (Pycha and Dahan 2016). However, the learner group did not show native-like reduction of the temporal distance between the nucleus and the offglide and spectral peripheralization in the offglide. Although the non-native speakers in this study exhibited some durational changes as a function of coda voicing, they did not learn the fine phonetic details regarding the gestural timing and spectral patterns in the diphthong. A considerable degree of individual variation in the learner group and speaking context effects were also found. It was interpreted that the Hyperarticulation hypothesis (Thomas 2000; Moreton 2004) and the Gestural Timing hypothesis (Pycha and Dahan 2016) could provide indices modelling the non-native phenomena found in this study as not attaining native-like phonetic values and gradual approach to the values concerning the spectral and durational aspects in /aɪ/ as a function of coda voicing. **(Ewha Womans University)**

Keywords American English, diphthong, /aɪ/, coda voicing, nucleus, offglide, F1, F2, duration, native Korean speaker, hyperarticulation, gestural timing

[*] I am grateful to anonymous reviewers for their helpful comments and suggestions. This work was supported by the CORE program funded by the Korea government (MOE).

1. Introduction

It is well known that coda voicing affects the duration of a preceding vowel in that the vowel duration becomes shorter before a voiceless consonant than before a voiced counterpart (e.g., House and Fairbanks 1953; Peterson and Lehiste 1960; Chen 1970). For example, Chen (1970) reported that native speakers of English, French, Russian, and Korean consistently produced shorter vowel duration before voiceless than before voiced consonants. Another aspect related to vowel duration is that when a vowel becomes shorter, in general, the centralization phenomenon is observed in the vowel space. According to Lindblom (1963), when a vowel becomes shorter, the time required to achieve articulatory targets of the vowel becomes insufficient. Thus, target values can be undershot with the vowel tending to become centralized. It has been reported that vowel shortening caused by omitting stress or faster speaking rates tends to cause lower F1 values (i.e., tongue body raising) in the English low vowels /æ/ or /ɑ/, centralizing the vowels (e.g., Gay 1978; Summers 1987).

From the two phenomena described above (i.e., vowel shortening before a voiceless consonant and vowel centralization with a shorter duration), it is predicted that before a voiceless consonant, a vowel becomes shorter and more centralized. However, as several studies have demonstrated, vowels show more peripheralization before a voiceless consonant than before a voiced consonant, which is puzzling (e.g., Wolf 1978; Summers 1987; Crowther and Mann 1992; Thomas 2000; Moreton 2004; Pycha and Dahan 2016). For example, although low monophthong vowels, such as /æ/ or /ɑ/ in English, become shorter before a voiceless consonant, speakers lower their jaw and produce even lower vowels (i.e., increasing F1 values) in these contexts (e.g., Wolf 1978; Summers 1987; Crowther and Mann 1992).

There have been studies concerning perception of English monophthong vowels as a function of coda voicing. For example, Wolf (1978) reported that stimuli with only the first half of the vowel in the /æC/ context (with vowel duration being equal) affected the voicing judgment of the coda consonant, indicating that spectral differences in that part of the vowel systematically vary depending on coda voicing. Wardrip-Fruin (1982) investigated the American English vowels /i, ɪ, ɛ, ɑ, u/ and reported that when 1/3 of the front part of vowels was deleted, the voicing judgment was not reduced. However, when 1/3 of the rear part of the vowels was deleted, the voicing judgment was reduced. This indicates that more perceptual cues for coda voicing are contained in the rear part of vowels. Summers (1988) elicited more voiceless judgments when the F1 steady state is higher in the American English vowel /æ/.

There have been several studies concerning the spectral characteristics of the American English diphthong /aɪ/ as a function of coda voicing (e.g., Thomas 2000; Moreton 2004; Pycha and Dahan 2016). The offglide F1 is lower and the offglide F2 is higher before a voiceless consonant compared to a voiced consonant. That is, the offglide /ɪ/ is peripheralized with the tongue body raising (lower F1) and fronting (higher F2) in the vowel space. Thomas (2000) investigated the production and perception of the American English /aɪ/ by central Ohioans and Mexican Americans from southern Texas. Results confirmed that both Ohio and Texas subjects showed the expected spectral differences (i.e., lower F1 and higher F2 before a voiceless consonant) in the diphthong offglide due to coda voicing, but the degree of spectral differences was significantly smaller for the Texas subjects. Regarding the perception, the subjects of both dialects used the offglide spectral differences as perceptual cues for coda voicing, eliciting more voiceless judgments with more peripheral offglides, although the Texas subjects used them to a lesser degree. Since there were differences between dialects, it was interpreted that the offglide spectral differences were not a consequence of articulatory constraints relating to the voicing of the following consonant (e.g., “pharyngeal expansion associated with voiced consonants”; Thomas 2000: 2), but resulting from phonetic grammar.

Moreton (2004) reported that more peripheral F1 and F2 patterns in the diphthong offglides before voiceless consonants were also found in the production of other American English diphthongs /eɪ, aʊ, ɔɪ/ as well as /aɪ/. A perception experiment was also performed in which synthetic stimuli manipulated offglide F1, offglide F2, and nucleus duration were used for the diphthongs /aɪ/ and /eɪ/. More voiceless responses were elicited for the stimuli with lower F1, higher F2, and shorter nuclei. It was concluded that the listeners used the spectral and durational cues for /eɪ/ as well as /aɪ/.

Two types of hypotheses have been suggested to model these seemingly contradictory phenomena (i.e., vowel peripheralization with a shorter duration in the voiceless context). First, Moreton (2004) explained the phenomena in terms of hyperarticulation in that the vowel peripheralization plays a role of reinforcing the otherwise weakened perceptual cues for the subsequent voiceless consonants supplementing the shortened duration. Thomas (2000) had a similar view regarding the phenomena, stating that “speakers compensate for the shorter duration of pre[-voice] /aɪ/, which ought to cause more truncation of the diphthong, by exaggerating the glide gestures” (p. 2). This “Pre-voiceless Hyperarticulation hypothesis” (Moreton 2004: 3) accounts for both the tongue body lowering of the low monophthongs /æ/ and /ɑ/ and the tongue body raising of the high offglide /ɪ/ in the diphthongs /aɪ/ and /eɪ/ before voiceless consonants. However, the nucleus /a/ as well as the offglide /ɪ/ in /aɪ/ is reported to be raised in the voiceless context, and this cannot be explained in terms of the Hyperarticulation hypothesis since the low nucleus /a/ is expected to be lowered in the voiceless context. Moreton (2004) further hypothesizes that this is due to coarticulation in that the voicing effect in the offglide (i.e., tongue body raising) continues into the nucleus, resulting in a raised /a/.

On the other hand, Pycha and Dahan (2016) explained the durational and spectral differences of the American English diphthong /aɪ/ as a function of coda voicing resulting from a temporal reorganization of articulatory gestures within the framework of Articulatory Phonology (e.g., Browman and Goldstein 1990). According to the Gestural Timing hypothesis, before a voiceless consonant, the gesture of the tongue body lowering for the nucleus /a/ and the gesture of the tongue body raising for the offglide /ɪ/ become closer, resulting in the gestures of the offglide /ɪ/ and the following consonant becoming more separated (Table 1). As a result, enough time is secured to realize the vowel target values in the offglide (i.e., peripheralization). On the contrary, before a voiced consonant, the gestures of /a/ and /ɪ/ become farer apart, resulting in the gestures of /ɪ/ and the following consonant becoming more overlapped (Table 1). As a result, the time necessary to realize the vowel target values would be insufficient in the offglide (i.e., centralization).

Table 1. Gestural timing in /aɪ/ as a function of coda voicing (Pycha and Dahan 2016).

Before	**Gestures of /a/ & /ɪ/**	**Gestures of /ɪ/ & C**	**Target values of /ɪ/**
Voiceless C	Closer	More separated	Peripheralized
Voiced C	Farer apart	More overlapped	Centralized

There have been studies concerning the monophthong vowels of English produced by non-native speakers. Crowther and Mann (1992) explored how native Japanese and native Mandarin Chinese speakers realize the duration of the vowel /ɑ/ and the F1 at the vowel offset in “pod” (/pɑd/) and “pot” (/pɑt/) of American English. A production test, an identification test using natural stimuli, and an identification test using synthesized stimuli to manipulate vowel duration and F1 offset frequency were performed. Results found that native English listeners used the vowel duration cue most sensitively, and native Chinese listeners used it least sensitively. As for the use of the offset F1 cue, native English listeners used the cue most sensitively, but differences among language groups were smaller compared to the use of the duration cue and there was no difference between the Japanese and Chinese groups. These results indicate that although the non-native listeners tended to perceive a voiceless consonant for stimuli with higher F1 and a voiced consonant for stimuli with lower F1, the F1 is a weaker cue than the vowel duration. It was interpreted that the Japanese and Chinese subjects who did not have native experience with word-final stops had difficulty using the vocalic cues for the coda voicing. In addition, the more sensitive use of the vowel duration cue by the Japanese group compared to the Chinese group resulted from the experience of Japanese speakers in using the vowel duration phonemically (i.e., long *vs.* short vowel contrasts) in their native language. Choi, Kim, and Cho (2016) investigated how native speakers of Korean realized the durational and spectral differences in the English monophthongs /ɛ/ and /æ/ according to coda voicing, and reported that they produced vowel duration differences but not spectral differences (F1 and F2) as a function of coda voicing.

This study aimed to investigate how native speakers of Korean who learned English as a second language realize the spectral differences in the American English diphthong /aɪ/ as a function of coda voicing. The Korean language has neither a voicing contrast in the word-final position nor a phonemic length contrast in vowels. Crowther and Mann (1992) and Choi, Kim, and Cho (2016) are studies on English monophthong vowels produced by non-native speakers. In this study, it will be investigated how Korean learners of English realize even finer phonetic details of the diphthong /aɪ/ with more complex spectral and durational patterns than monophthong vowels. It will also be explored whether data from non-native speakers can be modelled from the indices of the Hyperarticulation hypothesis (i.e., lowered F1 and raised F2 in the offglide before a voiceless consonant and the spectral coarticulation of the nucleus with the offglide; sections 3.1, 3.2, and 3.3) and the Gestural Timing hypothesis (i.e., smaller duration ratios of [nucleus/offglide] and larger spectral ratios of [F2/F1] in the offglide before a voiceless consonant; sections 2, 3.4, and 3.5) in a parallel way.

2. Experimental methods

Eighteen male speakers participated in this production study. Participants were divided into a native group and a learner group. The learner group was comprised of ten participants who were native speakers of Seoul Korean and learned English as a second language (age range 19~27; mean age 23.20). All Korean learners of English were born and raised in Seoul, were undergraduate or graduate students at Seoul National University at the time of recording, had no or almost no frequency of conversation with native English speakers in everyday life, and had no experience residing in an English-speaking country except two speakers who had resided in an English-speaking country less than one year. The native group included eight participants who were native speakers of American English (age range 18~22; mean age 20.25) and were undergraduate students at Stanford University at the time of recording.

Regarding test materials, the monosyllabic words “bite” (/baɪt/) and “bide” (/baɪd/) were used, containing the voiced bilabial stop in the initial position, the diphthong /aɪ/ in the medial position, and the voiced or voiceless alveolar stop in the final position. It is relatively easier to measure the acoustic values of the diphthong /aɪ/ compared to other diphthongs of American English in that movements in the tongue height and backness are clear from the low nucleus /a/ to the high front offglide /ɪ/ (e.g., Thomas 2000; Moreton 2004; Pycha and Dahan 2016). The test words were read in isolation and in the carrier sentence of ‘Say “___” to me.’ It will be examined whether the acoustic patterns in the diphthong were more native-like in either the isolation or the sentence context. Twelve filler words and sentences were also recorded together with the test materials.

Recordings were made using a MARANTZ PMD661 recorder and a Shure KSM32 microphone in sound-attenuated rooms in phonetics laboratories located in the two universities attended by the speakers of the two groups. The microphone was placed approximately 10 cm from the speakers' mouth. The speakers read the materials first for practice and then twice for recording. The test and filler materials were presented on a computer screen one at a time at an interval of two seconds. The recordings were digitized at a sampling rate of 44,100 Hz and stored as WAV files. Data was analyzed using the speech analysis program Praat (Boersma and Weenink 2016).

Measurements of /aɪ/ were made on waveforms and spectrograms, using a cursor in the following order: (1) At the maximum F1 value (where the tongue body is at the lowest for /a/), the F1 and F2 values were measured as spectral values of the nucleus. (2) Following the release of the word-initial stop to the maximum F1 were measured as duration of the nucleus. (3) From the maximum F1 to the beginning of the closure of the word-final stop were measured as duration of the offglide. (4) At the maximum F2 (where the tongue body is at the most front for /ɪ/), the F1 and F2 were measured as spectral values of the offglide (e.g., Moreton 2004; Pycha and Dahan 2016). The measurements of the formant values were made manually with reference to formant trajectories. The total duration of /aɪ/ was calculated by adding the duration of the nucleus and the offglide.

Pycha and Dahan (2016) used duration ratios of the nucleus to the offglide ([nucleus/offglide]) as an index of temporal distance between the two gestures of /a/ (tongue body lowering) and /ɪ/ (tongue body raising and fronting) (sections 1 and 3.4). As summarized in Table 2, smaller duration ratios indicate that the gestures of /a/ and /ɪ/ are more overlapped, shortening the duration of the nucleus (in the voiceless context), and larger duration ratios indicate that the two gestures are more separated (in the voiced context). Also, Pycha and Dahan (2016) used spectral ratios of F2 to F1 ([F2/F1]) in the offglide as an index of the degree of spectral peripheralization in /ɪ/ as a function of coda voicing (sections 1 and 3.5). As in Table 2, smaller spectral ratios indicate that F1 and F2 values are closer together with higher F1 and lower F2, centralizing the offglide in the vowel space (in the voiced context). Larger spectral ratios indicate that F1 and F2 values are farer apart with lower F1 and higher F2 (i.e., with tongue body raising and fronting for /ɪ/), peripheralizing the offglide (in the voiceless context).

Table 2. Indices of gestural timing and spectral peripheralization in /aɪ/ as a function of coda voicing (Pycha and Dahan 2016).

	Index	**Smaller ratio**	**Larger ratio**
Duration ratio [nucleus/offglide]	Temporal distance between two gestures of /a/ & /ɪ/	- More overlapped - Before /t/	- More separated - Before /d/
Spectral ratio in the offglide [F2/F1]	Spectral (F1, F2) peripheralization in /ɪ/	- More centralized - Before /d/	- More peripheralized - Before /t/

3. Experimental results

3.1 Spectral values of the diphthong /aɪ/ as a function of coda voicing

Figure 1 presents schematic representations of the F1 and F2 movements in the nucleus and the offglide of /aɪ/ in the contexts of /d/ (gray dotted lines) and /t/ (black solid lines) as a function of the speaker group (native *vs.* learner) and the speaking context (isolation *vs.* sentence). F1 (circles) and F2 (triangles) values are the group means with standard deviations. Concerning the native group, the mean spectral values were in the expected directions in that the F1 values were smaller and the F2 values were larger before /t/ than /d/ both in the nucleus and in the offglide. In isolation, the mean absolute differences in F1 between the /t/ and /d/ contexts were 91 Hz (= ∣663 Hz − 754 Hz∣) in the nucleus and 85 Hz (= ∣354 Hz − 439 Hz∣) in the offglide. The mean F2 differences were 146 Hz (= 1413 Hz − 1267 Hz) in the nucleus and 221 Hz (= 2127 Hz − 1906 Hz) in the offglide. In the sentence context, the mean F1 differences between the /t/ and /d/ contexts were 122 Hz (= ∣648 Hz − 770 Hz∣) in the nucleus and 80 Hz (= ∣344 Hz − 424 Hz∣) in the offglide. The mean F2 differences were 121 Hz (= 1417 Hz − 1296 Hz) in the nucleus and 266 Hz (= 2187 Hz − 1921 Hz) in the offglide.

Regarding the learner group, any noticeable changes in the spectral values were not found as a function of coda voicing, except the peripheralization of F2 before /t/ in the offglide in the isolation context. For the nucleus, the mean F1 and F2 values were slightly smaller before /t/ than /d/ both in isolation (mean absolute difference in F1 between the /t/ and /d/ contexts 14 Hz [= ∣802 Hz − 816 Hz∣]; mean difference in F2 5 Hz [= ∣1292 Hz − 1297 Hz∣]) and in sentence (mean difference in F1 11 Hz [= ∣797 Hz − 808 Hz∣]; mean difference in F2 4 Hz [= ∣1301 Hz − 1305 Hz∣]). Concerning the offglide, the mean F1 values were smaller and the mean F2 values were larger before /t/ than /d/ both in isolation (mean difference in F1 8 Hz [= ∣357 Hz − 365 Hz∣]; mean difference in F2 76 Hz [= 2188 Hz − 2112 Hz]) and in sentence (mean difference in F1 11 Hz [= ∣378 Hz − 389 Hz∣]; mean difference in F2 33 Hz [= 2150 Hz − 2117 Hz]).

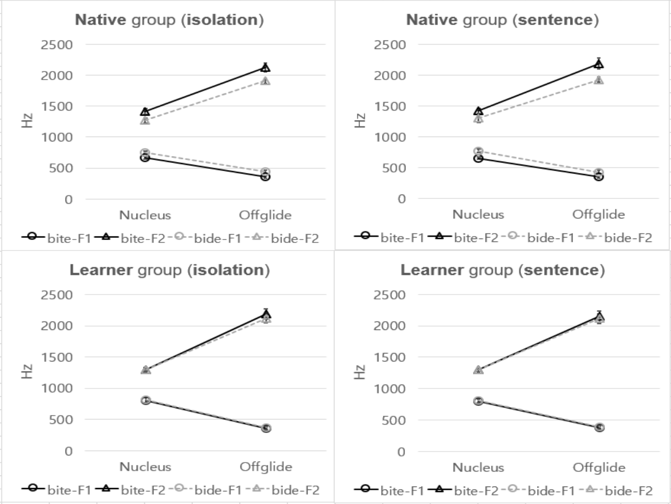

Figure 1. Schematic representations of F1 and F2 movements in /aɪ/ as a function of coda voicing, speaker group, and speaking context. F1 (circles) and F2 (triangles) values are group means with standard deviations.

Four repeated-measures ANOVAs for nucleus F1, nucleus F2, offglide F1, and offglide F2 were conducted to investigate the effects of the speaker group and the speaking context, using the mean spectral (F1 *or* F2) differences in the contexts of /t/ and /d/ for each speaker as the input data. Group served as the between-subjects factor, and speaking context (isolation *vs*. sentence) as the within-subjects factor. Group differences in the spectral differences between the /t/ and /d/ contexts were statistically significant all for the nucleus F1 ($F(1, 16) = 33.109$, $p < 0.001$), the nucleus F2 ($F(1, 16) = 19.809$, $p < 0.001$), the offglide F1 ($F(1, 16) = 25.732$, $p <$

0.001), and the offglide F2 (F(1, 16) = 27.261, p < 0.001). No significant interactions occurred except for the offglide F2 (F(1, 16) = 9.232, p = 0.008).

Regarding the Pre-voiceless Hyperarticulation hypothesis, the native group showed hyperarticulation in the offglide and coarticulation of the nucleus with spectral changes in the offglide. On the other hand, the learner group showed neither statistically significant hyperarticulation in the offglide nor coarticulation of the nucleus with the offglide.

3.2 Spectral and durational ratios of the voiceless to the voiced contexts

This section reports the spectral and durational ratios of the voiceless to the voiced context in the nucleus and the offglide of /aɪ/ as a function of the speaker group and the speaking context (cf., Thomas 2000). First, F1 and F2 ratios of the voiceless to the voiced context were calculated by (1) calculating the mean values of F1 and F2 for each speaker, (2) calculating F1 and F2 ratios of the voiceless to the voiced context from the mean values of each speaker, and (3) calculating group mean ratios and standard deviations from the individual mean ratios. It was expected that the F1 ratios would be smaller than 1, and the F2 ratios would be larger than 1, since F1 was lowered and F2 was raised before the voiceless compared to the voiced context by the native group (Figure 1). Figure 2 presents the group (native *vs.* learner) mean F1 and F2 ratios and standard deviations for the nucleus and the offglide. Table 3 reports 95% confidence intervals for the mean F1 and F2 ratios.

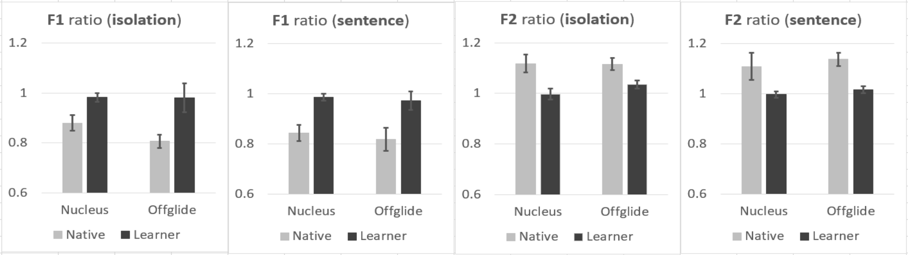

Figure 2. Mean F1 and F2 ratios of the voiceless to the voiced contexts and standard deviations for the nucleus and the offglide in /aɪ/ as a function of speaker group and speaking context.

Table 3. Mean F1 and F2 ratios, standard deviations, and 95% confidence intervals for the nucleus and the offglide in /aɪ/ as a function of speaker group and speaking context.

				Mean ratio	SD	95% CI	CI including 1.000?
Native	F1	Isolation	Nucleus	0.880	0.062	0.828 ~ 0.932	Excluding
			Offglide	0.807	0.053	0.763 ~ 0.851	Excluding
		Sentence	Nucleus	0.845	0.066	0.790 ~ 0.899	Excluding
			Offglide	0.819	0.091	0.743 ~ 0.895	Excluding
	F2	Isolation	Nucleus	1.118	0.071	1.059 ~ 1.177	Excluding
			Offglide	1.116	0.047	1.077 ~ 1.155	Excluding
		Sentence	Nucleus	1.109	0.107	1.019 ~ 1.198	Excluding
			Offglide	1.137	0.052	1.093 ~ 1.181	Excluding
Learner	F1	Isolation	Nucleus	0.982	0.036	0.957 ~ 1.008	Including
			Offglide	0.981	0.117	0.898 ~ 1.065	Including
		Sentence	Nucleus	0.987	0.026	0.968 ~ 1.006	Including
			Offglide	0.973	0.073	0.921 ~ 1.025	Including
	F2	Isolation	Nucleus	0.997	0.044	0.966 ~ 1.028	Including
			Offglide	1.035	0.030	1.014 ~ 1.057	Excluding
		Sentence	Nucleus	0.997	0.024	0.979 ~ 1.014	Including
			Offglide	1.016	0.028	0.995 ~ 1.036	Including

Concerning the F1 ratios, the native group showed mean ratios smaller than 1 for the nucleus and the offglide both in isolation and in sentence. The mean ratios were larger for the nucleus than the offglide, indicating that the F1 changes due to the coda voicing were larger for the offglide. On the other hand, the F1 ratios of the learner group were close to 1 for both the nucleus and the offglide. For the F2 ratios, the native group showed mean ratios larger than 1 for both the nucleus and the offglide. On the other hand, the F2 ratios of the learner group were closer to 1 for both the nucleus and the offglide. The spectral ratios of the learner group being close to 1 indicate that the learners did not produce native-like changes in the spectral values for the diphthong /aɪ/ as a function of coda voicing. As for the 95% confidence intervals for F1 and F2 as a function of the speaker group (Table 3), all the confidence intervals for the native group excluded 1, indicating that the spectral values were significantly different between the /t/ and /d/ contexts. All the confidence intervals for the learner group included 1 except the offglide F2 in isolation, indicating that the spectral values were not significantly different between the /t/ and /d/ contexts.

The duration ratios of the voiceless to the voiced contexts were calculated separately for the total, the nucleus, and the offglide duration by (1) calculating mean values of the duration for each speaker, (2) calculating the duration ratios of the voiceless to the voiced contexts from the mean values of each speaker, and (3) calculating group mean ratios and standard deviations from the individual mean ratios. Table 4 presents the mean duration ratios and standard deviations as a function of the speaker group and the speaking context.

Table 4. Mean duration ratios of the voiceless to the voiced contexts and standard deviations (parentheses) as a function of speaker group and speaking context.

Group	Isolation			Sentence		
	Total	Nucleus	Offglide	Total	Nucleus	Offglide
Native	0.688 (0.134)	0.577 (0.167)	0.754 (0.123)	0.726 (0.136)	0.498 (0.063)	0.875 (0.219)
Learner	0.790 (0.065)	0.728 (0.170)	0.847 (0.102)	0.862 (0.045)	0.844 (0.155)	0.890 (0.110)

Both the native and the learner groups showed ratios smaller than 1 in all cases of the total, nucleus, and offglide duration. The duration ratios of the native group were smaller than those of the learner group, indicating that native speakers produced larger differences in duration between the /d/ and /t/ contexts compared to learners. It is also notable that both the native and the learner groups produced smaller duration ratios for the nucleus than the offglide, indicating that the shortening before /t/ affected the nucleus to a larger extent than the offglide. However, the differences between the nucleus and the offglide ratios were larger for the native group (0.177 in isolation; 0.377 in sentence) than the learner group (0.119 in isolation; 0.046 in sentence). For the native group, the difference between the nucleus and the offglide ratios was larger in sentence (0.377) than in isolation (0.177).

The 95% confidence intervals for all the data in Table 4 excluded 1.000, except the offglide in sentence for the native group, indicating that the durational values produced by the native and the learner groups were significantly different between the /d/ and /t/ contexts. Three repeated-measures ANOVAs for the total, nucleus, and offglide duration were conducted to investigate the effects of speaker group and speaking context, using the mean durational ratios of the voiceless to the voiced contexts for each speaker as the input data. Group served as the between-subjects factor, and speaking context (isolation *vs.* sentence) as the within-subjects factor. Group differences were statistically significant for the total duration ($F(1, 16) = 8.473$, $p = 0.010$) and the nucleus duration ($F(1, 16) = 26.149$, $p < 0.001$), but not for the offglide duration ($F(1, 16) = 0.972$, $p = 0.339$). The interactions between speaking context and group were not significant for all the total, nucleus, and offglide duration.

3.3 Individual speaker patterns in spectral values due to coda voicing

Figure 3 presents the scatterplots of mean formant values (F1 *or* F2) in the contexts of /d/ (x-axis) *vs.* /t/ (y-axis) produced by individual native speakers (asterisks) and learners (circles). Results of the nucleus and the offglide were presented separately. Above is the isolation context, and below is the sentence context. Data located on the y=x line on the graphs indicate the formant values being identical in the contexts of /d/ and /t/ (Moreton 2004). Data points to the left or right of the y=x line indicate the formant values being larger or smaller, respectively, in the contexts of /t/ than /d/. Farer apart from the y=x line, data points represent larger spectral differences due to the voicing of the subsequent consonant (Moreton 2004).

All the data points of the native speakers were located to the right of the y=x line for F1, and to the left for F2, both in the nucleus and in the offglide. This indicates that all speakers in the native group produced smaller F1 and larger F2 in the contexts of /t/ than /d/ with one exception (i.e., in the nucleus F2 in sentence). On the other hand, the data points of the learners were inconsistent, demonstrating individual variations. Some data points were to the left or the right of the y=x line within each graph. That is, some learners produced the formant values in the native-like directions (i.e., smaller F1 and larger F2 values in the voiceless context), but others in the opposite directions (i.e., larger F1 and smaller F2 values in the voiceless context). It is also notable that the data points of the learners were in general closer to the y=x line compared to those of the native speakers, indicating that the learners produced spectral changes as a function of coda voicing to a lesser extent than the native speakers.

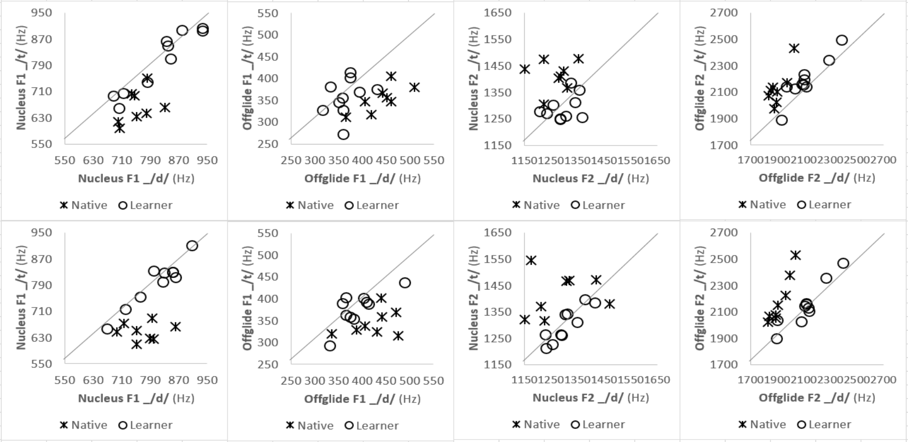

Figure 3. Scatterplots of mean formant (F1 or F2) values in the contexts of /d/ (x-axis) vs. /t/ (y-axis) by individual native speakers (asterisks) and learners (circles): The isolation (above) vs. sentence (below) context.

For the native group, the differences in the spectral values between the /d/ and /t/ contexts were statistically significant all for the nucleus F1 ([t = 5.123, p = 0.001] in isolation; [t = 6.073, p = 0.001] in sentence), the nucleus F2 ([t = −5.156, p = 0.001] in isolation; [t = −2.852, p = 0.025] in sentence), the offglide F1 ([t = 8.417, p < 0.001] in isolation; [t = 4.999, p = 0.002] in sentence), and the offglide F2 ([t = −6.616, p < 0.001] in isolation; [t = −6.681, p < 0.001] in sentence) by paired t-tests. For the learner group, the differences in the spectral values between the /d/ and /t/ contexts were not statistically significant for the nucleus F1 ([t = 1.579, p = 0.149] in isolation; [t = 1.577, p = 0.149] in sentence), the nucleus F2 ([t = 0.276, p = 0.788] in isolation; [t = 0.465, p = 0.653] in sentence), and the offglide F1 ([t = 0.598, p = 0.565] in isolation; [t = 1.286, p = 0.230] in sentence). For the offglide F2 of the learner group, the difference was significant in the isolation context (t = −3.879, p = 0.004), but not in the sentence context (t = −1.810, p = 0.104).

3.4 Duration ratios of nucleus to offglide as a function of coda voicing

This section examines the learners' duration ratios of the nucleus to the offglide in /aɪ/ as an index of gestural timing as a function of coda voicing (section 2; Pycha and Dahan 2016). Table 5 presents the mean duration of the nucleus and the offglide, and the mean duration ratios of [nucleus/offglide] as a function of coda voicing, speaker group, and speaking context. The duration ratios of [nucleus/offglide] represent values that are rounded off mean values calculated from individual means, and therefore, the ratio values presented in Table 4 can differ from the ratio values calculated from the rounded nucleus mean and the rounded offglide mean.

Table 5.. Mean duration (ms) of the nucleus and the offglide, and mean duration ratios of [nucleus/offglide] as a function of coda voicing, speaker group, and speaking context (standard deviations in parentheses).

Context		**Native** group			**Learner** group		
		Nucleus	Offglide	**Ratio**	Nucleus	Offglide	**Ratio**
Isolation	"bite"	57 (16)	133 (21)	**0.424** (0.098)	72 (19)	142 (21)	**0.518** (0.159)
	"bide"	100 (19)	177 (26)	**0.578** (0.145)	103 (32)	169 (29)	**0.633** (0.265)
Sentence	"bite"	41 (12)	115 (16)	**0.352** (0.074)	59 (21)	125 (20)	**0.484** (0.213)
	"bide"	83 (20)	137 (36)	**0.616** (0.140)	70 (22)	144 (34)	**0.536** (0.279)

The mean total duration (i.e., [nucleus + offglide]) was shorter before /t/ than /d/ for both the native group (189 ms *vs*. 278 ms in isolation; 155 ms *vs*. 220 ms in sentence) and the learner group (213 ms *vs*. 271 ms in isolation; 184 ms *vs*. 214 ms in sentence). The ratios of the total duration before /t/ to the total duration before /d/ were smaller for the native group (0.680 in isolation; 0.705 in sentence) than the learner group (0.786 in isolation; 0.860 in sentence), indicating that the learner group produced less durational differences due to coda voicing compared to the native group.

The duration ratios of [nucleus/offglide] were smaller before /t/ than /d/ for both the native group (0.424 *vs*. 0.578 in isolation; 0.352 *vs*. 0.616 in sentence) and the learner group (0.518 *vs*. 0.633 in isolation; 0.484 *vs*. 0.536 in sentence). The smaller ratio values before /t/ indicate the temporal distance being closer between the two gestures of the nucleus and the offglide (section 2; Pycha and Dahan, 2016).

However, the ratio differences between the /d/ and /t/ contexts were smaller for the learner group (0.115 [= 0.633 − 0.518] in isolation; 0.052 [= 0.536 − 0.484] in sentence) than the native group (0.154 [= 0.578 − 0.424] in isolation; 0.264 [= 0.616 − 0.352] in sentence). These results indicate that while both the native and learner groups reduced the temporal distance between the nucleus and the offglide in the /t/ context, the learner group exhibited this effect to a lesser extent. It is also notable that the ratio differences between the /d/ and the /t/ contexts were larger in sentence (0.264 [= 0.616 − 0.352]) than in isolation (0.154 [=0.578 − 0.424]) for the native group, but were smaller in sentence (0.052 = [0.536 − 0.484]) than in isolation (0.115 [= 0.633 − 0.518]) for the learner group. These results indicate that native speakers further adjusted the gestural timing between the nucleus and the offglide to a larger extent in the faster speaking rate in the sentence context of this study, while the learner group did not show this effect.

A repeated-measures ANOVA for the ratio differences between the /d/ and /t/ contexts was conducted to investigate the effects of speaker group and speaking context, using the mean ratio differences between the /d/ and /t/ contexts for each speaker as the input data. Group served as the between-subjects factor, and speaking context (isolation *vs*. sentence) as the within-subjects factor. The group difference was statistically significant ($F(1, 16) = 6.795$, $p = 0.019$), and the interaction between speaking context and group was not significant. By paired *t*-tests, the differences in the duration ratios of [nucleus/offglide] between the /d/ and /t/ contexts were statistically significant for the native group both in isolation ($t = -3.646$, $p = 0.008$) and in sentence ($t = -6.015$, $p = 0.001$). For the learner group, the differences were statistically significant neither in isolation ($t = -1.927$, $p = 0.086$) nor in sentence ($t = -1.047$, $p = 0.322$).

Figure 4 presents the relative duration ratios of the nucleus and the offglide which were normalized to sum to 1 for the native (left) and the learner (right) groups. For example, the relative duration ratios of the nucleus (mean duration 100 ms) and the offglide (mean duration 177 ms) in the context of /d/ produced in isolation by the native group were computed as 0.361 [= 100 / (100 + 177)] for the nucleus and 0.639 for the offglide [= 177 / (100 + 177)].

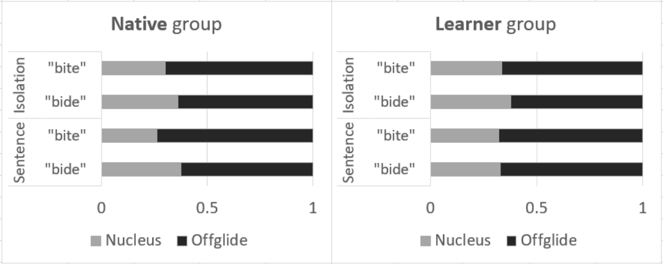

Figure 4. Relative duration ratios of the nucleus and the offglide normalized to sum to 1 as a function of coda voicing and speaker group.

For both the native and learner groups, the nucleus ratios were smaller before /t/ than before /d/. For the native group, the nucleus ratios were 0.300 before /t/ and 0.361 before /d/ in isolation, and 0.263 before /t/ and 0.377 before /d/ in sentence. For the learner group, the nucleus ratios were 0.336 before /t/ and 0.379 before /d/ in isolation, and 0.321 before /t/ and 0.327 before /d/ in sentence. The differences in the relative ratios between the /d/ and /t/ contexts were larger for the native group (0.061 [= 0.361 − 0.300] in isolation; 0.114 [= 0.377 − 0.263] in sentence) than the learner group (0.043 [= 0.379 − 0.336] in isolation; 0.006 [= 0.327 − 0.321] in sentence). These results indicate that the learners, as well as the native speakers, reorganized the gestural timing in the voiceless context by reducing the relative duration ratios for the nucleus, although the degree of the gestural reorganization was smaller for the learners. It is also notable that the native group showed a larger ratio difference between the /d/ and /t/ contexts in sentence (0.114) than in isolation (0.061), but the learner group showed a minimal difference between the /d/ and /t/ contexts in sentence (0.006) compared to the isolation context (0.042).

3.5 Spectral ratios of F2 to F1 in the offglide as a function of coda voicing

Table 6 presents the mean F1 and F2 values of the offglide in /aɪ/ and mean spectral ratios of [F2/F1] as a function of speaker group, speaking context, and coda voicing. The [F2/F1] ratios are rounded off mean values calculated from individual means, and therefore, the ratio values presented in Table 6 can differ from the ratio values calculated from the rounded F1 mean and the rounded F2 mean (section 3.4). According to Pycha and Dahan (2016), the [F2/F1] ratios index the degree of peripheralization in the spectral values, in that larger values infer a greater degree of spectral peripheralization (i.e., smaller F1 and larger F2) in the offglide (section 2).

Table 6. Mean F1 and F2 values (Hz) of the offglide in /aɪ/ and mean spectral ratios of [F2/F1] as a function of coda voicing, speaker group, and speaking context (standard deviations in parentheses).

Context		**Native** group			**Learner** group		
		F1	F2	**Ratio**	F1	F2	**Ratio**
Isolation	"bite"	354 (33)	2127 (139)	**6.069** (0.696)	357 (53)	2188 (156)	**6.262** (1.027)
	"bide"	439 (54)	1906 (82)	**4.419** (0.734)	365 (39)	2112 (140)	**5.853** (0.765)
Sentence	"bite"	344 (38)	2187 (177)	**6.459** (1.122)	378 (42)	2150 (170)	**5.768** (0.855)
	"bide"	424 (63)	1921 (82)	**4.685** (1.185)	389 (47)	2117 (152)	**5.526** (0.859)

The mean ratios of [F2/F1] in the offglide were larger before /t/ than /d/ for both the native group (6.069 *vs*. 4.419 in isolation; 6.459 *vs*. 4.685 in sentence) and the learner group (6.262 *vs*. 5.853 in isolation; 5.768 *vs*. 5.526 in sentence). However, the ratio differences between the /t/ and /d/ contexts were smaller for the learner group (0.409 [= 6.262 − 5.853] in isolation; 0.242 [= 5.768 − 5.526] in sentence) than the native group (1.650 [= 6.069 − 4.419] in isolation; 1.774 [= 6.459 − 4.685] in sentence). These results indicate that while both the native and learner groups showed peripheralization in the vowel space before /t/, the learner group showed the effect to a lesser extent. It is also notable that the native group showed larger [F2/F1] ratios in sentence than in isolation for both /t/ (6.459 *vs*. 6.069) and /d/ (4.685 *vs*. 4.419) contexts. On the other hand, the learner group showed smaller [F2/F1] ratios in sentence than in isolation for both /t/ (5.768 *vs*. 6.262) and /d/ (5.526 *vs*. 5.853) contexts. These results indicate that native speakers demonstrated further peripheralization in further vowel shortening that occurred in sentence, but the learners did not realize this fine phonetic detail.

A repeated-measures ANOVA for the ratio differences between the /d/ and /t/ contexts was conducted to investigate the effects of speaker group and speaking context, using the mean ratio differences between the /d/ and /t/ contexts for each speaker as the input data. Group served as the between-subjects factor, and speaking context (isolation *vs*. sentence) as the within-subjects factor. The group difference was statistically significant ($F(1, 16) = 42.238$, $p < 0.001$), and the interaction between speaking context and group was not significant. By paired *t*-tests, the differences in the spectral ratios of [F2/F1] in the offglide between the /d/ and /t/ contexts were statistically significant for the native group both in isolation ($t = 11.801$, $p < 0.001$) and in sentence ($t = 6.166$, $p < 0.001$). For the learner group, the difference was not statistically significant in isolation ($t = 1.346$, $p = 0.211$), and was significant in sentence ($t = 2.374$, $p = 0.042$).

Figure 5 presents the scatterplots of the duration ratios of [nucleus/offglide] and the spectral ratios of [F2/F1] in the offglide produced by the native (left) and learner (right) groups. According to Pycha and Dahan (2016), the larger duration ratios represent more separated gestural timing between /a/ and /ɪ/ (before /d/), and the larger spectral ratios represent greater peripheralization of F1 and F2 (before /t/) (sections 2, 3.4, and 3.5). Therefore, a negative correlation between these two indices is expected (Pycha and Dahan 2016). As seen in Figure 5, the native group showed a statistically significant negative correlation between the duration ratios and the spectral ratios (Pearson's $r \risingdotseq -0.383$, $p = 0.002$), while the learner group showed a slight positive correlation between these ratios with no significance (Pearson's $r \risingdotseq 0.204$, $p = 0.070$).

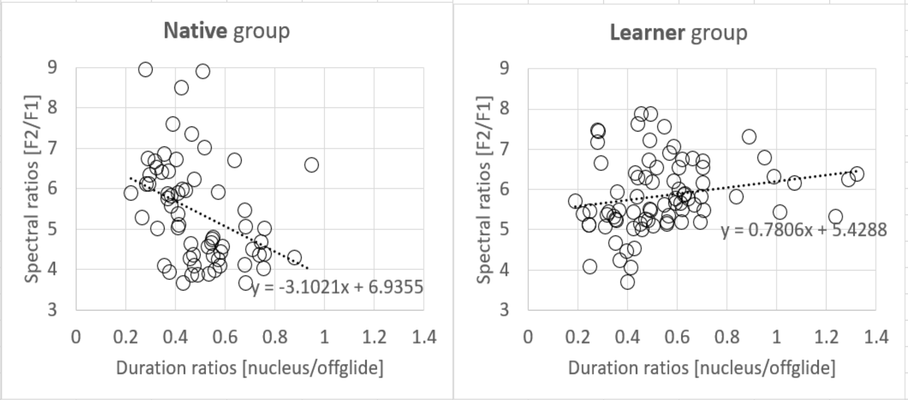

Figure 5. Scatterplots of duration ratios [nucleus/offglide] and spectral ratios [F2/F1] in the offglide as a function of speaker group.

4. Summary and discussion

The experimental results of this study can be summarized as follows: (1) The native group demonstrated significantly smaller F1 and larger F2 before /t/ than /d/ both in the nucleus and the offglide. The offglide data for the native group can be modelled in terms of the Pre-voiceless Hyperarticulation hypothesis and coarticulation of the nucleus with the spectral changes in the offglide (Thomas 2000; Moreton 2004). (2) The learner group showed neither statistically significant hyperarticulation in the offglide nor the coarticulation of the nucleus with the offglide. (3) The native group significantly reduced the temporal distance between the nucleus and the offglide, and showed greater spectral peripheralization in the offglide before /t/ than /d/, which can be the result of the Gestural Timing hypothesis (Pycha and Dahan 2016). (4) The learner group did not show native-like reduction of the temporal distance between the nucleus and the offglide, as well as spectral peripheralization in the offglide. (5) It can be interpreted that the Hyperarticulation hypothesis and the Gestural Timing hypothesis provide indices modelling the non-native phenomena found in this study as not attaining native-like phonetic parameters concerning the spectral and durational aspects in /aɪ/ as a function of coda voicing.

The learner group in this study demonstrated a considerable degree of individual variation, with some producing formant values in native-like directions (i.e., smaller F1 and larger F2 in the voiceless context), and others producing formant values in the opposite directions (i.e., larger F1 and smaller F2 in the voiceless context), as shown in Figure 3. In addition, the learners displayed individual differences in the duration ratios of [nucleus/offglide]. Two learners in the isolation context, and four learners in the sentence context, showed larger duration ratios of [nucleus/offglide] before /t/ than /d/, which contrasts with the native-like pattern. This learner phenomenon regarding individual variation is in contrast with findings with native speakers in this study and in Pycha and Dahan (2016), in which all subjects displayed larger duration ratios of [nucleus/offglide] before a voiced compared to a voiceless consonant without exception. The learner variation indicates that non-native speakers need to gradually learn the fine phonetic details in the spectral and durational cues as a function of coda voicing in the diphthong of English. It can be assumed that the two hypotheses discussed in this study provide appropriate indices modelling gradual approach to the native-like values in the relevant phonetic properties.

This study found speaking context effects. First, regarding the duration ratios of the voiceless to the voiced contexts, the native group produced a larger difference between the nucleus and the offglide ratios in sentence than in isolation, indicating that the faster speaking rate in the sentence context produced a shorter proportion of nucleus duration (section 3.2). Secondly, the differences between the /d/ and /t/ contexts in the duration ratios of [nucleus/offglide] were larger in sentence than in isolation for the native group, but were smaller in sentence than in isolation for the learner group (section 3.4). These results indicate that the native speakers further adjusted the gestural timing between the nucleus and the offglide to a larger extent for the faster speaking rates in the sentence context, while the learner group did not show this effect. Another speaking context effect was found for the spectral ratios of [F2/F1] in that the native group showed larger [F2/F1] ratios in sentence than in isolation, but the learner group showed smaller [F2/F1] ratios in sentence than in isolation for both /t/ and /d/ contexts (section 3.5). This was also interpreted as the native speakers demonstrated further peripheralization in the further vowel shortening that occurred in sentence. The learners in this study did not realize the native-like fine phonetic details regarding the speaking context effects on the spectral and durational aspects of the diphthong /aɪ/.

Pycha and Dahan (2016) reported in a perception study that incongruent duration ratios of [nucleus/offglide] in /aɪ/ (e.g., smaller duration ratios before /d/ or larger duration ratios before /t/) delayed listener identification of the coda voicing, even when other durational and spectral (i.e., the degree of peripheralization) factors were maintained. This was interpreted as listeners being sensitive to the different gestural timing due to coda voicing. Thomas (2000) reported that listeners use offglide spectral differences as perceptual cues when distinguishing between words. In Crowther and Mann (1992), not only native listeners, but also non-native listeners whose native language was Japanese or Mandarin Chinese, made more voiceless responses with higher F1 offsets in /ɑC/ context of American English. This result indicates that non-native listeners also use perceptual cues for vowels as a function of coda voicing despite using them to a lesser degree compared to native listeners. In this study, the F1 and F2 ratios of the voiceless to the voiced contexts produced by the learner group were close to 1, and it is possible that non-native tokens do not provide listeners with reliable spectral cues for the diphthong as a function of coda voicing. A perception study is necessary to investigate how native listeners judge non-native tokens with non-native spectral values. How non-native listeners process native perceptual cues regarding the gestural timing and the offglide spectral patterns for the diphthong /aɪ/ as a function of coda voicing also requires further examination.

References

Boersma, Paul and David Weenink. 2016. Praat: Doing phonetics by computer. Software retrieved from [http://www.fon.hum.uva.nl/praat/].

Browman, Catherine P. and Louis Goldstein. 1990. Gestural specification using dynamically defined articulatory structures. *Journal of Phonetics* 18: 299-320.

Chen, Matthew. 1970. Vowel length variation as a function of the voicing of consonant environment. *Phonetica* 22: 129-159.

Choi, Jiyoun, Sahyang Kim, and Taehong Cho. 2016. Phonetic encoding of coda voicing contrast under different focus conditions in L1 vs. L2 English. *Frontiers in Psychology* 7: 1-17.

Crowther, Court S. and Virginia Mann. 1992. Native-language factors affecting use of vocalic cues to final-consonant voicing in English. *Journal of the Acoustical Society of America* 92: 711-722.

Gay, Thomas. 1978. Physiological and acoustic correlates of perceived stress. *Language and Speech* 21: 347-353.

House, Arthur S. and Grant Fairbanks. 1953. The influence of consonantal environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America* 25: 105-113.

Lindblom, Björn. 1963. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35: 1773-1781.

Moreton, Elliott. 2004. Realization of the English postvocalic [voice] contrast in F_1 and F_2. *Journal of Phonetics* 32: 1-33.

Peterson, Gordon E. and Ilse Lehiste. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32: 693-703.

Pycha, Anne and Delphine Dahan. 2016. Differences in coda voicing trigger changes in gestural timing: A test case from the American English diphthong /aɪ/. *Journal of Phonetics* 56: 15-37.

Summers, W. Van. 1987. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analysis. *Journal of the Acoustical Society of America* 82: 847-863.

Summers, W. Van. 1988. F_1 structure provides information for final-consonant voicing. *Journal of the Acoustical Society of America* 84: 485-492.

Thomas, Erik R. 2000. Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics* 28: 1-25.

Wardrip-Fruin, Carolyn. 1982. On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *Journal of the Acoustical Society of America* 71: 187-195.

Wolf, Catherine G. 1978. Voicing cues in English final stops. *Journal of Phonetics* 6: 299-309.

Eunjin Oh
Department of English Language and Literature
Ewha Womans University
52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea
Email: ejoh@ewha.ac.kr

Received: 2018. 02. 08.
Revised: 2018. 03. 13.
Accepted: 2018. 03. 13.