http://dx.doi.org/10.17959/sppm.2016.22.1.147

Pattern recognition modeling of American English vowel identification by four different identification-proficiency levels of Korean listeners[*] [**]

Soonhyun Hong
(Inha University)

Soonhyun Hong. 2016. Pattern recognition modeling of American English vowel identification by four different identification-proficiency levels of Korean listeners. *Studies in Phonetics, Phonology, and Morphology* 22.1. 147-175. It has been recently reported through pattern recognition or synthesized vowel perception studies that American English listeners categorize vowels using dynamic spectral properties and duration rather than static spectral properties only (Hillenbrand et al. 1995, Hillenbrand 2013). However, Hong (2015) showed through pattern recognition modeling that Korean listeners used static spectral properties and duration when identifying English vowels. The present study extended Hong (2015) and built a logistic regression classification model to investigate which acoustic cues (static or dynamic spectral features, duration or F0) four different identification-proficiency levels of 133 Korean listeners may use to identify American English monophthongs in /hVd/ syllables. It turned out that the two upper-level groups of Korean listeners used dynamic spectral properties and duration just like American English listeners, whereas the other two lower-level groups used static spectral properties and duration. **(Inha University)**

Keywords: English vowels, vowel perception, pattern recognition modeling, logistic regression classification of vowels, different proficiency levels

1. Introduction

In the literature on the perception or production of American English (AE) vowels, it has been implicitly assumed that AE vowels can be properly characterized by steady-state Formant 1 (F1) and Formant 2 (F2) measurements in the acoustic vowel space

[*] This work was supported by INHA UNIVERSITY Research Grant.
[**] I would like to thank anonymous reviewers for their insightful comments and suggestions on this paper.

(Ladefoged and Johnson 2011).

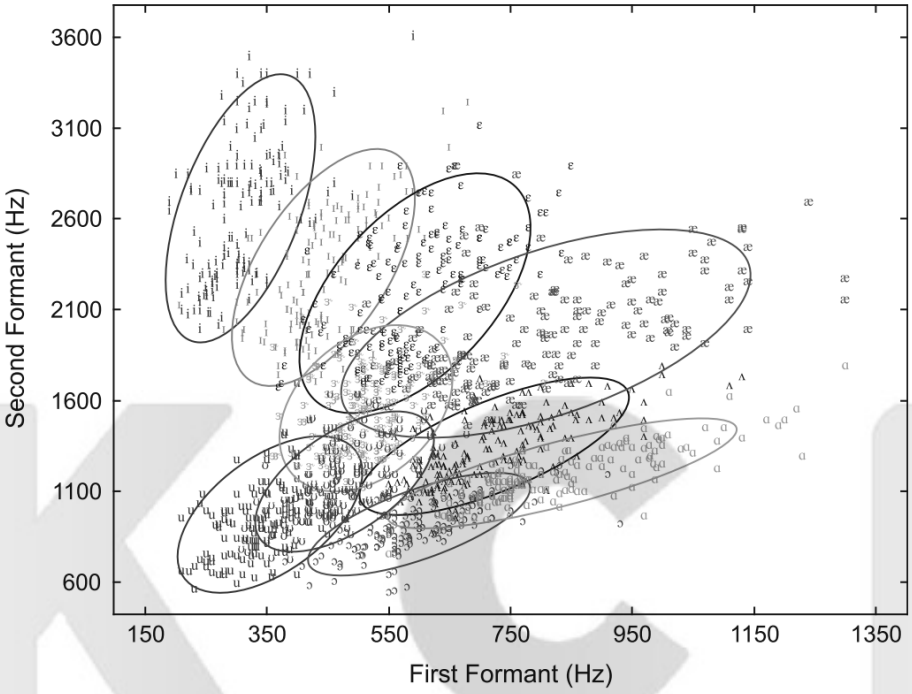

Figure 1. Plots of steady-state F1 and F2 measurements of English vowels from Peterson and Barney (1952) (from Hillenbrand 2013: 11)

However, Peterson and Barney (1952) recorded ten English vowels in /hVd/ syllables from 33 men, 2 women, and 15 children. Figure 1 shows the plots of the measurements of F1 and F2 sampled at steady-state, revealing extensive spread and overlap between vowel types in the acoustic space (Hillenbrand et al. 1995). Despite such extensive overlap and lengthy spread, all the vowel signals were identified as intended vowel types surprisingly over 94% of the time.

Hillenbrand et al. (1995) also recorded signals of twelve vowel types (/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, eɪ, oʊ, ɚ/) in /hVd/ syllables produced by 45 males, 48 females, and 46 children from the Upper Midwest in the U.S., the Michigan area. These vowel signals were identified as intended vowels more than 90% by another twenty AE listeners. Plots of ten vowel types (excluding /eɪ, oʊ/) with the measurements of F1 and F2

sampled at steady state, also showed similar lengthy spread and extensive overlap between vowel types as shown in Figure 2.

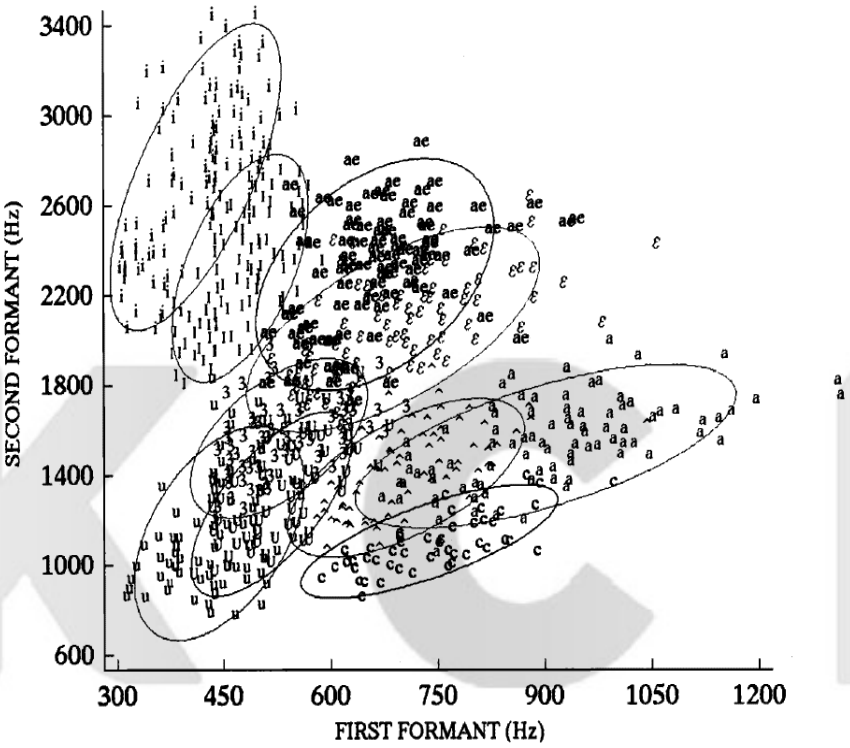

Figure 2. Values of F1 and F2 for 45 men, 48 women, and 46 children for 10 vowels with ellipses fit to the data ("ae"=/æ/, "a"=/ɑ/, "c"=/ɔ/, "^"=/ʌ/, "3"=/ɚ/) (from Hillenbrand et al. 1995: 3104)

The observation that AE listeners identified these vowel signals as intended vowels with over 90% accuracy, also indicates that AE listeners do not identify AE vowels only with static spectral measurements sampled at steady state. Possible major acoustic cues may be duration, steady-state F3 measurements and/or dynamic spectral cues like F1, F2 and F3 measurements sampled at various vowel duration points.

Hillenbrand et al. (1995) and Nearey and Assmann (1986) drew acoustic vowel charts with means of F1 and F2 measurements sampled at 20% and 80% of vowel

duration. They showed that AE monophthongs have sharp spectral property change just like diphthongs along the F1 and F2 axis. Namely, AE monophthongs pattern together with diphthongs in spectral change.

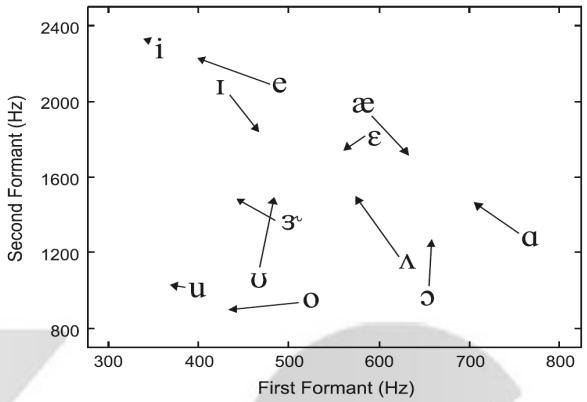

Figure 3. F1 and F2 measurements sampled at 20% and 80% of vowel duration for vowels in /hVd/ syllables spoken by 45 men from the Upper Midwest (Hillenbrand 2013: 13) (/e/=/eɪ/ and /o/=/oʊ/)

In Figure 3, /i/ and /u/ show almost no spectral movement in the acoustic space. However, all the other monophthongs show dynamic spectral change like diphthongal /eɪ/ and /oʊ/. Nearey and Assmann (1986) and Hillenbrand et al. (1995) argued that dynamic spectral properties should also be considered as major cues to explain AE listeners' vowel identification.

2. Previous studies

Hillenbrand et al. (1995) raised two questions as to AE listeners' vowel perception. First, do AE listeners pick up dynamic spectral properties when they perceive AE vowels? Second, if they do, how are dynamic spectral properties best characterized with the least number of spectral measurements along the vowel duration? The first question is about whether AE listeners use static or dynamic spectral properties for vowel perception. If AE listeners use dynamic spectral properties, the second question to be addressed is how many spectral measurements along the vowel duration are needed to properly characterize dynamic spectral properties of AE

vowels. Hillenbrand et al. (1995) trained a pattern recognition quadratic discriminant classifier model in a supervised learning mode to categorize 12 vowel types (/i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u, eɪ, oʊ, and ɝ/) in /hVd/ syllables, produced by 45 males, 48 females, and 46 children. The model was fitted to the AE vowel signals with various combinations of steady-state F0, duration, and static or dynamic measurements of F1-F3. The static spectral measurements refer to the measurements sampled once at steady state whereas the dynamic ones (1 samples), the measurements sampled twice at 20% and 80% of vowel duration (2 samples) or sampled three times at 20%, 50% and 80% (3 samples) in Table 1. Note in the table that F0 refers to the measurement at steady state and the results were based on a data set that did not include tokens with error rates of 15% or greater (11.5% of the tokens in the database).

Table 1. Quadratic discrimination results for the data showing the effect of including duration and spectral change information on classification accuracy ("NoDur"=vowel duration not included; "Dur"=vowel duration included)

Parameters	1 samples		2 samples		3 samples	
	NoDur	Dur	NoDur	Dur	NoDur	Dur
F1, F2	71.4	80	90.8	93.6	90.7	93.3
F1-F3	85.3	89.1	95.4	96.2	95.3	95.8
F0, F1, F2	82.3	86.3	95.5	96.3	94.8	96
F0, F1-F3	88.7	91.6	97.3	97.8	96.6	97.3

Partially from Hillenbrand et al. (1995: 3109)

As seen in Table 1, model classification improved substantially from 1 samples to 2 samples along with duration. However, almost no improvement was observed from 2 samples to 3 samples with duration. The best parameter in this model turned out to be a combination of measurements of steady-state F0, duration, 2 samples of F1, F2 and F3. Hillenbrand et al (1995) concluded that dynamic spectral properties can be best characterized by 2 samples of formant frequencies (with duration and steady-state F0). They argued that the pattern recognition model with 2 samples of formant measurements could model AE listeners' identification best. They concluded that the dynamic spectral properties of AE vowels can be best characterized with spectral measurements sampled twice at 20% and 80% of vowel duration and AE listeners use dynamic spectral properties, duration, and steady-state F0 for AE vowel identification.

Other pattern recognition studies (Nearey and Assmann 1986, Zahorian and Jagharghi 1993, Hillenbrand et al. 1995, Hillenbrand and Nearey 1999, Morrison 2013) also demonstrated that AE vowels could be classified more accurately and robustly when dynamic spectral measurements (i.e., two samples of spectral information) were used as spectral cues than when static spectral measurements were used. These results also indicated that "vowel inherent spectral change" (VISC[1]), namely, dynamic spectral features, should be considered to explain dynamic spectral movements of AE monophthong signals. These pattern recognition modeling studies with spectral change (and duration) agreed better with AE listeners' vowel identification (Hillenbrand 2013).

Studies using synthesis methods also reported that AE vowels can be more properly characterized by spectral trajectories in the acoustic space. Jenkins et al. (1983), Nearey (1989) and Nearey and Assmann (1986) observed the highest identification rates by AE listeners for the "silent-center" vowel signals in which the middle 50-60% was replaced with silence in /bVb/ signals than for the variable-center signals in which variable length of both of the beginning and end of the signals was excised. In addition, Hillenbrand and Gayvert (1993) found that steady-state vowels synthesized from the measurements in Peterson and Barney (1952), were poorly identified by AE listeners. These studies also suggest that dynamic spectral features are crucial in AE vowel identification.

The importance of the duration cue is also found in the literature. Conducting an AE listeners' identification experiment with synthesized AE vowels, Hillenbrand et al. (1995) showed that a pattern recognition model excelled with duration in addition to dynamic spectral change. Hillenbrand et al. (2000) conducted pattern-recognition modeling of AE listeners' identification of the vowels of variable length which were manipulated through sinusoidal synthesis. They modeled the identity shifting of those vowels as a function of manipulated duration with a quadratic discriminant classifier. The model was fitted to the vowel signals with the measurements of duration, steady-state F0, and F1-F3 sampled at 20% and 80% of vowel duration and then was tested

[1] "[VISC] refers to the changes in spectral properties over the time course of a vowel which are characteristic of vowel-phoneme identity. It refers not only to the widely-recognized spectral changes found in diphthongs and triphthongs, but also to the less-well-recognized spectral changes which are characteristic of vowel-phonemes which have traditionally been called monophthongs in some dialects of some languages, particularly in North American English" (Assmann and Morrison 2013: 1).

on the same measurements but with modified duration for the synthesized signals. The optimal model showed vowel classification shifts similar to AE listeners', though it was more sensitive to duration than AE listeners.

Table 2. Frequency of vowel identity change resulting from sinusoidal-synthesized vowel shortening or lengthening

	Vowel shift		% of vowel shifts by AE listeners	% of vowel shifts by model
	From	To		
Vowel shortening effects	/ɔ/	/ɑ/ or /ʌ/	43	54.2
	/æ/	/ɛ/	20.7	25
	/ɑ/	/ʌ/	9.4	8
Vowel lengthening effects	/ʌ/	/ɑ/ or /ɔ/	36	60
	/ɛ/	/æ/	18.9	33

Table merged from Table II and III in Hillenbrand et al. (2000: 3019-3020)

As shown in Table 2, AE listeners identified 43% of shortened /ɔ/ signals as /ɑ/ or /ʌ/ and 20.7% of shortened /æ/ signals as /ɛ/ whereas 36% of lengthened /ʌ/ as /ɑ/ or /ɔ/ and 18.9% of lengthened /ɛ/ as /æ/. This suggests that duration also plays an important role in AE vowel identification.

There have been some studies on Korean listeners' perception of AE vowels. Yun (2005) conducted an experiment with synthesized English /i-ɪ/ and /ɪ-ɛ/ continua and reported that Korean listeners (hereafter, K listeners) relied both on vowel duration and (static) spectral cues to identify the /i-ɪ/ continuum (Ingram and Park 1997; Flege et al. 1997 for different views[2]) while AE listeners relied on (static) spectral cues only. However, K listeners used the (static) spectral cues for /ɛ/-like sounds in the /ɪ-ɛ/ continuum like AE listeners. Though his study was restricted to limited AE vowel types and to their static spectral properties, it still indicated that K listeners use different acoustic cues from AE listeners for AE vowel identification due to the interference from their native Korean vowels[3] (Iverson et al. 2003).

---

[2] Ingram and Park (1997) and Flege et al. (1997) independently argued that K listeners relied only on vowel duration.

[3] Long/short vowels are distinctive in word-initial syllables in Korean (Korean Ministry of Education 1988: 102).

Assuming that such pattern recognition modeling studies of AE listeners' vowel identification are on the right track to uncover AE listeners' use of (dynamic) spectral properties, duration, and steady-state F0 as major cues, Hong (2015) examined whether K listeners' difficulty differentiating among AE vowels may result from the hypothesis that they might not use major acoustic cues the way AE listeners do. He administered a forced-choice vowel identification task to 57 K listeners, using AE vowel signals partially selected from Hillenbrand et al. (1995). Then a 10-fold cross-validation logistic regression classifier model was fitted to the measurements of the AE vowel signals based on K listeners' identification with the same types of parameters as in Hillenbrand et al. (1995). The fitting procedure was repeated ten times for each parameter. Note that Table 3 and Figure 4 are from Hong (2015). Note also that 57 K listeners' overall correct identification rate was 61.49% (SD=11.65, N=57).

Table 3. Accuracy means for parameters with or without duration and with a combination of steady-state or dynamic spectral cues for K listeners' identification of vowels after 10 times of trials (partially modified from Table 6 in Hong 2015: 229)

Parameters	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0 F11F22
Mean	62.73	59.40	62.89	62.74	60.49
SD	.12	.09	.04	.11	.09
N	10	10	10	10	10

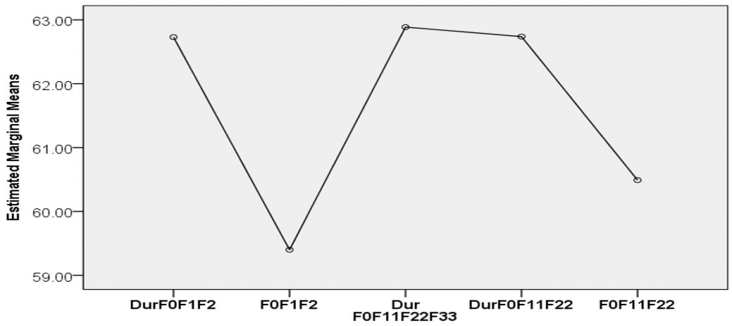

Figure 4. Comparison between estimated marginal means of parameters to model K listeners' identification of vowels (Hong 2015: 229)

Figure 4 shows that DurF0F1F2, DurF0F11F22F33 and DurF0F11F22 showed significantly more classification improvement than F0F1F2 and F0F11F22[4]. Hong (2015) concluded that K listeners in general used duration as a major cue for AE vowel identification like AE listeners. However, he observed that K listeners appealed to static spectral information unlike AE listeners. Namely, no model performance difference was found among the three best performing parameters (DurF0F1F2, DurF0F11F22F33 and DurF0F11F22), which included either static or dynamic spectral information inside the parameter. The parameter with static spectral information and the ones with dynamic spectral information resulted in equivalent model fitting improvement. This means that dynamic spectral information was redundant in modeling K listeners' AE vowel identification. Therefore, DurF0F1F2 which uses the least number of acoustic cue measurements, excels by Occam's Razor. For this reason, Hong (2015) concluded that K listeners used static spectral properties (but not dynamic properties) along with duration when they identified AE vowels.

Hong (2015) administered a forced-choice AE vowel identification test to 57 K listeners as one single group and then compared their use of acoustic cues with five AE listeners'. However, K listeners are different in their English proficiency, more specifically in their AE vowel perception abilities. This suggests that K listeners of different AE vowel identification-proficiency levels resulting from an AE vowel identification test, may use different combinations of acoustic cues. Unlike Hong (2015), the present study divided 133 K listeners into four different groups (Advanced, Upper-intermediate, Lower-intermediate, and Beginner's) based on the performance in the same forced-choice AE vowel identification task in Hong (2015), to see if different groups use a parameter different from or the same as AE listeners' group does. We assumed, as in Hong (2015), that K and AE listeners' identification can be properly modeled with a combination of cues like steady-state F0, duration and static or dynamic F1-F3. The present study hypothesized that if AE listeners use dynamic spectral cues for AE vowel identification, the four different levels of K listeners may pick up different combinations of acoustic cues for AE vowel identification. More specifically, K listeners of higher levels of identification may pick up dynamic spectral cues like AE listeners while lower levels of K listeners

[4] "F1" refers to F1 measurements at steady state whereas "F11" to F1 measurements sampled at 20% and 80% of vowel duration. For more information, refer to Table 6.

static spectral cues due to interference from their native Korean vowels. The present study also wondered whether different levels of K listeners use duration differently for AE vowel identification. Through the pattern recognition modeling of K listeners' identification results, we would like to see what combinations of acoustic cues these four different groups of K listeners use, compared to the acoustic cues five AE listeners use.

3. Experiment

3.1 Subjects

The K listeners' group consisted of 133 college-level K listeners (103 females and 30 males) whose age varied from 20 to 27 (mean=21.2). They had been learning English for at least 6 years. The other subject group consisted of 5 AE listeners (2 males and 3 females) whose age varied from 19 to 23. All of them were born near Washington D.C. or Rockville, Maryland[5]. All the subjects had no reported history of speech or hearing problems.

3.2 Stimuli

3.2.1 The database in Hillenbrand et al. (1995) which the AE vowel stimuli were selected from

The same set of /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/ signals in /hVd/ syllables which was used in Hong (2015), was also used in the present experiment but with different Korean listener subjects. The English monophthong signals were carefully chosen among the signals produced by male or female talkers in the database in Hillenbrand et al. (1995),

The database in Hillenbrand et al. (1995) which the vowel stimuli were selected from, have audio recordings of 12 vowels in /hVd/ syllables read by AE talkers: /i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ, eɪ, oʊ/ in "heed," "hid," "head," "had," "hod," "hawed," "hood," "who'd," "hud," "heard," "hayed," and "hoed," respectively. These recordings were

[5] The five AE listeners were the same subjects as in Hong (2015).

from 139 talkers (45 males, 48 females and 46 children), all of whom were raised in the Michigan State. The signals were accompanied with F1-F3 frequencies of the steady-state vowel tracks, F1-F3 frequencies sampled at 10 to 80% of vowel duration by 10% increment, and vowel duration and steady-state F0 frequencies. The signals also came with correct identification rates by another group of twenty phonetically trained AE listeners. 89% of the vowel signals were identified as intended vowels at 90% or above. Unfortunately, however, as pointed out in Hong (2014, 2015), the authors failed to take seriously the fact that some of the talkers of General American English do not contrast between /ɑ/ and /ɔ/. And this phenomenon is observed with the AE talkers in the database. In the database, intended /ɔ/ signals were heard as /ɔ/ 82.0%, as /ɑ/ 13.8% whereas intended /ɑ/ was heard as /ɑ/ 92.3%. Hillenbrand et al. (1995) reported that /ɔ/ signals consistently heard as an unintended vowel, were "misproduced" vowels, a typical characteristic (neutralization of /ɔ/ to /ɑ/) in General American English. They admitted that nearly 14% of the attempts at /ɔ/ were production "errors."

In the database, all the vowel types except /ɔ/ were identified as intended vowels at 90% or above: From the highest 99.6% for /i/ to the lowest 90.8% for /ʌ/, compared to 82.0 for /ɔ/. For this reason, Hong (2015) assumed that if /ɔ/ signals are produced correctly, they should be heard as /ɔ/ by more than 90% of the listeners, as all the other vowels were heard as intended vowels by more than 90% of the listeners. In order to pick up convincing /ɔ/ signals, correct identification rates of all the vowel types in the database were compared with one another. Hong (2015) used in the identification test the vowel signals which were produced by talkers whose /ɔ/ signals were agreed on the intended /ɔ/ by 90% or above of the 20 listeners (see Hillenbrand et al. (1995) for the misproduction of /ɔ/ and Hong (2014, 2015)).

3.2.2 Stimuli in the test

Target vowels were nine English monophthong vowels (/i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/) in /hVd/ syllables. The total number of vowel signals for the test was 180 (= (8 males + 12 females) * 9 vowels), and AE listeners identified 97.97% (SD[6]=3.20) of the stimuli vowels as intended vowels according to the database (Hong 2015).

---

[6] SD refers to Standard Deviation.

Table 4. Means and SDs of accuracy rates for nine vowel types of stimuli identification by 20 AE listeners

Vowels	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	Total
Mean	99.50	99.00	98.50	97.75	96.50	97.00	96.50	98.00	99.00	97.97
SD	1.54	2.62	2.86	3.02	3.28	3.40	5.16	2.51	2.05	3.20
N	20	20	20	20	20	20	20	20	20	180

Hong (2015: 218)

3.3 Procedures

The forced-choice /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/ identification protocol in Hong (2015) was used in the present experiment, based on Alvin 2.0 (Hillenbrand and Gayvert 2005). In the protocol, the 180 signals were randomized without repetition. After listening to each stimulus presentation, the subject was forced to click on a vowel icon with a "hVd" word and a vowel phonetic symbol inside on a computer screen. The interval between the click and the next stimulus presentation was set to 500ms. The subject could get back to the previous presentation to make a readjustment click if s/he made an error click. Furthermore, listeners were allowed to listen to each signal up to three times. It took about 20 minutes to complete the task, which was administered at a phonetics lab.

4. Results

All the five AE listeners identified 100% of all the vowel signals as intended vowels. They had no problem differentiating between /ɑ/ and /ɔ/. However, 133 K listeners varied drastically in vowel identification rates and they were divided into four groups of almost equal number of members (Advanced, Upper-intermediate, Lower-intermediate, and Beginner's groups), based on their identification performance. Four Korean listener groups' mean correct identification rates and SD figures are shown in Table 5.

Table 5. Means and SDs of correct identification rates for the four different K listeners' groups

Groups	Mean (%)	SD	N
Advanced	92.42	3.28	33
Upper-intermediate	75.70	7.57	34
Lower-intermediate	59.56	2.73	33
Beginner's	49.60	4.98	33
Total	69.37	17.02	133

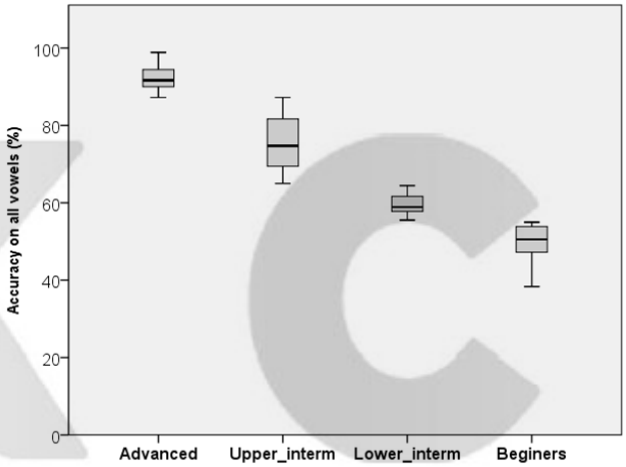

Figure 5. Boxplots of the correct AE vowel identification rates by the four different K listeners' groups

The boxplots in Figure 5 show that 133 K listeners were divided into four groups based on the correct identification rates from the identification test: $F(3, 129)= 460.53$, $p<0.01$, $r^2=.915$). Posthoc Bonferroni showed that pairwise comparisons between the groups were all significant ($p<0.001$).

5. Discussion

5.1 The acoustic characteristics of the tested vowel signals

When all the tested vowel signals were plotted with the measurements sampled at steady-state measurements of F1 and F2 on the acoustic space, a lot of extensive overlapped areas between vowel types were observed and plots of each of the vowel types showed wide spread in the acoustic space, as shown in Figure 6. The fact that all five AE listeners successfully identified the stimuli vowels as intended vowels, strongly suggested that steady-state F1 and F2 measurements are not enough to characterize these vowels.

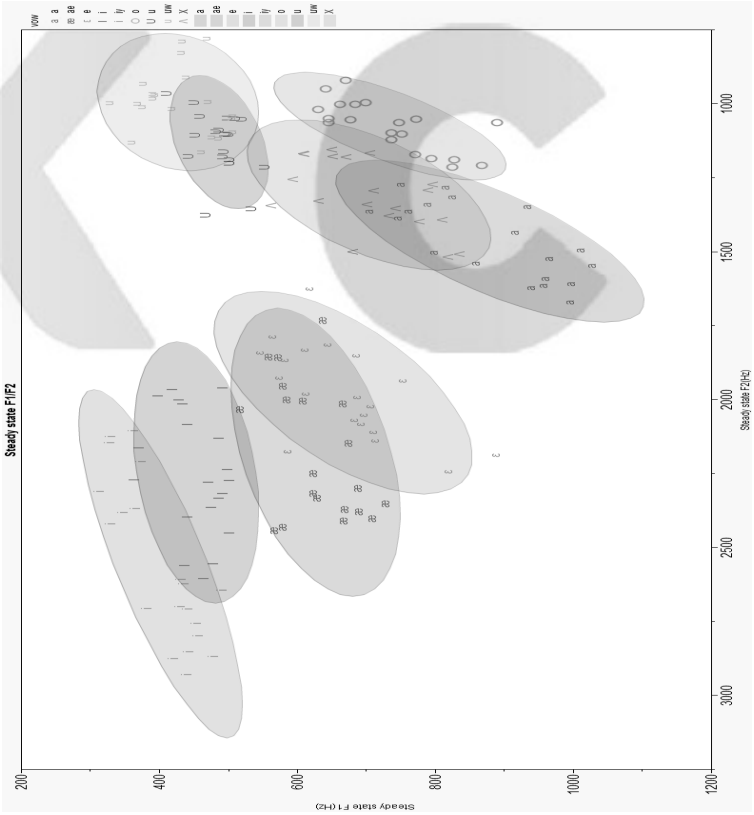

Figure 6. Plots of F1 and F2 measurements of all the tested vowel signals sampled at steady state (Hong 2015: 223)

When means of F1 and F2 measurements of the tested vowel signals of males' and females' sampled at steady state were plotted, the nine monophthong vowel types were unfortunately not so equally spaced as implicitly assumed in the literature on English vowel production (Ladefoged and Johnson 2011), as shown in Figure 7. This suggests that static spectral features along are not enough to explain how AE listeners' perceive vowels[7].

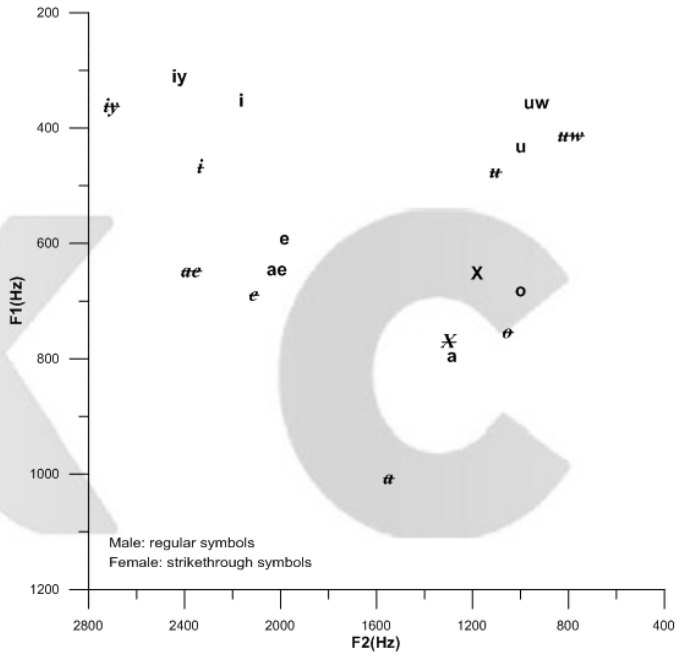

Figure 7. Plots of mean F1 and F2 measurements of the tested vowel signals across vowel types sampled at steady state

On the other hand, when means of F1 and F2 measurements of the tested vowel signals of males' and females' sampled at 20% and 80% of vowel duration, were plotted (all the data and Figure 8 from Hong 2015), all monophthong types except for /i/ and /u/ showed dynamic spectral properties which are observed in diphthongs, exactly as were demonstrated in Hillenbrand et al. (1995) and Hillenbrand (2013).

---

[7] Due to incompatibility with the statistical package, /iy, i, e, æ, a, o, x, u, uw/ in the figures in the present study refer to /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/, respectively.

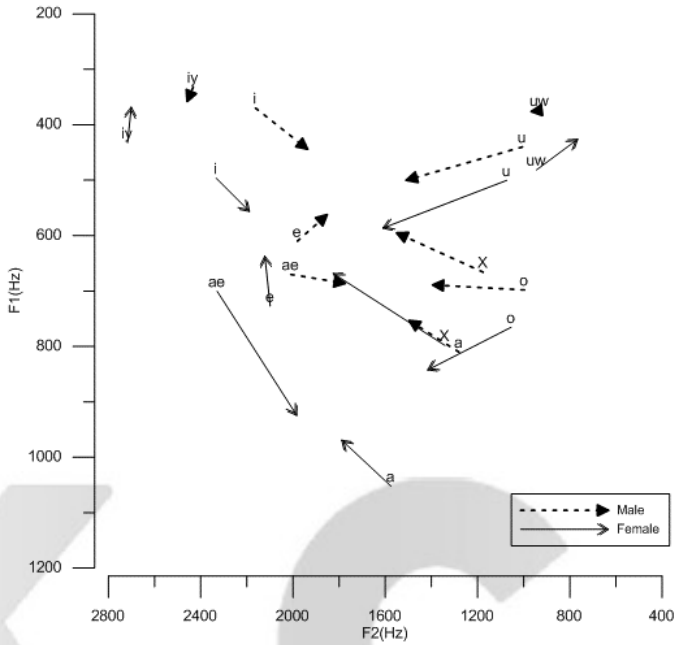

Figure 8. Plots of means of F1 and F2 measurements of the tested vowel signals across vowel types sampled at 20% and 80% of vowel duration (Hong 2015: 224)

5.2 Modeling the identification of AE vowels by AE and K listeners with a pattern recognition classifier

5.2.1 Modeling the identification of AE vowels by five AE listeners

A 10-fold cross-validation logistic regression classifier (le Cessie and van Houwelingen 1992, Hall et al. 2009) was built to categorize the identified nine vowel types of the tested vowel signals in a supervised learning mode. In the present study, the recognition model was fitted to five AE listeners' identification results with various combinations of spectral features sampled once at steady state, twice at 20% and 80% of vowel duration, steady-state F0, and duration. Each of the model training processes with different parameters was repeated ten times for statistically valid comparison between parameters. The purpose of modeling AE listeners' vowel identification performance was that the resulting fitted model for AE listeners' vowel

identification constitutes a reference to be compared to the fitted models for the vowel identification results from four different groups of K listeners. Through model comparison, the perception of four different groups of K listeners could be legitimately evaluated on their use of different cues with reference to the AE listeners' group.

Table 6 and Figure 9 show means of model classification rates with different parameters and the boxplots of the classification rates from ten times of model training and testing, respectively. Note that five AE listeners' mean correct identification rate was 100%.

Table 6. Means of classification rates for parameters with various acoustic cues to model AE listeners' identification of vowels after 10 times of trials

Parameters	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0 F11F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean	84.72	86.28	85.5	91.28	92.33	88.5	90.61
SD	1.44	1.34	.72	.83	.68	1.23	.49
N	10	10	10	10	10	10	10

Dur = duration; F0 = steady-state F0
F1, F2, F3 = F1, F2, F3 measurements sampled at steady state
F11 = F1 measurements sampled at 20% and 80% of vowel duration
F22 = F2 measurements sampled at 20% and 80% of vowel duration
F33 = F3 measurements sampled at 20% and 80% of vowel duration

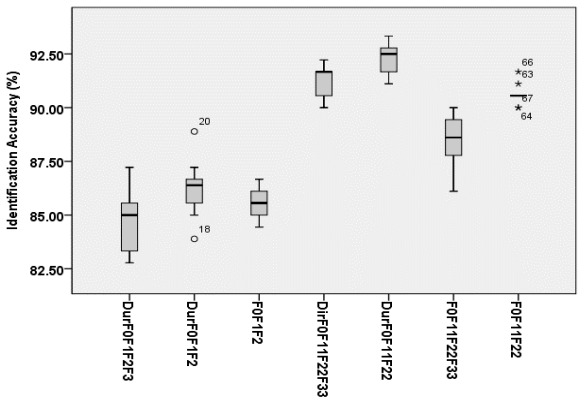

Figure 9. Boxplots of model performance for different parameters for AE listeners' model

One-way ANOVA showed that classification rates were significantly different across parameters: $F(6, 63)=93.90$, $p<0.001$. Posthoc Bonferroni revealed that more statistically substantial model performance improvement was seen with the parameters with duration, F0, and dynamic spectral frequencies (DurF0F11F22 and DurF0F11F22F33) than with duration, F0, and static spectral frequencies, as posthoc Bonferroni pairwise comparison shows in Table 7. However, no statistical accuracy improvement difference was found between DurF0F11F22 and DurF0F11F22F33. As the inclusion of dynamic spectral information of F3 did not result in substantial improvement, F33 seems to be redundant in the current model fitting. Therefore, it will be assumed in the present study that the best fit parameter for AE listeners' identification is DurF0F11F22 by Occam's Razor[8]. This suggests that AE listeners may use as major acoustic cues duration, steady-state F0, and dynamic spectral properties. And the dynamic spectral properties can be best characterized by measurements of F1 and F2 sampled at 20% and 80% of vowel duration.

Table 7. Posthoc Bonferroni pairwise comparison of parameters to model AE listeners' identification

(I) Parameter	DurF0F11F22					
(J) Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	F0F11 F22F33	F0 F11F22
Mean Diff. (I-J)	7.61*	6.06*	6.83*	1.06	3.84*	1.72*
S.E.	.46	.46	.46	.46	.46	.46
Sig.	.00	.00	.00	.50	.00	.01

5.2.2 Modeling the identification of AE vowels by Advanced K listeners' group

A 10-fold cross-validation logistic regression classifier was built to categorize the nine vowel types of the tested vowel signals. The recognition model was fitted to Advanced K listeners' identification results with the parameters of various acoustic cues. Each of the model training and testing processes was repeated ten times. Note that Advanced K listener's mean correct identification rate was 92.42% (SD=3.28, N=33)

[8] DurF0F11F22 uses less number of acoustic cues in modeling than DurF0F11F22F33 and therefore, Occam's Razor says that the former is better than the latter.

Table 8. Means of classification rates for parameters with various acoustic cues to model Advanced K listeners' identification of vowels after 10 times of trials

Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean	88.65	88.83	84.78	91.40	91.40	91.07	90.48
SD	.15	.12	.06	.03	.04	.04	.02
N	10	10	10	10	10	10	10

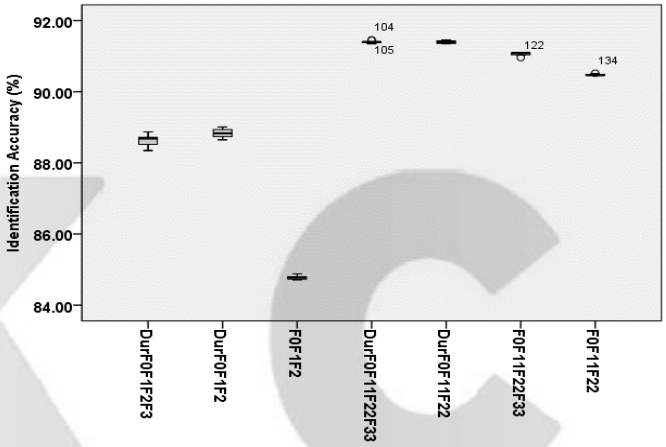

Figure 10. Boxplots of model performance for different parameters for Advanced K listeners' model

One-way ANOVA showed that classification rates were significantly different across parameters: $F(6, 63)=8640.33$, $p<0.001$. Posthoc Bonferroni revealed that DurF0F11F22 and DurF0F11F22F33 showed the best model fit, as shown in Table 9. As the two parameters showed equivalent model performance, F33 seems to be redundant in model fitting. Therefore, DurF0F11F22 turned out to be the best parameter by Occam's Razor, which was also assumed to be the best model parameter. This means that Advanced K listeners used dynamic spectral features and duration as major cues just like AE listeners did.

Table 9. Posthoc Bonferroni pairwise comparison of parameters to model Advanced K listeners' identification

(I) Parameter	DurF0F11F22					
(J) Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	F0F11 F22F33	F0 F11F22
Mean Diff. (I-J)	2.75*	2.57*	6.62*	1.11E-16	.33*	.92*
S.E.	.04	.04	.04	.04	.04	.04
Sig.	.00	.00	.00	1.00	.00	.00

5.2.3 Modeling the identification of AE vowels by Upper-intermediate K listeners' group

A 10-fold cross-validation logistic regression classifier was built to categorize the nine vowel types of the tested vowel signals. The 10-fold cross-validation logistic regression classifier model was fitted to Upper-intermediate K listeners' identification results with the same parameters. Note that Upper-intermediate K listener's mean correct identification rate was 75.70% (SD=7.57, N=34).

Table 10. Means of classification rates for parameters with various acoustic cues to model Upper-intermediate K listeners' identification of vowels after 10 times of trials

Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean	75.11	75.05	70.98	76.23	76.32	75.17	74.03
SD	.06	.05	.10	.14	.04	.11	.06
N	10	10	10	10	10	10	10

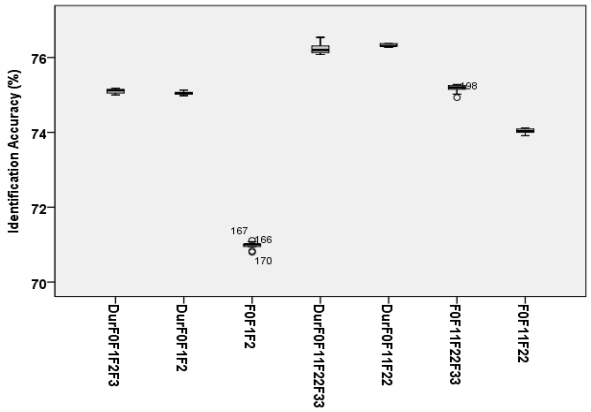

Figure 11. Boxplots of model performance for different parameters for Upper-intermediate K listeners' model

One-way ANOVA showed that classification rates were significantly different across parameters: $F(6, 63)=4296.46$, $p<0.001$. Potshot Bonferroni showed that DurF0F11F22 and DurF0F11F22F33 parameters excelled in model performance. Since the two parameters showed equivalent model performance, F33 in DurF0F11F22F33 seems to be redundant in model fitting, and hence DurF0F11F22 turned out to be the best parameter by Occam's Razor. This means that Upper-intermediate K listeners also used dynamic spectral properties and duration as major cues despite their relatively poorer identification rates than AE and Advanced K listeners. It might be the case that they knew the importance of dynamic spectral properties but they made lots of perceptual errors while they were using dynamic spectral cues.

Table 11. Posthoc Bonferroni pairwise comparison of parameters to model Upper-intermediate K listeners' identification

(I) Parameter	DurF0F11F22					
(J) Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	F0F11 F22F33	F0 F11F22
Mean Diff. (I-J)	1.22*	1.27*	5.34*	0.10	1.16*	2.30*
S.E.	.04	.04	.04	.04	.04	.04
Sig.	.00	.00	.00	.41	.00	.00

5.2.4 Modeling the identification of AE vowels by Lower-intermediate K listeners' group

The 10-fold cross-validation logistic regression classifier model was fitted to Lower-intermediate K listeners' identification results with the same parameters. Note that Lower-intermediate K listeners' mean correct identification rate was 59.56% (SD=2.73, N=33).

Table 12. Means of classification rates for parameters with various acoustic cues to model Lower-intermediate K listeners' identification of vowels after 10 times of trials

Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean	65.34	65.61	60.30	65.36	65.30	62.99	62.51
SD	.08	.09	.11	.11	.15	.16	.18
N	10	10	10	10	10	10	10

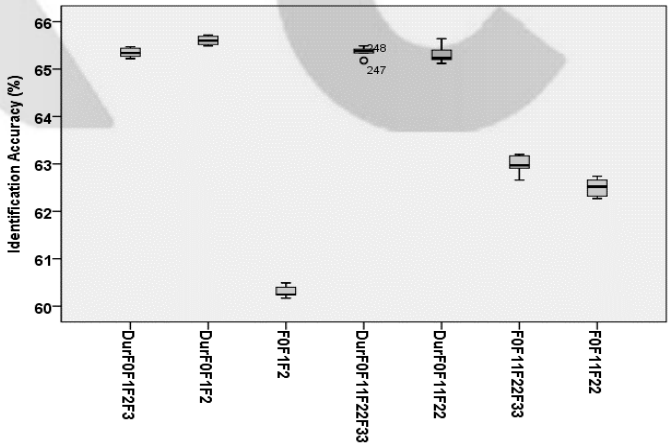

Figure 12. Boxplots of model performance for different parameters for Lower-intermediate K listeners' model

One-way ANOVA showed that classification rates were significantly different across parameters: $F(6, 63)=2375.72$, $p<0.001$. Posthoc Bonferroni showed that

DurF0F1F2 revealed the best model performance fit. This means that Lower-intermediate K listeners also used duration like AE listeners. However, they used static spectral properties as major cues unlike AE listeners. Note that AE listeners and Advanced and Upper-intermediate K listeners used dynamic spectral properties and duration as major cues.

Table 13. Posthoc Bonferroni pairwise comparison of parameters to model Lower-intermediate K listeners' identification

(I) Parameter	DurF0F1F2					
(J) Parameter	DurF0 F1F2F3	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean Diff. (I-J)	.26*	5.30*	.25*	.31*	2.62*	3.10*
S.E.	.06	.06	.06	.06	.06	.06
Sig.	.00	.00	.00	.00	.00	.00

5.2.5 Modeling the identification of AE vowels by Beginner's K listeners' group

The 10-fold cross-validation logistic regression classifier model was fitted to Beginner's K listeners' identification results with the same parameters. Note that Beginner's K listeners' mean correct identification rate was 49.60% (SD=4.98, N=33).

Table 14. Means of classification rates for parameters with various acoustic cues to model Beginner's K listeners' identification of vowels after 10 times of trials

Parameter	DurF0 F1F2F3	DurF0 F1F2	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean	57.92	58.28	54.4	57.95	57.68	56	55.56
SD	.09	.12	.15	.07	.13	.15	.09
N	10	10	10	10	10	10	10

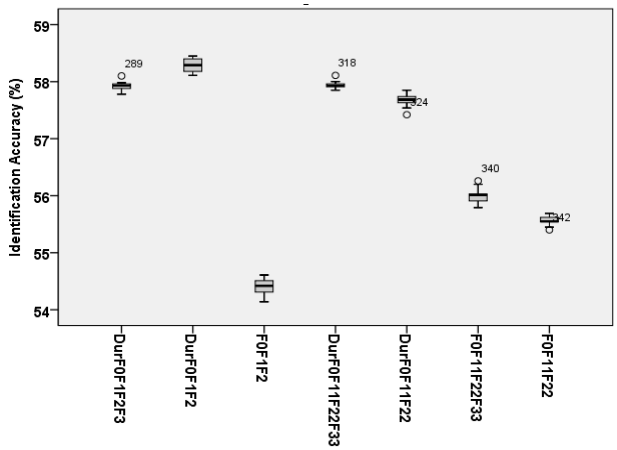

Figure 13. Boxplots of model performance for different parameters for Beginner's model

One-way ANOVA showed that classification rates were significantly different across parameters: $F(6, 63)=1636.53$, $p<0.001$. Posthoc Bonferroni showed that DurF0F1F2 revealed the best model performance fit. This means that Beginner's K listeners also used duration like AE listeners. However, they used static spectral properties as major cues like Lower-intermediate K listeners. However, the two groups of K listeners were different form AE listeners, and Advanced and Upper-intermediate K listeners in that the former two groups used static spectral properties but the latter three groups, dynamic spectral properties.

Table 15. Posthoc Bonferroni pairwise comparison of parameters to model Beginner's K listeners' identification

(I) Parameter	DurF0F1F2					
(J) Parameter	DurF0 F1F2F3	F0 F1F2	DurF0F11 F22F33	DurF0 F11F22	F0F11 F22F33	F0 F11F22
Mean Diff. (I-J)	.36*	3.89*	.34*	.61*	2.29*	2.73*
S.E.	.05	.05	.05	.05	.05	.05
Sig.	.00	.00	.00	.00	.00	.00

6. Summary and Conclusion

It has been shown that four different levels of K listeners appealed to either almost the same or totally different AE vowel identification strategies in the ways to use duration, F0 and dynamic or static spectral cues when compared with AE listeners' strategy, as summarized in Table 16.

Table 16. Usage of cues as major cues for the AE vowel identification across subject groups

Groups	Duration	Static spectral cues	Dynamic spectral cues
AE	O		O
Advanced Korean	O		O
Upper-intermediate Korean	O		O
Lower-intermediate Korean	O	O	
Beginner's Korean	O	O	

Advanced and Upper-intermediate K listeners used the same strategy as AE listeners for AE vowel identification, since all the three groups used dynamic spectral cues. However, Lower-intermediate and Beginner's K listeners used static spectral cues. However, all five groups used duration as a major cue. Lower-intermediate and Beginner's K listeners did not recognize that dynamic spectral properties are important to identify AE vowels correctly. Among 133 K listeners, only those who scored identification accuracy above the average, could pick up dynamic spectral properties.

Table 17 below compares mean correct identification rates of all five human listeners' groups along with their corresponding mean model classification rates. AE listeners in the experiment identified the tested vowels as intended vowels 100% correct. However, the model classified the vowel signals as intended vowels with only 92.33% accuracy with DurF0F11F22. Unfortunately, the model classification rate was lower than AE listeners'. However, the mean model classification rates for Advanced and Upper-intermediate K listeners were quite comparable to human listeners' mean identification rates (92.42% for Advanced K listeners and 91.40% for their model; 75.70% for Upper-intermediate and 76.32% for their model). However, the present models overclassified the vowel signals more than Lower-intermediate and Beginner's K listeners (59.56% for Lower-intermediate K listeners and 65.61%

for their model; 49.60% for Beginner's and 58.28% for their model), suggesting that K listeners with below-average identification scores made more random identification errors than the other K listeners, which the models could not handle properly in classification.

Table 17. Comparison between AE and K listeners' identification and the model's classification

Groups	Mean correct Identification (%)	Best parameter	
		Parameter	Model Classification (%)
AE	100 (0)	DurF0F11F22	92.33 (.68)
Advanced Korean	92.42 (3.28)	DurF0F11F22	91.40 (.04)
Upper-intermediate Korean	75.70 (7.57)	DurF0F11F22	76.32 (.04)
Lower-intermediate Korean	59.56 (2.73)	DurF0F1F2	65.61 (,09)
Beginner's Korean	49.60 (4.98)	DurF0F1F2	58.28 (.12)

In addition to the model's overclassification, the modeling analysis in the present study has many other limitations. The focus of the present study has been placed on AE and K listeners' identification of AE monophthong vowel signals in /hVd/ syllables in a laboratory environment. Even though AE vowel category identification by AE and K listeners could be more or less safely modeled with dynamic or static spectral change and duration, the modeling in the current study has been restricted only to identification of the AE vowel signals in citation form, which tend to be realized longer than those in connected and casual speech. Furthermore, due to the restriction on the phonological environments before and after the vowel signals (namely, /h/ and /d/), coarticulatory effects on the target vowel have been completely ignored, which may turn out to be more severe in connected and casual speech than in citation form. Therefore, the pattern recognition modeling approach tried in the present study may not properly characterize the AE and K listeners' perception of AE vowels in variable environments and situations. These problems are pending questions for further study.

REFERENCES

ASSMANN, PETER F. and GEOFFREY S. MORRISON. 2013. Introduction. In Geoffrey S. Morrison and Peter F. Assmann (eds.). *Vowel Inherent Spectral Change, Modern Acoustics and Signal Processing*, 1-6. Berlin: Springer.

FLEGE, JAMES E., OCKE-SCHWEN BOHN and SUNYOUNG JANG. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 437-470

HALL, MARK, EIBE FRANK, GEOFFREY HOLMES, BERNHARD PFAHRINGER, PETER REUTEMANN and IAN H. WITTEN. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11, 10-18.

HILLENBRAND, JAMES M. 2013. Static and dynamic approaches to vowel perception. In Geoffrey S. Morrison and Peter F. Assmann (eds.). *Vowel Inherent Spectral Change, Modern Acoustics and Signal Processing*, 9-30. Berlin: Springer.

HILLENBRAND, JAMES M., MICHAEL J. CLARK and ROBERT A. HOUDE. 2000. Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America* 108, 3013-3022.

HILLENBRAND, JAMES M. and ROBERT T. GAYVERT. 1993. Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *Journal of the Acoustical Society of America* 94, 668-674.

_____. 2005. Open source software for experiment design and control. *Journal of Speech, Language, and Hearing Research* 48, 45-60.

HILLENBRAND, JAMES M., LAURA A. GETTY, MICHAEL J. CLARK and KIMBERLEE WHEELER. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97, 3099-3111.

HILLENBRAND, JAMES M. and TERRANCE M. NEAREY. 1999. Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America* 105, 3509-3523.

HONG, SOONHYUN. 2014. Training effects after training Korean listeners for the contrast of /ɑ, ɔ, ʌ/. *Language and Linguistics* 65, 299-329.

_____ 2015. Pattern recognition modeling of Korean listeners' perception of American English monophthongs. *Language and Linguistics* 68, 209-239.

INGRAM, JOHN and SEE-GYOON PARK. 1997. Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics* 25, 343-370.

IVERSON, PAUL, PATRICIA K. KUHL, REIKO AKAHANE-YAMADA, EUGEN DIESCH, YOH'ICH TOHKURA, ANDREAS KETTERMANN and CLAUDIA SIEBERT. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47-B57.

JENKINS, JAMES J., WINIFRED STRANGE and THOMAS R. EDMAN. 1983. Identification of vowels in 'vowelless' syllables. *Perception & Psychophysics* 34, 441-450.

KOREAN MINISTRY of EDUCATION. 1988. Pyojune Kyojung ("Rules in Standard Korean"). *Kuke Sanghwal* 13, 79-108.

LADEFOGED, PETER and KEITH JOHNSON. 2011. *A Course in Phonetics* (6th edition). Boston: Cengage Learning.

LE CESSIE, SASKIA and J. C. VAN HOUWELINGEN. 1992. Ridge estimators in logistic regression. *Applied Statistics* 41, 191-201.

MORRISON, GEOFFREY S. 2013. Theories of vowel inherent spectral change: A review. In Morrison Geoffrey S. and Peter F. Assmann (eds.). *Vowel Inherent Spectral Change, Modern Acoustics and Signal Processing*, 31-47. Berlin: Springer.

NEAREY, TERRANCE M. 1989. Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America* 85, 2088-2113.

NEAREY, TERRANCE M. and PETER ASSMANN. 1986. Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America* 80, 1297-1308.

PETERSON, GORDON E. and HAROLD L. BARNEY. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175-184.

YUN, YUNG-DO. 2005. Korean listeners' perception of English /i/, /ɪ/, and /ɛ/. *Speech Science* 12.1, 75-87.

ZAHORIAN, STEPHEN A. and AMIR JALALI JAGHARGHI. 1993. Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America* 94, 1966-1982.

Soonhyun Hong
Department of English Language and Literature
Inha University
100 Inharo, Nam-gu, Incheon
Korea 22212
e-mail: shong@inha.ac.kr

received: February 14, 2016
revised: March 14, 2016
accepted: March 31, 2016