http://dx.doi.org/10.17959/sppm.2016.22.3.583

Vowel inherent spectral properties characterized in Korean and American English talkers' English vowel signals: A production-based pattern recognition modeling study* **

Soonhyun Hong
(Inha University)

Hong, Soonhyun. 2016. Vowel inherent spectral properties characterized in Korean and American English talkers' English vowel signals: A production-based pattern recognition modeling study. *Studies in Phonetics, Phonology and Morphology* 22.3. 583-609. Hillenbrand et al. (1995) showed in a pattern recognition modeling study of American English listeners' vowel perception that American English vowel signals are characterized by dynamic spectral properties, and that American English listeners' vowel perception can be modeled best with American English vowel inherent dynamic spectral properties. The present study, on the other hand, investigated the production side of the vowel inherent spectral properties of American English talkers' and Korean talkers' English vowel signals, by fitting four different production-based pattern recognition classification models (K-Nearest Neighbors, Random Forests, Support Vector Machine and Logistic Regression) directly to American English and Korean talkers' English vowel signals. The results showed that American English talkers' vowel signals were best characterized by dynamic spectral properties but Korean talkers' English vowel signals by static spectral properties. **(Inha University)**

Keywords: Vowel inherent spectral change, English vowel production, pattern recognition modeling, classification of English vowels

1. Introduction

Under the influence of Peterson and Barney (1952), it was implicitly assumed in the literature on vowel production that American English (AE) vowels are well characterized by steady-state or vowel-midpoint Formant 1 (F1) and 2 (F2)

* This work was supported by INHA UNIVERSITY Research Grant.

** I would like to thank anonymous reviewers for their insightful comments and suggestions for this paper.

measurements (Ladefoged and Johnson 2011). The following illustrate plots of steady-state F1 and F2 measurements of signals of twelve AE vowel types (/i, ɪ, ɛ, a, ɑ, ɔ, ʊ, u, ʌ, eɪ, oʊ, ɚ/) in /hVd/ syllables produced by 45 males, 48 females, and 46 children from the Michigan area in the U.S. in Hillenbrand et al. (1995).

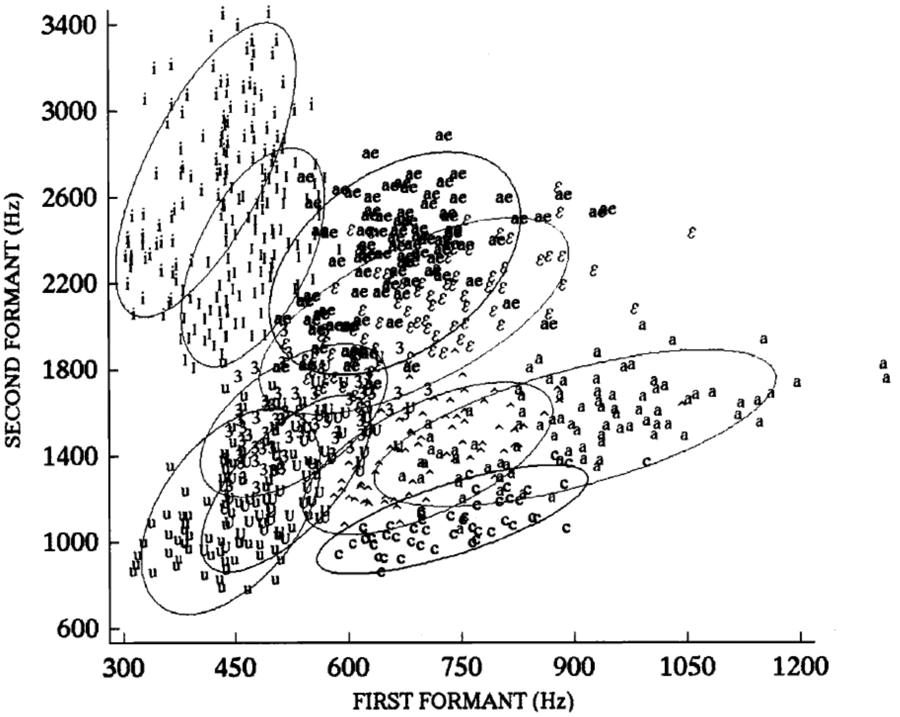

Figure 1. Steady-state values for 45 men, 48 women, and 46 children for 10 vowels with ellipses fit to the data ("ae"=/a/, "a"=/ɑ/, "c"=/ɔ/, "^"=/ʌ/, "3"=/ɚ/) (from Hillenbrand et al. 1995: 3104)

The plots in Figure 1 show a long and wide spread of each vowel type and extensive overlaps between vowel types, which cannot explain the observation that these vowel signals were identified as intended vowels by more than 90% of AE listeners. This indicates that AE listeners do not perceive AE vowels only with static spectral measurements sampled once at steady state.

Hillenbrand et al. (1995) and Nearey and Assmann (1986) further showed that AE monophthong vowels have drastic spectral change like diphthongs in the F1/F2 plane.

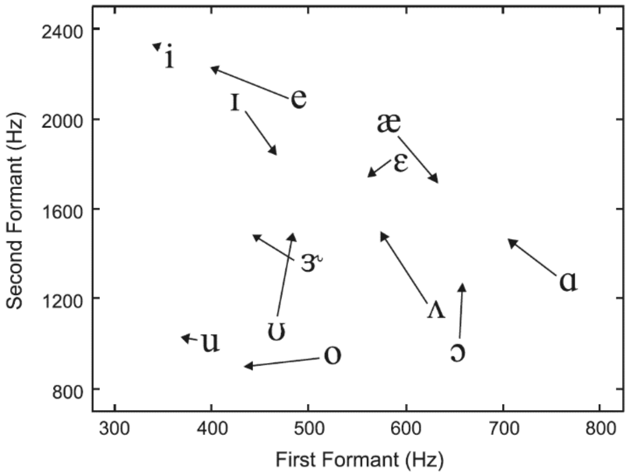

Figure 2. F1 and F2 measurements sampled at 20% and 80% of vowel duration for vowels in /hVd/ syllables spoken by 45 men from the Upper Midwest (Hillenbrand 2013: 13) (/e/=/eɪ/ and /o/=/oʊ/)

In Figure 2, all the monophthongs except for /i/ and /u/ show drastic spectral change like diphthongal /e/ (/eɪ/) and /o/ (/oʊ/). Hillenbrand et al. (1995) and Nearey and Assmann (1986) argued that AE vowels may be characterized better by dynamic spectral properties than static spectral properties.

2. Background of the study

Hillenbrand et al. (1995) proposed through pattern recognition modeling that AE listeners perceive vowels by picking up the dynamic spectral properties of AE vowels, which can be best captured with spectral measurements sampled twice at 20% and 80% of vowel duration. They conducted a perception experiment to verify the proposal. First, they recorded vowel signals of 12 vowel types (/i, ɪ, ɛ, a, ɑ, ɔ, ʌ, ʊ, u, eɪ, oʊ, ɝ/) in /hVd/ syllables produced by 45 males, 48 females, and 46 children. Then, they forced 20 AE listeners to identify these vowel signals and selected only AE vowel signals that were identified as intended vowels by 85% or greater number of the AE listeners (88.5% of the tokens in the database). As a next step, they trained and tested a pattern recognition quadratic discriminant classifier with a “Jackknife” technique to 20 AE listeners’ identification results in supervised learning mode, with various selections of measurements of F0 at steady state, duration, and F1-F3 sampled at 20%, 50%, and 80% of vowel duration fitted to the model. The results showed that model classification excelled with measurements of F1-F3 sampled twice at 20% and 80% of vowel duration with F0 and duration (2 samples) (97.8%) and with measurements of F1-F3 sampled three times at 20%, 50% and 80% of vowel duration with F0 and duration (3 samples) (97.3%). However, there was almost no identification accuracy difference between models with 2 samples and 3 samples. On the other hand, the model with steady-state F1-F3 with F0 and duration (91.6%) performed worst. By Occam’s Razor[1], the 2 samples parameter set turned out to be the best for the model of AE listener’ vowel perception: steady-state F0, duration and 2 samples of F1-F3. Namely, the pattern recognition model with 2 samples of formant measurements could model AE listeners’ correct identification best. They concluded that AE listeners pick up the dynamic spectral properties of AE vowels when they identify vowels, and the dynamic spectral properties can be best characterized by spectral measurements sampled twice at 20% and 80% of vowel duration along with measurements of duration and steady-state F0.

Nearey and Assmann (1986), Nearey (1989), Zahorian and Jagharghi (1993), Hillenbrand et al. (1995), Hillenbrand and Nearey (1999), Morrison (2013) also independently pointed out that AE vowels could be classified more accurately and robustly when dynamic spectral measurements were used as spectral cues than when static spectral measurements were used. All of these studies indicated that “vowel inherent spectral change” (VISC[2]), namely, dynamic spectral features of AE vowels, plays an important role in AE listeners’ vowel perception.

Pattern recognition modeling studies on Korean (K) listeners’ perception of AE

[1] Occam’s Razor says that when two models with different number of parameters show the same performance, the model with less parameters are better than the one with more parameters

[2] “[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).”

vowel signals are found in Hong (2015, 2016), who tried to explain whether K listeners use dynamic spectral properties or static spectral properties when they perceive AE vowels. Hong (2015) conducted a pattern recognition modeling study on K listeners' English vowel perception. He puzzled whether K listeners' difficulties differentiating among AE vowels may result from the hypothesis that they might not use spectral vowel change the way AE listeners do as in Hillenbrand et al. (1995). He administered a forced-choice vowel identification task to 57 K listeners, using AE vowel signals selected from Hillenbrand et al. (1995). Then a 10-fold cross-validation[3] Logistic Regression classification model was fitted to K listeners' correct identification results with the same parameter sets used in Hillenbrand et al. (1995). He found that DurF0F1F2_ST[4] (62.73%), DurF0F1F2F3_2080[5] (62.89%) and DurF0F1F2_2080 (62.74%) showed the best performance. No model performance difference was found among the three best performing parameter sets: static DurF0F1F2_ST (62.73%) and dynamic DurF0F1F2F3_2080 (62.89%) and DurF0F1F2_2080 (62.74%). The model showed equivalent model fitting improvement both with the static and dynamic parameter sets. This means that dynamic spectral information was redundant in modeling K listeners' AE vowel perception. Therefore, the model with DurF0F1F2 turned out the best results by Occam's Razor. Hong (2015) concluded that K listeners used static spectral properties (but not dynamic properties) along with duration and F0 unlike AE listeners, when they identified AE vowels.

Hong (2016) also conducted a similar perception-based Logistic Regression pattern recognition modeling study based on K listeners' English vowel identification results. He divided 133 K listeners into four groups based on their correct identification performance on AE vowel signals in /hVd/ syllables. It was found that the two upper-level groups of K listeners used dynamic spectral properties when they perceived English vowels whereas the two lower-level groups used static spectral

[3] "10-fold cross-validation" refers to the classification procedure in which the training sample is classified (or tested) based on rules or functions obtained from all the other cases in the data, namely, one part of samples out of 10 parts of the whole data is "held out" for classification while the other parts are trained, one at a time and hence 10 times total.

[4] DurF0F1F2_ST refers to the parameter set of measurements of duration, mean F0 and F1 and F2 sampled once at steady state.

[5] DurF0F1F2F3_2080 refers to the parameter set of measurements of duration, mean F0 and F1-F3 sampled twice at 20% and 80% of vowel duration.

properties.

The previous pattern recognition modeling studies in Hillenbrand et al (1995) and Hong (2015, 2016), were focused on AE and K listeners’ perception of AE vowels. They conducted perception-based pattern recognition modeling studies on AE and K listeners’ identification of AE vowel signals, and drew the conclusion that AE listeners pick up dynamic spectral properties when they identify AE vowels whereas K listeners, static spectral properties.

On the other hand, studies on K talkers’ English vowel production in terms of spectral change are found in Oh (2013) and Chung et al. (2010). Oh (2013) conducted a study on K talkers’ production of English vowels with a focus on spectral change. She measured F1-F3 values of six K talkers’ and eight AE talkers’ English front vowel signals /i, ɪ, ɛ, æ/ in /bVt/ syllables, at 11 points along vowel duration. Then she compared K talkers’ English vowel spectral trajectory patterns with AE talkers’. She observed that K talkers’ English vowel spectral change did not pattern together with AE talkers’. Unfortunately, however, the focus of her study was restricted only to front vowels and she did not spell out in what way AE and K talkers’ English front vowel signals were different in terms of spectral change. Chung et al. (2010) also tried to compare six K talkers’ K vowel trajectory patterns after /s, ʃ, s’, ʃ’ / with ten AE talkers’ English vowel trajectory patterns after /s, ʃ/. They pointed out that Korean vowel signals showed minimal spectral change compared with AE vowel signals. This suggests that Korean talkers may produce English vowels with minimal spectral change unlike AE talkers.

The present pattern recognition modeling study on AE and K talkers’ English vowel production is different from previous modeling studies (Hillenbrand et al. 1995 and others mentioned previously) in that the former is purely based on English vowel production whereas the latter are on English vowel perception. We tried to investigate whether K talkers’ English vowel signals are best characterized by dynamic spectral properties or not. We began with the assumption that AE vowels’ spectral changes are major cues to perceive AE vowels (Hillenbrand et al. 1995, Hong 2015, 2016). Through production-based pattern recognition modeling of the AE talkers’ vowel production, we tried to investigate whether AE talkers really use vowel spectral changes as major cues in production. Such an inquiry is warranted because by assumption, vowel spectral changes in production will enable AE listeners to perceive AE vowel signals distinctively. In addition, we also would like to find whether K talkers use vowel spectral changes as major cues in their English vowel production.

For these purposes, four perception-based pattern recognition classification models for AE talkers' vowel signals, were built based on a selection of the speech data in Hillenbrand et al. (1995) in Orange 3.3.7 (Demsar et al. 2013): K-Nearest Neighbors[6] (kNN), Random Forests[7] (RF), Support Vector Machine[8] (SVM), and Logistic Regression (LR) classifier. Then the models were fitted with a 10-fold cross-validation technique to different selections of major acoustic cue measurements (used in previous studies) of AE talkers' vowel signals, in a supervised learning mode. Through the predicted identification accuracy comparison between the models with different parameter sets, we derived the models with the best parameter sets for AE talkers' vowel production by checking which parameter sets (dynamic or static parameter sets) best improve the models' performance on AE vowel signals. Then we let the best-fit production-based AE models (hereafter, production-based AE models) directly identify K talkers' English vowel signals to see K talkers use dynamic spectral cues like AE talkers.

The hypotheses to be tested in the present study are as follows:

1. The production-based AE models for AE talkers' vowel signals improve best with dynamic parameter sets.
2. The proposed production-based AE models (with dynamic parameter sets) show very poor performance in identifying K talkers' English vowel signals as intended English vowels.
3. K talkers' English vowel signals are best characterized by static parameter sets rather than dynamic parameter sets unlike AE talkers'.

[6] According to K-Nearest Neighbor classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

[7] Random Forests classification is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).

[8] Given a set of training data points with each marked as belonging to one or the other categories, Support Vector Machine builds a model that assigns new data points to one category or the other. The resulting model turns into non-probabilistic binary linear classifier.

The first two hypotheses may be addressed by building the AE models with the best parameter sets for AE talkers' vowel signals. Then, we let the AE models to identify K talkers' English vowel signals. Our prediction is that the AE models with dynamic parameter sets will perform best on AE talkers' vowel signals but perform quite poorly on K talkers' signals. The third hypothesis will be verified by building production-based Korean models, which will be fitted directly for K talkers' English vowel signals with either dynamic or static parameter sets. If we are on the right track, it is predicted that the K models' performance on K talkers' English vowel signals excels with static parameter sets unlike AE models' on the same signals. It is also predicted that the correct identification performance of the Korean models with static parameter sets (hereafter, K models) on K talkers' signals, are far better than that of the AE models with dynamic parameter sets on the same K signals. If these predictions are correct, verified is the third hypothesis that K talkers' English vowel signals are best characterized by static spectral properties.

3. Experiment

3.1 Subjects

The subject group consisted of 18 college-level K talkers (11 females and 7 males) whose age varied from 19 to 25 (mean=21.7). They had been learning English at least 9 years. All the subjects had no reported history of speech or hearing problems. They received a partial credit for a Phonetic course by taking part in the experiment.

3.2 Stimuli

A set of English monophthong /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/ signals in /hVd/ syllables was used in the present K talkers' vowel production experiment. A total of 90 vowel signals (10 signals for each of the 9 monophthong vowel types) which were identified by more than 85% of 20 AE listeners, were chosen among the signals produced by male or female talkers in the database in Hillenbrand et al. (1995), Each of these vowel signals was to be presented to K subjects before they produced the same English vowel type.

3.3 Procedure

A recording protocol was built in Alvin 2.0 (Hillenbrand and Gayvert 2005) to be used to record K talkers' English vowels during the recording session. It randomized the 90 AE vowel signals in /hVd/ syllables before recording. The subject with a pair of headphones on, clicked on the "Start" button on the computer screen while the author was monitoring the whole recording procedure. The carrier sentence, e.g., "Say /hæd/ now" with the phonetic symbol in the /hVd/ syllable and the corresponding word "had" prompted on the computer screen. At the same time, the /hæd/ signal was presented to the subject. Then s/he was forced to read the carrier sentence through the microphone (Infrasonic's UFO) while recording at the sampling rate of 16kHz was in progress. After reading the sentence, s/he clicked on the "End" button on the computer screen to finish recording. Then s/he clicked on the "Next" button for the next sentence. This procedure was repeated for all 90 AE vowel signals. It took less than 8 minutes for each recording session, which was administered at a phonetics lab. The total number of the recorded signals is 1620 (=10 signals for each vowel * 9 vowels * 18 subjects).

4 Results

A Praat script was written to automatically measure vowel duration and the mean of F0 measurements sampled at 20%, 50% and 80% of vowel duration in the recorded English vowel signals of K talkers', once the vowel portion was segmented by the author. It also automatically measured F1-F3 of the vowel signals sampled at 20%, 50% and 80% of vowel duration based on the Burg algorithm. Then the author checked all the cue measurements on each of the speech signals and corrected mis-measurements due to the algorithm's mistracking by readjusting the maximum formant value and the permitted number of formants in the Burg algorithm.

Table 1. The mean measurements of duration, F0, and the mean spectral measurements sampled at 20%, 50%, and 80% of vowel duration of K female talkers' English vowel signals with SD[9]

K Females		Dur	F0	F1_20	F2_20	F3_20	F1_50	F2_50	F3_50	F1_80	F2_80	F3_80	N
ɑ	Mean	234	242	890	1257	3005	895	1273	3031	844	1401	2993	110
	SD	94.3	23.8	111.1	138.4	310.4	116. 9	159.6	326. 7	88.8	226.9	322.0	
æ	Mean	262	235	889	2043	3034	904	2011	3048	846	1997	3049	110
	SD	95.9	24.8	109.6	199.3	264.6	119.2	227.2	292.7	107.7	225.0	279.1	
ɛ	Mean	223	241	831	2059	3038	825	2040	3055	756	2030	3074	110
	SD	38.6	20.6	100.0	191.8	259.5	97.2	180.2	254.1	84.9	190.3	257.2	
ɪ	Mean	209	237	442	2759	3545	438	2772	3555	436	2739	3516	110
	SD	79.0	27. 5	77.2	255.8	445.5	75.8	239.3	460.2	78.3	252.0	431.5	
i	Mean	255	260	369	2885	3716	379	2879	3702	388	2845	3684	110
	SD	92.1	28.1	84.3	199.0	358.9	92.3	194.9	358.8	103.6	214.0	388.8	
ɔ	Mean	272	240	814	1207	3034	807	1209	3071	759	1274	3062	110
	SD	66.7	23.5	99.3	161.5	324.2	99.6	157.8	331.0	97.7	201.9	330.0	
ʊ	Mean	189	256	510	1355	2961	505	1348	2974	488	1505	3000	110
	SD	42.1	25.7	78.5	166.7	285.1	82.7	193.0	257.8	66.1	221.6	235.3	
u	Mean	263	250	465	1292	2877	455	1233	2893	445	1302	2920	110
	SD	55.1	23.6	71.5	191.6	277.3	73.0	184.3	271.3	70.7	195.3	280.7	
ʌ	Mean	214	243	811	1293	3032	804	1326	3041	739	1440	3048	110
	SD	42.2	21.8	77.5	163.2	361.4	69.9	189.4	351.9	80.3	236.5	336.6	

[9] F1_20, F1_50, and F1_80 refer to F1, F2 and F3 measurements sampled at 20%, 50% and 80% of vowel duration, respectively.

Table 2. The mean measurements of duration, F0, and the mean spectral measurements sampled at 20%, 50%, and 80% of vowel duration of K male talkers' English vowel signals with SD

K Males		Dur	F0	F1_20	F2_20	F3_20	F1_50	F2_50	F3_50	F1_80	F2_80	F3_80	N
ɑ	Mean	211	136	748	1166	2744	730	1161	2756	684	1211	2702	70
	SD	54.7	10.3	98.9	115.8	259.7	99.3	134.8	247.1	75.3	123.0	226.9	
æ	Mean	241	133	688	1860	2648	705	1820	2661	681	1723	2653	70
	SD	52.1	11.0	88.9	94.6	213.6	80.7	90.2	203.9	86.3	110.1	179.4	
ɛ	Mean	204	135	628	1853	2625	639	1822	2634	608	1747	2639	70
	SD	42.0	11.4	53.3	85.6	186.7	55.8	88.7	177.8	49.8	103.2	154.4	
ɪ	Mean	202	136	394	2173	2873	407	2183	2868	412	2136	2802	70
	SD	47.9	12.1	46.3	121.7	272.8	40.9	117.9	260.1	40.5	129.8	231.0	
i	Mean	214	148	339	2249	2961	342	2264	2972	356	2235	2892	70
	SD	38.0	10.4	48.7	128.1	252.8	43.8	114.0	274.8	39.5	114.4	251.2	
ɔ	Mean	260	133	703	1095	2676	684	1077	2693	617	1103	2673	70
	SD	81.3	10.5	70.7	84.2	237.3	77.1	75.8	272.9	66.8	86.3	285.5	
ʊ	Mean	184	148	436	1192	2514	427	1185	2550	419	1314	2549	70
	SD	39.1	13.0	39.7	199.7	173.9	42.2	205.8	179.5	40.2	151.2	146.2	
u	Mean	232	139	391	1149	2461	394	1131	2497	387	1215	2528	70
	SD	54.6	12.4	44.5	211.3	182.8	34.3	188.1	164.4	37.9	172.2	166.7	
ʌ	Mean	206	136	679	1139	2721	680	1143	2732	621	1214	2710	70
	SD	50.5	12.6	73.7	97.3	281.8	70.8	104.9	280.0	60.9	96.0	268.0	

When mean F1/F2 measurements of the recorded K male and female talkers' English vowel signals sampled at 50% of vowel duration were plotted, /ɪ, i/, /ʊ, u/, /ɛ, æ/, and /ɑ, ɔ, ʌ/ formed their own groups. Due to incompatibility with the graphing software, /iy, i, e, æ, a, o, x, u, uw/ in the figures in this paper refer to /i, ɪ, ɛ, æ, ɑ, ɔ, ʌ, ʊ, u/, respectively.

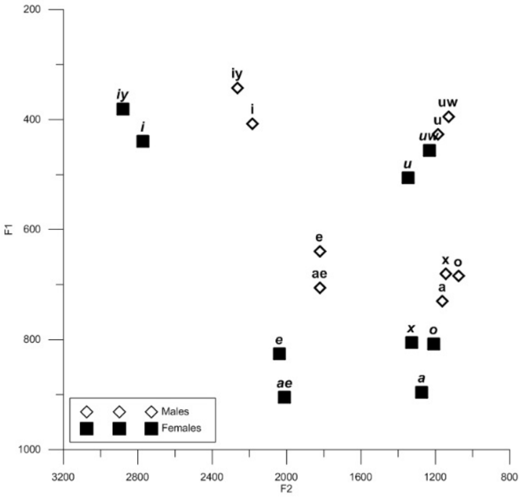

Figure 3. Plots of mean F1/F2 measurements of K male and female talkers' AE vowel signals across vowel types, sampled at 50% of vowel duration

As /ɪ, i/, /ʊ, u/, /ɛ, æ/ are perceptually nondistinctive in Korean (Hong 2012), it is naturally predicted that the English vowels of these pairs are produced similarly and are represented close together in the F1/F2 plane. It was already observed in K listeners' AE vowel perception studies in Hong (2013, 2014) that K listeners have perceptual difficulties distinguishing between /ɑ, ɔ, ʌ/. This is the same case in K talkers' English vowel production. The plotted K talkers' /ɑ, ɔ, ʌ/ signals represented as a single group in the F1/F2 plane is naturally expected. This means that K male and female talkers did not distinguish between English /ɪ, i/, /ʊ, u/, /ɛ, æ/, and /ɑ, ɔ, ʌ/, respectively, in their English vowel production, either.

Note that Hillenbrand et al. (1995) reported that dynamic spectral change was found in AE male and female talkers' vowel signals. To check whether there is drastic spectral change along vowel duration in K male and female talkers' English vowel signals, mean F1/F2 measurements sampled twice at 20% and 80% of vowel duration were plotted.

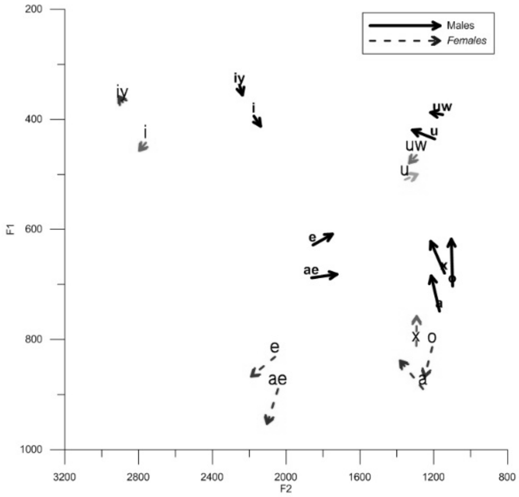

Figure 4. Plots of mean F1/F2 measurements of K male and female talkers' AE vowel signals across vowel types, sampled at 20% and 80% of vowel duration

In Figure 4, relatively little spectral changes were observed in K male and female talkers' vowel signals when compared with those in AE talkers' in Hillenbrand et al (1995) in Figure 2 and those in male and female AE talkers' vowel signals partially selected from Hillenbrand et al (1995) for the present study (see Figure 6 below).

5. Discussion

5.1 Building a corpus of AE talkers' vowel signals

In order to build production-based pattern recognition classification models for AE talkers' vowel production, an AE vowel signal corpus was built by collecting all the AE males' and females' vowel signals from the database in Hillenbrand et al. (1995) that were identified as intended vowels by 85% or greater of AE talkers. The following are plots of means of F1/F2 measurements of the vowel signals in the corpus sampled at 50% of AE vowel duration (Figure 5) and plots of means of F1/F2 measurements of the same AE vowel signals sampled twice at 20% and 50% of vowel duration (Figure 6):

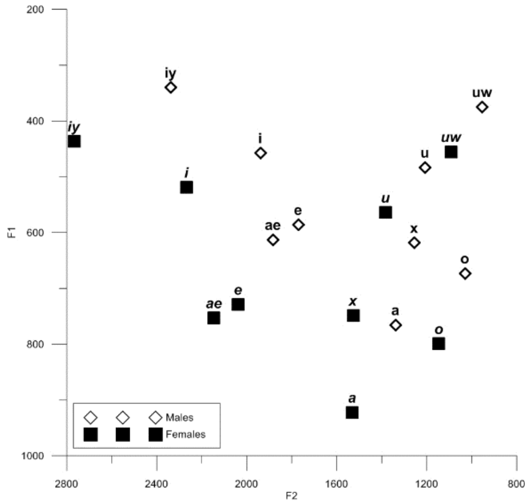

Figure 5. Plots of means of F1/F2 measurements of AE male and female talkers' monophthong vowel signals, sampled at 50% of vowel duration. All of these vowel signals were identified as intended vowels by more than 85% of AE listeners according to Hillenbrand et al. (1995)

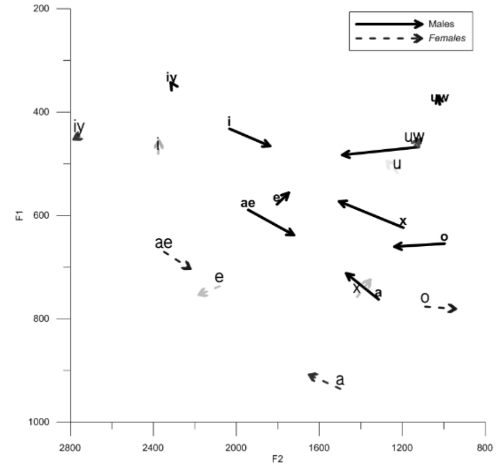

Figure 6. Plots of mean F1/F2 measurements of AE male and female talkers' monophthong vowel signals, sampled at 20% and 80% of vowel duration. All of these vowel signals were identified as intended vowels by more than 85% of AE listeners according to Hillenbrand et al. (1995)

When the spectral change of K talkers' vowel signals in Figures 4, were compared with AE talkers' in Figure 6, K talkers' English vowel signals are relatively stable in spectral change. In the next subsection, we are going to build the pattern recognition models with selective cue parameters fitted to AE talkers' vowel production, based on these AE vowel signals.

5.2 Modeling AE talkers' English vowel production: AE models

Production-based kNN, RF, SVM, and LR classifiers were built in Orange 3.3.7 (Demsar et al. 2013) to categorize the 9 English vowel types from the AE vowel signals in the corpus. In 10-fold cross-validation supervised learning mode, all the four pattern recognition models were fitted to various selections of the cue measurements which included vowel duration and F1, F2 and F3 sampled once at 50% of vowel duration, twice at 20% and 80% of vowel duration, and mean F0 which averaged F0 measurements sampled at 20%, 50% and 80% of vowel duration: DurF0F123_50, DurF0F12_50, F0F123_50, F0F12_50, DurF0F123_2080, DurF0F12_2080, F0F123_2080, and F0F12_2080[10].

Table 3. The identification accuracies by the four AE models with static and dynamic parameter sets, on AE talkers' English vowel signals in the corpus through 10-fold cross-validation

AE vowel signals predicted	Static parameter sets				Dynamic parameter sets			
	DurF0F 123_50*	DurF0F 12_50	F0F123 _50	F0F12_ 50	DurF0F12 3_2080	DurF0F 12_2080	F0F123_ 2080**	F0F12_ 2080
kNN	88.7	90.9	83.2	85.3	93.8	96.9	92.2	94.9
RF	87.9	89.8	84.7	83.2	93	92.8	92.2	92
SVM	90.6	92.2	87.9	87.3	94.6	96.8	94.5	96.5
LR	83	83.6	78.8	71.7	94.5	92.2	93.4	91.4

[10] DurF0F123_50 refers to the parameter set of duration, mean F0, and F1-F3 sampled once at 50% of vowel duration whereas DurF0F123_2080 refers to the parameter set of duration, mean F0, and F1-F3 sampled twice at 20% and 80% of vowel duration.

Mean	87.55	89.13	83.65	81.88	93.98	94.68	93.08	93.7
*DurF0F123_50 refers to the parameter set of vowel duration, mean F0, and F1, F2 and F3 measures sampled once at 50% of vowel duration. **DurF0F123_2080 refers to the parameter set of vowel duration, mean F0, and F1, F2, and F3 measurements sampled twice at 20% and 80% of vowel duration.								

According to the results of the model fitting in Table 3, the model performance excelled with dynamic parameter sets. The models for the best correct classification performance are kNN (96.9%) and RF (96.8%) with DurF0F12_2080. Models with other dynamic parameter sets also performed far better than with static parameter sets. Hereafter, those models with dynamic parameter sets fitted to AE vowel signals are to be called "AE models." Dynamic spectral parameter sets contributed much to the model performance in all models. This means that AE talkers' production of AE vowels can be best modeled with dynamic spectral properties (e.g., AE models), regardless of the talkers' gender.

5.3 The AE models' prediction on K talkers' English vowel signals

The proposed AE models (with dynamic parameter sets) explained up to 97% of AE talkers' vowel signals. Now, we would like to know the correct identification performance of the AE models on 18 K male and female talkers' English vowel signals to see how many signals among K talkers' English vowel signals the AE models identified as intended AE vowel signals. All the dynamic parameter sets of cue measurements of K talkers' English vowel signals were fed into the proposed four AE models for identification accuracies.

Table 4. The identification accuracies of the four AE models (with dynamic parameter sets) on K talkers' and AE talkers' English vowel signals

K talkers' English vowel signals predicted	Dynamic parameter sets							
	DurF0F123_2080		DurF0F12_2080		F0F123_2080		F0F12_2080	
	K	AE	K	AE	K	AE	K	AE
kNN	50.42	93.8	54.08	96.9	50.37	92.2	52.21	94.9
RF	54.38	93	53.63	92.8	51.91	92.2	51.23	92

SVM	44.93	94.6	53.48	96.8	45.56	94.5	50.50	96.5
LR	46.86	94.5	49.17	92.2	43.58	93.4	43.52	91.4
Mean	49.15	93.98	52.59	94.68	47.85	93.08	49.37	93.7

The proposed AE models disappointingly identified 49-53% of K talkers' English vowel signals as intended vowels, when considering the AE models' 93-94% identification accuracies on AE talkers' vowel signals. Such poor performance may result partly from the fact that K talkers produced each type of English vowels with slightly different and inconsistent F1/F2/F3 values.

It is interesting to see what types of K talkers' English vowel signals were identified well or badly by the AE models. The kNN AE model with DurF0F123_2080 identified disappointing 15.00% of K talkers' /ʌ/ signals as intended (39.40% of /ʌ/ identified as /ɔ/), 22.80% of /ɪ/ as intended (77.20% of /ɪ/ identified as /i/), 24.40% of /æ/ as intended (57.20% of /æ/ as /ɛ/), 38.90% of /ɑ/ as intended (49.40% of /ɑ/ as /ɔ/), and 49.40% of /ʊ/ as intended (37.80% of /ʊ/ as /u/), whereas 92.20% of /i/ signals, 75% of /u/ and 73.30% of /ɛ/ were identified as intended vowels. This suggests that K talkers had difficulties producing English /ʌ, ɪ, æ, ɑ, ʊ/ correctly whereas they have relatively less difficulties producing /i, u, ɛ/. Most of K talkers' /ʌ, ɪ, æ, ɑ, ʊ/ were identified as /ɑ or ɔ, i, ɛ, ɔ, ʊ or u/, respectively, by the kNN model. K talkers had less difficulties correctly producing high tense vowels /i, u/ than lax /ɪ, ʊ/. These results were already observed in Flege et al. (1997) and Cho and Jeong (2013). However, K talkers had less difficulties producing lax /ɛ/ than tense /æ/, which does not agree with the observation in Cho and Jeong (2013) that K talkers had more difficulties producing correct /ɛ/ than /æ/.

Table 5. The confusion matrix by kNN AE models with DurF0F123_2080 on K talkers' English vowels

kNN (%)	ɑ	æ	ɛ	ɪ	i	ɔ	ʊ	u	ʌ	N
ɑ	38.90	0.00	0.00	0.00	0.00	49.40	0.00	8.90	2.80	180
æ	17.80	24.40	57.20	0.00	0.00	0.00	0.00	0.00	0.60	180
ɛ	6.70	17.20	**73.30**	2.20	0.00	0.00	0.00	0.00	0.60	180

ɪ	0.00	0.00	0.00	22.80	77.20	0.00	0.00	0.00	0.00	180
i	0.00	0.00	0.00	7.80	**92.20**	0.00	0.00	0.00	0.00	180
ɔ	16.70	0.00	0.00	0.00	0.00	62.80	0.00	20.00	0.60	180
ʊ	0.00	0.00	7.20	0.00	0.00	0.60	49.40	37.80	5.00	180
u	0.00	0.00	3.90	0.00	0.00	0.60	18.90	**75.00**	1.70	180
ʌ	31.10	0.00	0.00	0.00	0.00	39.40	0.00	14.40	15.00	180
Rows represent presented vowels whereas columns, identified vowels by the kNN AE model								Total	50.42	1620

In this subsection, it has been shown that the AE models identified only half of K talkers' English vowel signals as intended AE vowels. It was suggested that one of the major reasons for this is that K talkers used incorrect and inconsistent vowel spectral cue values in their English vowel production. In the next subsection, it will be shown that there is another major reason as to why half of K talker's English vowel signals were misidentified by the AE models.

5.4 Modeling and testing K talkers' English vowel production: K models

The following two figures of plots of F1/F2 measurements sampled at 50% of vowel duration of K male talkers' English vowel signals (Figure 7) and K female talkers' (Figure 8), show serious overlaps between different vowel types and a long and wide spread of signals of each vowel type.

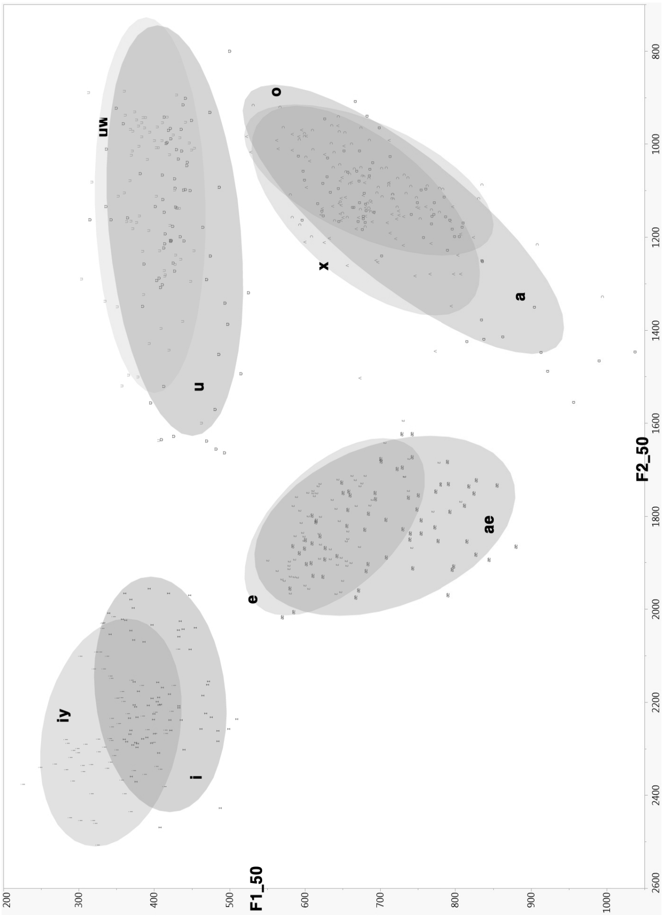

Figure 7. Plots of F1/F2 measurements of K male talkers' English vowel signals, sampled at 50% of vowel duration with ellipses fit to the vowel signals

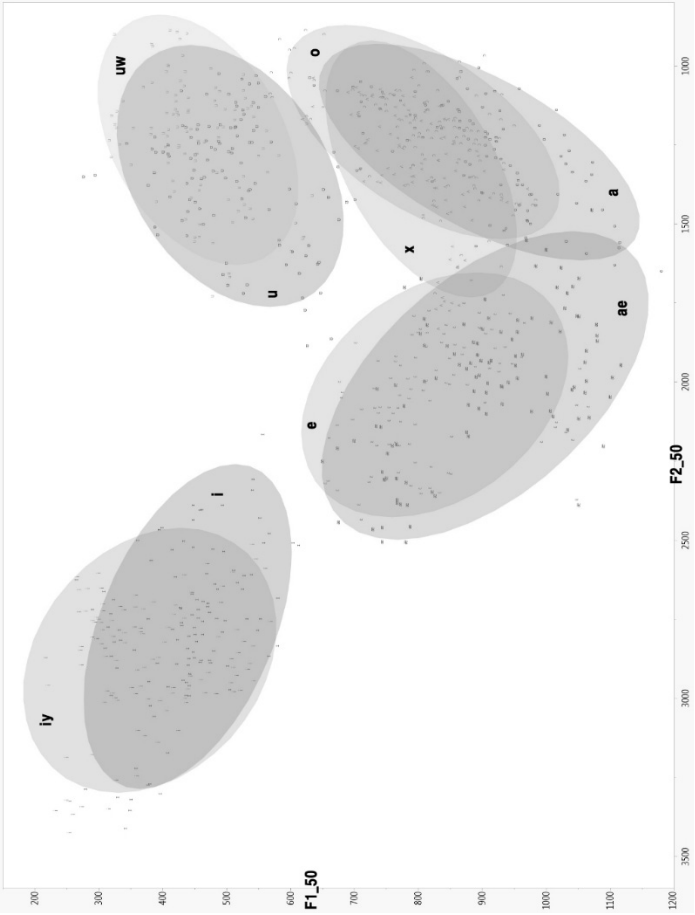

Figure 8. Plots of F1/F2 measurements of K female talkers' English vowel signals, sampled at 50% of vowel duration with ellipses fit to the vowel signals

We are going to show that AE models' poor identification accuracies on K signals were due to the fact that K talkers' English vowel signals cannot be well characterized by dynamic parameter sets unlike AE talkers' vowel signals. Figure 7 and 8 above suggest any pattern recognition model fitted only with static F1 and F2 measurements of K talkers' vowel signals might not properly identify K talkers' English vowel signals. However, we are going to propose that K talkers' English vowel signals can be best characterized by the pattern recognition models which were fitted with static parameter sets along with duration and F0. We are going to demonstrate that such models identified about 70% of K talkers' English vowel signals as intended vowels, which are far better than AE models' performance (about 50% identified as intended vowels).

We built the same four pattern recognition models for K talkers' vowel signals. This time, however, we set up the same models to be fitted to K talkers' vowel signals with dynamic or static parameter sets in order to verify whether K talkers produce English vowels with dynamic or static cue parameter sets. The model fitting processes used the same 10-fold cross-validation supervised learning mode.

It turned out, as shown in Table 6, that there was almost no model identification accuracy difference on the K talkers' English vowel signals between the models with static parameter sets and the ones with dynamic parameter sets. The K models with static DurF0F123_50 and DurF0F12_50 correctly identified 71.23% (mean) and 71% (mean) of K talkers' English vowel signals, respectively, while the models with dynamic DurF0F123_2080 and DurF0F12_2080, 72.75% (mean) and 72.9% (mean) of the same signals, respectively. The lack of the difference in identification accuracies between K models with static and dynamic parameter sets, indicates that the K models with static parameter sets turned out to be better models for K talkers' vowel production according to Occam's Razor, which says that when models show the same performance, the models with less parameters are better than the ones with more parameters. Therefore, the K models can be optimized best with static parameter sets. Hereafter, "K models" refer to only the K models with static parameter sets. Note for comparison that the AE models were optimized best with dynamic parameter sets.

Table 6. Identification accuracies on K talkers' English vowel signals by models fitted to K talkers' English vowel signals through 10-fold cross-validation

K English signals predicted	Static parameter sets				Dynamic parameter sets			
	DurF0F123_50	DurF0F12_50	F0F123_50	F0F12_50	DurF0F123_2080	DurF0F12_2080	F0F123_2080	F0F12_2080
kNN	69.8	72.7	66.6	66.6	70.2	70.2	66.2	69.3
RF	76.1	73.1	67.5	66.2	75.9	76.5	71.5	70.1
SVM	76	75.7	63.9	63.9	80	80	71.4	70.2
LR	63	62.5	52.8	52.8	64.9	64.9	58.6	57
Mean	71.23	71	62.7	62.38	72.75	72.9	66.93	66.65

5.5 Comparison of identification accuracies between the AE and the K models

Note that the K models (with static parameter sets) correctly identified about mean 71% of K talkers' English vowel signals. These identification accuracies of the K models on K talkers' English vowel signals, turned out to be far better than the identification accuracies of the AE models (with dynamic parameter sets) on the same signals (52.59% with DurF0F12_2080 and 49.15% with DurF0F123_2080). Note that AE models identified about mean 93% of AE talkers' vowel signals correctly.

Table 7. Identification accuracies predicted on K talkers' English vowel signals by models fitted to AE talkers' vowel signals and by models fitted to K talkers' English vowel signals

K signals predicted	AE Models fitted to AE talkers' English vowel signals with dynamic parameter sets				Fitted to K talkers' English vowel signals with static parameter sets			
	DurF0F123_2080	DurF0F12_2080	F0F123_2080	F0F12_2080	DurF0F123_50	DurF0F12_50	F0F123_50	F0F12_50
kNN	50.42	54.08	50.37	52.21	69.8	72.7	66.6	66.6
RF	54.38	53.63	51.91	51.23	76.1	73.1	67.5	66.2
SVM	44.93	53.48	45.56	50.50	76	75.7	63.9	63.9
LR	46.86	49.17	43.58	43.52	63	62.5	52.8	52.8
Mean	49.15	52.59	47.85	49.37	71.23	71	62.7	62.38

The K models' far higher identification accuracies on K talkers' English signals than the AE models' on the same signals, strongly suggest that K talkers' English vowel signals are better characterized by static spectral properties. This strongly suggests that K talkers produced English vowels with less spectral change than AE talkers. We suggest that this is another reason for poor identification accuracies by the AE models on K talkers' English vowel signals.

6. Conclusion

It has been shown through perception-based pattern recognition modeling of AE talkers' signals that AE talkers' vowel signals were best characterized by dynamic parameter sets. However, the AE models (with dynamic parameter sets) identified K talkers' English vowel signals as intended vowels very poorly, suggesting that K talkers' signals could not be well characterized by dynamic parameter sets. However, when K talkers' signals were directly fitted to by the same pattern recognition classification models, the models with static parameter sets (K models) identified K talkers' signals best. And the K models' identification performance on K talkers' English vowel signals were far better than AE models' on the same signals. This strongly suggests that K talkers produced English vowels with static spectral properties. As a result, we suggest that K talkers' English vowel signals do not sound like English vowels due to the lack of spectral properties with incorrect and inconsistent spectral values throughout vowel duration.

Before we conclude, we are going to compare the production-based AE models' identification performance on AE talkers' vowel signals in the present study with the perception-based AE models' on AE talkers' perception of vowel signals in Hong (2015). Note however, that though different vowel data were used in the two studies and hence, direct comparison may not be allowed, such model performance comparison offers some implication to us, though no further discussion will be offered here, since more detailed studies on this topic are pending.

Table 8. Identification accuracies on AE vowel signals identified by the four production-based AE models (with dynamic parameter sets) and by a perception-based AE model (with dynamic parameter sets) in Hong (2015)

AE talkers' vowel signals predicted	The present production-based AE models with dynamic parameter sets tested on AE talkers' vowel signals			The perception-based AE model with dynamic parameter sets tested on AE talkers' vowel identification in Hong (2015)		
	DurF0F123_2080	DurF0F12_2080	F0F12_2080	DurF0F123_2080	DurF0F12_2080	F0F12_2080
LR	94.5	92.2	91.4	91.84	92.72	90.33

Only LR AE model was built in Hong's (2015) study on AE listeners' vowel perception and we compare identification accuracies on AE signals tendered by production-based and perception-based LR. In Table 8, production-based and perception-based AE models do not show big difference in correct identification performance (production-based DurF0F12_2080: 92.2% vs. perception-based DurF0F12_2080: 92.72%; production-based DurF0F123_2080: 94.5% vs. perception-based DurF0F123_2080: 91.84%).

It is interesting to see that production-based K models' identification performance with static DurF0F12 (62.5%) turned out to be almost the same as perception-based K models' (62.73%), as shown in Table 9.

Table 9. Identification accuracies on K talkers' English vowel signals identified by the four production-based K models (with static parameter sets) and by a perception-based K model (with static parameter sets) in Hong (1995)

K talkers' English vowel signals predicted	The present production-based K models with static parameter sets tested on K talkers' English vowel signals		The perception-based K model with static parameter sets tested on listeners' English vowel identification in Hong (2015)	
	DurF0F12_50	F0F12_50	DurF0F12_ST*	F0F12_ST
LR	62.5	52.8	62.73	59.40
DurF0F12_ST = Measurements of duration, mean F0, and F1 and F2 sampled at steady state				

These results may weakly suggest that production may be closely related with perception in second language acquisition. However, we admit that such direct comparison between the two different types of models would not be warranted since K subjects are different in the two studies. In addition, more detailed studies on this question are pending.

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.

Demsar, Janez, Tomaz Curk, Ales Erjavec, Crt Gorup, Tomaz Hocevar, Mitar Milutinovic, Martin Mozina, Matija Polajnar, Marko Toplak, Anze Staric, Miha Stajdohar, Lan Umek, Lan Zagar, Jure Zbontar, Marinka Zitnik and Blaz Zupan. 2013. Orange: Data Mining Toolbox in

Python. *Journal of Machine Learning Research* 14, 2349–2353.

CHO, MI-HUI and SOONYONG JUNG. 2013. Perception and production of English vowels by Korean learners: A case study. *Studies in Phonetics, Phonology and Morphology* 19.1, 155-177. The Phonology-Morphology Circle of Korea.

CHUNG, HYUNJU, EUN JONG KONG and GARY WEISMER. 2010. Vowel formant trajectory patterns for shared vowels of American English and Korean. *Phonetics and Speech Sciences* 2.4, 67-74.

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.

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. and ROBERT T. GAYVERT. 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. 2012. The relative perceptual easiness between perceptually assimilated vowels for university-level Koran learners of American English and measurement bias in an identification test. *Studies in Phonetics, Phonology and Morphology* 18.3, 491-511. The Phonology-Morphology Circle of Korea.

_______________. 2013. Korean talkers' cue weighting perception strategies in perceiving English /ɑ/, /ɔ/ and /ʌ/ in comparison with American talkers'. *Studies in Phonetics, Phonology and Morphology* 19.3, 529-554. The Phonology-Morphology Circle of Korea.

_______________. 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.

_______________. 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. The Phonology-Morphology Circle of Korea.

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

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.

OH, EUNJIN. 2013. Dynamic spectral patterns of American English front monophthong vowels produced by Korean-English bilingual speakers and Korean later learners of English. *Studies in Phonetics, Phonology and Morphology* 30.2, 293-312. The Phonology-Morphology Circle of Korea.

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.

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

received: Octorber 31, 2016
revised: December 6, 2016
accepted: December 7, 2016