

The relative contributions of non-spectral cues to static and dynamic spectral model identification of Korean monophthong signals in Seoul Corpus*

Soonhyun Hong
(Inha University)

Hong, Soonhyun. 2020. The relative contributions of non-spectral cues to static and dynamic spectral model identification of Korean monophthong signals in Seoul Corpus. *Studies in Phonetics, Phonology, and Morphology* 26.1. 159-184. Though the importance of spectral characteristics at the steady-state central sections of Korean monophthongal signals in the hVd syllable has been amply reported in the literature, it has been rarely studied whether dynamic spectral measurements sampled multiply across the temporal dimension can better characterize Korean vowels in spontaneous speech than static spectral measurements at a (steady-state) central section. Furthermore, the perceptual influence of non-spectral cues on the spectral properties of vowels in vowel perception has been frequently reported in the literature, but few reports have been released on the relative amount of the individual perceptual contributions of non-spectral cues (e.g., gender, speaking rate, duration, F0, place and manner of the flanking phones, etc.) on the spectral properties of vowels in vowel perception. Neural Network pattern recognition modeling on spectral identification of Korean monophthong signals in Seoul Corpus showed that dynamic spectral models fitted to non-spectral cues, identified vowel signals better than static spectral models. Furthermore, flanking phone identities, and manner and place of flanking phones (i.e., coarticulation information) were the most contributive to spectral vowel identification. However, F0, speaking rate, duration, gender, and speaker's age showed little or almost no contribution. **(Inha University, Professor)**

Key words: pattern recognition modeling of Korean vowels, perceptual influence of cues, coarticulation effects

1. Introduction

Peterson and Barney (1952) pointed out between-speakers spectral variation of

* This work was supported by INHA UNIVERSITY Research Grant.
Many thanks should go to Jongpyo Hong (CDS) for Python coding.

English vowels. They showed that the plots of F1/F2 frequency measurements sampled at the steady-state central section of English vowel signals in the hVd syllable, produced by males, females and children, are scattered wide-spread on the acoustic vowel space, and some vowel signals of different types were overlapped with each other. Despite such between-speakers acoustic variations, American English (AE) listeners perceived them as intended vowels 95% correct (Hillenbrand et al. 1995). There is a wide gap between acoustically variable characteristics and perceptually rather constant characteristics of vowels. Such between-speakers spectral variation of vowels has become a major research topic to explain listeners' perception of vowels.

One approach to resolve the gap is the speaker normalization approach, which says that the static acoustic values of F1 and F2 (and also F3) at a steady-state central section of vowel duration are perceptually modulated by other acoustic or non-acoustic properties of vowels such as F0, gender, prosodic position of the vowel, manner and place of the preceding or following consonant, or speaking rates (Johnson 2008). Namely, perceptual values of acoustic formant frequencies may vary as a function of acoustic or non-acoustic cues such as the flanking phone identities, manner and place of the flanking phones, gender, prosody, duration, speaking rate, F0, etc.

Another approach to fill the gap is the Vowel Inherent Spectral Change (VISC) approach (Lindblom and Studdert-Kennedy 1967, Nearey and Assmann 1986, Strange et al. 1983, Hillenbrand et al. 1995, Assmann and Morrison 2013, Hillenbrand 2013), which tries to find a solution in the dynamic spectral characteristics of the vowels across the vowel length. It says that even AE monophthongs show dynamic spectral movement across the vowel length like diphthongs, and spectral measurements at multiple sections of the vowel duration are enough to explain listeners' perception, demonstrating that the perceptual influence of non-spectral cues such as duration and F0, is relatively minimal (Hillenbrand et al. 1995). Hillenbrand et al. reported that a discriminant pattern recognition model with one-sampled F1-3 at steady state identified the vowel signals in the hVd syllable 81% correct whereas the one with F1-3 frequencies sampled at 20% and 80% of vowel duration, identified the signals 91.6% correct. As 20 AE listeners' correct identification rate was 95.4%, the identification accuracy enhancement from 81% for a static model to 91.6% for a dynamic model indicated that the identification performance of the dynamic spectral model almost reached the level of listeners'

vowel perception level. These results suggested that between-speakers variation for the vowel signals in the hVd syllable may be better explained perceptually by VISC than the speaker normalization approach.

However, the previous VISC studies suffered from problems despite high correct vowel identification accuracies. Their studies were limited to the vowel signals in the hVd syllable. It has not been shown how VISC handles vowel signals in spontaneous speech, where neighboring phones and other non-spectral cues like F0 and gender may affect spectral model identification of vowels.

The first objective of the present study is to verify the validity of the VISC approach. The target vowels of the present study are Korean monophthong signals in spontaneous speech in Seoul Corpus (Yun et al. 2015). For pattern recognition model classification, a Neural Network (NN) model with a multi-layer perceptron (MLP) algorithm with backpropagation in Orange 3.21¹ (Demsar et al. 2013, Rumelhart et al. 1986, Bottou 2010, Ng et al. 2011, LeCun et al. 1996, Kingma and Ba 2014, and Hong 2019 and Yoon 2019 for Neural Network (NN) modeling for Korean vowel categorization) is to be trained in a supervised mode and tested with seven different sets of dynamic or static spectral properties². Namely, a NN algorithm is to be fitted to F1-3 frequencies sampled at (1) 20% of the vowel duration (F123_20), (2) 50% (F123_50), (3) 80% (F123_80), (4) at 20% and 50% (F123_2050), (5) at 50% and 80% (F123_5080), (6) at 20% and 80% (F123_2080), and (7) at 20%, 50%, and 80% (F123_205080) to produce three simple static spectral models and four simple dynamic spectral models. The results of the seven NN spectral models should verify whether simple spectral models³ with spectral parameters sampled at multiple sections of the vowel duration show better classification accuracies than those

¹ In the present study, used was the NN widget which uses sklearn Multi-layer Perception algorithm that can learn non-linear models as well as linear: 100 Neurons per hidden layer, the rectified linear unit function for activation, stochastic gradient-based optimizer Adam as a solver for weight optimization, 0.00010 for L2 penalty (regularization term) parameter Alpha, and 200 maximum number of iterations (Hong 2019). For more information, refer to <https://docs.biolab.si//3/visual-programming/widgets/model/neuralnetwork.html>.

² Through stratified random sampling, the model is to be trained with two repetitions on 66% of the data and tested with two repetitions on the rest of the data.

³ These models fitted only to spectral properties are “simple” models in contrast with those “full” models fitted to the non-spectral properties affecting vowel identification as well as spectral properties of vowels.

sampled singly at one section.

The second objective of the present study is to verify different contributions of non-spectral cues to spectral identification of vowels when the seven simple spectral models are further fitted to non-spectral cue parameters. More specifically, the seven simple spectral models are to be further fitted to the measurements or nominals of the major non-spectral cues reported to affect spectral vowel identification in the literature. In the present study, the contribution of each of the non-spectral cues is assumed to be the difference between the identification accuracies before and after the removal of a cue parameter from the “full” spectral model fitted to all non-spectral cue parameters. Different amount of contributions of non-spectral cues such as F0 or speakers’ gender, if any, may constitute evidence for or counter evidence against the speaker normalization approach.

The non-spectral cue parameters which NN spectral models are to be fitted to, are the acoustic or non-acoustic cues which have been reported to affect vowel perception in the literature (Wright et al. 1997, Johnson 2008, and references therein).

2. The perceptual influence of non-spectral cues on spectral properties of vowels in the literature

The perceptual influence of non-spectral cues on vowel perception in addition to spectral properties has been amply reported in the literature. The difference in males’ and females’ acoustic vowel chart of F1 and F2 measurements, may be due to F0 difference between different genders. First, it was suggested in the literature that spectral properties of vowels may be modulated by F0. Miller (1953) observed that doubled F0 frequency resulted in the shifts of vowel category boundaries for most of the English vowels. Fujisaki and Kawashima (1968) and Slawson (1968) also reported that abrupt change of F0 resulted in perceptual change of F1 or F2. Furthermore, listeners experienced perceptual difficulties when they listened to children’s vowel signals with F0 replaced by male F0 (Lehiste and Metzger 1973). The vowel category boundary shifts were observed when listeners listened to vowel signals with an unpredictably abrupt F0 change (Johnson 1990).

Second, males’ and females’ vowel formant frequencies were normally realized quite similar at low formant frequencies with each other but quite differently at higher formant frequencies (Fant 1966). Johnson (2008) noted that gender was a

perceptually important cue due to anatomical differences between males and females in vocal tract size and shape (Fant 1966, Nordström 1977, Goldstein 1980, Traunmüller 1984) and also due to different speech styles or speech production patterns (Henton 1992, Byrd 1984). When speakers' gender was misidentified, the error rate in vowel identification went up to 25% from 5%.

Vowel duration may constitute an additional cue in vowel perception. Lehiste and Metzger (1973) showed that listeners' vowel identification accuracies deteriorated abruptly when they identified those vowels of fixed length with steady-state formant frequencies synthesized. Lehiste and Peterson (1961) reported that tense vowels were realized longer than lax vowels and low vowels were longer than high vowels.

Formants of vowels are rather unstable along the temporal dimension due to neighboring phones, especially flanking consonants. Neighboring phones directly affect the shape of a formant pattern, which results in a dynamic spectral change along the temporal dimension especially in casual or spontaneous speech. For this reason, the affected vowels sometimes do not hit the target formant frequencies which are observed in careful speech. Vowel undershoot (Fant 1960, Stevens and House 1963) was often observed at the beginning and the end of vowels due to the direct influence of the flanking consonants or of manner and/or place of the flanking consonants. Stevens and House (1963) and later by Hillenbrand et al. (2000) reported that average formant frequencies of English vowels /i, ɪ, ε, æ, ɑ ɒ u, ʌ/ changed as a function of the symmetrical flanking consonants' places (labial, post-dental, and velar).

The speaking rates on spectral change of vowels directly influence vowel undershoot and their influence on vowel perception has been extensively studied (Gay 1978, Engstrand 1988, Van Son and Pols 1992, Hirata and Tsukada 2004). Spectral undershoot may be due to flanking consonants, speaking rate, prosody, sentence structure, dialect, and individual speaking style (Lindblom 1963, Gay 1978). In addition, vowel duration and speaking style are variables in English stressed vowels that the extent of formant frequency change depends on (Moon and Lindblom 1994).

In the present study, the values of the following non-spectral cues which may affect vowel perception, were extracted from Korean Corpus of Spontaneous Speech (Yun et al. 2015) to see if they affect spectral identification of vowel signals by NN pattern recognition models (Hong 2019):

Table 1. The extracted thirteen non-spectral cues from Seoul Corpus which seven NN models are to be fitted to for vowel classification

Parameters	spkrSex	spkrAge	spRate	F0_ave	durTargetV
Description	Gender (female, male)	Talkers' exact age (from teens to forties)	Number of syllables spoken per second across three words with the target vowel in the 2 nd word	Average F0 across the vowel duration	Vowel duration
Parameters	locSylInWord	locSylInUtt	prevPho	nextPho	
Description	Syllable location within the word (word-initial, - medial, and -final syllable, and mono- syllable)	Syllable location within the utterance (utt-initial, -medial, and -final syllable, and mono-syllabic utt)	Identity of the preceding phone	Identity of the preceding phone	
Parameters	prevPhoPlace	prevPhoMan	nextPhoPlace	nextPhoMan	
Description	Place features of the previous phone (bilab, alveo, velar, palat, vocoid, others)	Manner features of the previous phone (stop, fric, affr, liquid, nas, vocoid, others)	Place features of the following phone (bilab, alveo, velar, palat, vocoid, others)	Manner features of the following phone (stop, fric, affr, liquid, nas, vocoid, others)	

3. Method

3.1 Subjects

The Korean Corpus of Spontaneous Speech (Seoul Corpus) (Yun et al. 2015) consisted of spontaneous speech signals produced by four groups of Seoul Korean talkers in their teens, twenties, thirties and forties. Each talker group consists of five males and five females. The forty-hours-long speech signals were recorded at 44kHz with 16-bit quantization with one hour for each talker, and then were segmented and labeled with Praat (Boersma and Weenink 2018).

3.2 Materials

3.2.1 Korean target monophthong signals

498,204 monophthong signals of seven vowel types in Seoul Corpus were the target signals under the modeling analysis, as shown in Table 2.

Table 2. The number of target monophthong signals of seven vowel types

IPA	Symbols in Seoul Corpus	Korean alphabet	No. of signals
i	ii	ㅣ	77,594
e	ee	ㅔ	73,217
a	aa	ㅏ	115,082
ɪ	xx	ㅡ	80,571
ə	vv	ㅓ	72,132
u	uu	ㅜ	36,009
o	oo	ㅛ	43,599
		total	498,204

3.2.2 Simple spectral models

The present study approached to the perceptual characteristics of vowels two ways. First, static and dynamic simple NN pattern recognition spectral models were built to check how well they categorized vowel signals. Simple static spectral models were composed of three simple models fitted to F1-3 measurements sampled at 20% (F123_20), 50% (F123_50), and 80% (F123_80) of the vowel duration. On the other hand, simple dynamic spectral models were four simple models to be fitted to measurements sampled at 20% and 50% (F123_2050), at 50% and 80% (F123_5080), at 20% and 80% (F123_2080), and finally at 20%, 50%, and 80% (F123_205080) of the vowel duration.

Table 3. Simple static and dynamic NN spectral models

Simple NN spectral models	F123_20	F123_50	F123_80	F123_2050	F123_5080	F123_2080	F123_205080
Static/ Dynamic	Static	Static	Static	Dynamic	Dynamic	Dynamic	Dynamic
Description	F1-3 sampled at 20% of vowel duration	F1-3 sampled at 50% of vowel duration	F1-3 sampled at 80% of vowel duration	F1-3 sampled at 20% and 50% of vowel duration	F1-3 sampled at 50% and 80% of vowel duration	F1-3 sampled at 20% and 80% of vowel duration	F1-3 sampled at 20%, 50%, and 80% of vowel duration

The next step was to feed in non-spectral parameters to these static and dynamic

spectral models to build up full spectral models to see how non-spectral parameters exerted perceptual influence on the static and dynamic spectral models based on model identification performance results.

3.2.3 Full spectral models to be fitted to non-spectral cue parameters

We began with checking correct identification performance of the seven simple dynamic and static simple spectral models. And then non-spectral cues which may help enhance model identification performance, constituted additional parameters which the seven simple spectral models were further fitted to for vowel classification performance. These models which are fitted to all parameters, will be called full spectral models in contrast with simple spectral models with no parameters fed in.

Then a specific parameter was to be removed from the full spectral model, which in turn was to be fitted to all the other parameters for identification performance. This procedure enabled us to see the perceptual contribution of the removed parameter. It was assumed that the degraded model identification performance difference between before and after the removal may be the perceptual contribution of the removed cue to the full spectral model.

From the target monophthong signals, extracted were F1-3 measurements at 20%, 50% and 80% of the vowel duration, and measurements or nominal values of the non-spectral parameters shown in Table 4 (refer to Table 1 for details of parameters), using a Praat script (Boersma and Weenink 2018).

Table 4. Seven simple NN spectral models to be fitted to two groups of non-spectral cue parameters for spectral identification of vowels, resulting in 14 full spectral models (14 full models: 7 simple spectral models fitted to 2 non-spectral parameter groups)

Seven NN simple spectral models	F123_20, F123_50, F123_80, F123_2050, F123_5080, F123_2080, and F123_205080	
	↑	↑
Non-spectral parameter groups	Group P (Phone identity group)	Group MP (Manner and place group)
Non-spectral cue parameters	Gender (spkrSex)	spkrSex
	Age (spkrAge)	spkrAge
	Speaking rate (spRate)	spRate

	Average F0 (F0_ave)	F0_ave
	Duration (durTargetV)	durTargetV
	Syllable location in word (locSylInWord)	locSylInWord
	Syllable location in utterance (locSylInUtt)	locSylInUtt
	Previous phone identity (prePho)	Manner of previous phone (prePhoMan)
	Next phone identity (nextPho)	Place of previous phone (prePhoPlace) Manner of next phone (nextPhoMan) Place of next phone (nextPhoPlace)

Notice that each of the seven simple spectral models was to be fitted to two non-spectral parameter groups, resulting in 14 full spectral models (seven simple spectral models were to be fitted to 2 different parameter groups (Group P and MP): 7 models*2 parameter groups=14 full models). Group MP included manners and places of the flanking phones whereas Group P the flanking phone identities, though the other parameters remained the same. Therefore, after seven simple spectral models were fitted to Group P and MP, identification accuracies of 14 full spectral models were attained.

Table 5. Notations to be used in the present study

Notations	Explanation
F123_50	Simple spectral model fitted to F1-3 sampled at 50% of the vowel duration
F123_50P++	Full F123_50 fitted to all parameters in Group P
F123_50P++spRate-	Full F123_50P++ with spRate removed
F123_50P++F0-	Full F123_50P++ with F0 removed
F123_50P++prePho-	Full F123_50P++ with prePho removed
F123_50P++nextPho-	Full F123_50P++ with nextPho removed
F123_50P++locSylInWord-	Full F123_50P++ with locSylInWord removed
F123_50P++locSylInUtt-	Full F123_50P++ with locSylInUtt removed
F123_50MP++	Full F123_50 fitted to all parameters in Group MP
F123_50MP++spRate-	Full F123_50MP++ with spRate removed
F123_50MP++prePhoPlace-	Full F123_50MP++ with prePhoPlace removed
F123_50MP++nextPhoPlace-	Full F123_50MP++ with nextPhoPlace removed
F123_50MP++prePhoMan-	Full F123_50MP++ with prePhoManner removed
F123_50MP++nextPhoManPlace-	Full F123_50MP++ with nextPhoManPlace removed
F123_50MP++prePhoManPlace-	Full F123_50MP++ with prePhoManPlace removed

Given the 14 full spectral models which were fitted to either Group P or MP parameters for vowel identification, the perceptual contribution of a given parameter was computed by removing the parameter from the parameter groups, in order to see how much the model identification performance was degraded. The performance degrade after the parameter removal was assumed to be the perceptual contribution of the removed parameter to the model's spectral identification of vowels.

For example, F123_50MP++ was a full spectral model with spectral samples at 50% of the vowel duration which was fitted to all parameters of Group MP. Now, one parameter, for example, prePhoMan (manner of the preceding phone) was removed from Group MP (e.g., F123_50MP++prePhoMan-), and the resulting model classification accuracy went down. The performance degrade resulting from the removal of prePhoMan ($= (\text{accuracy of "F123_50MP++"} - (\text{accuracy of "F123_50MP++prePhoMan-"}))$), was assumed to be the perceptual contribution of prePhoMan to the full model F123_50MP++. This way, each of the non-spectral parameters from Group P and Group MP was removed from the 14 full spectral models and compared were the identification accuracies before and after the removal. Additionally, also considered was the perceptual contributions of combinations of place and manner of the preceding or the following phone (i.e., prePhoManPlace and nextPhoManPlace) to be compared to those of the preceding or following phone identities (i.e., prePhone and nextPhone).

4. Results

4.1 Spectral characteristics of the Korean monophthong signals

Average F1 and F2 frequencies of all genders' signals of each vowel type sampled at 20%, 50% and 80% of the vowel duration are shown in Table 6.

Table 6. The average F1 and F2 frequencies sampled at 20%, 50%, and 80% of the vowel duration for each monophthong type⁴

Vowel	F1_20 ave	F2_20 ave	F1_50 ave	F2_50 ave	F1_80 ave	F2_80 ave
i (/ii/)	387	2126	381	2140	387	2126
e (/ee/)	404	1968	410	1971	391	1904
a (/aa/)	568	1428	575	1437	503	1460
i (/xx/)	410	1579	411	1580	395	1585
ə (/vv/)	402	1305	392	1297	376	1368
u (/uu/)	414	1211	415	1182	414	1211
o (/oo/)	363	1058	362	1043	355	1144

The average F1 and F2 frequencies of each of the vowels were plotted on the acoustic space as shown in Figure 1.

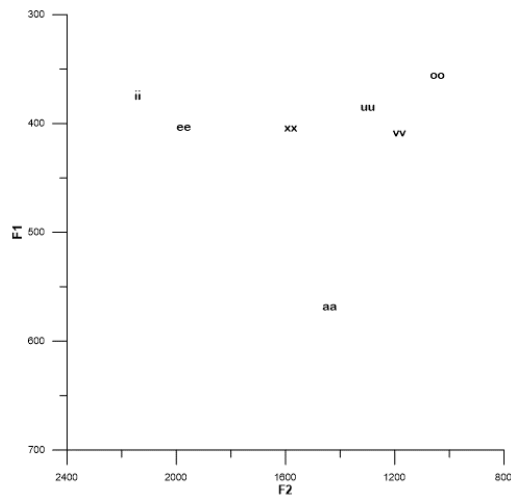


Figure 1. Acoustic vowel chart with average F1 and F2 frequencies of all genders' signals of each vowel type sampled at 50% of the vowel duration (Hong 2019)

⁴ "F1_20ave" refers to average F1 frequency sampled at 20% of the vowel duration, and "F2_20ave" to average F2 frequency sampled at 20% of the vowel duration: For example, $F1_20ave = (Male_F1_20 \text{ average} + Female_F1_20 \text{ average}) / 2$.

The acoustic vowel chart in Figure shows the plots of average F1/F2 values of signals of each vowel type produced by both males and females. From the perceptual point of view, however, the F1/F2 plots of vowels do not properly represent acoustic characteristics of vowels, since spectral properties are realized quite differently across different genders, as shown in Figure 2 below.

Table 7 shows the average F1 and F2 frequencies of vowel signals of males' and females' measured at 20%, 50% and 80% of the vowel duration and Figure 2 illustrates how males' and females' vowels are represented on the acoustic vowel chart.

Table 7. The average F1 and F2 frequencies for males and females sampled at 20%, 50%, and 80% of vowel duration for each monophthong type

Gender	Vowel	F1_20 ave	F2_20 ave	F1_50 ave	F2_50 ave	F1_80 ave	F2_80 Ave
M	i (/ii/)	380	1960	368	1972	380	1960
	e (/ee/)	371	1804	378	1805	369	1747
	a (/aa/)	509	1337	513	1343	458	1364
	i (/xx/)	391	1468	388	1465	375	1471
	ə (/vv/)	401	1267	384	1247	369	1311
	u (/uu/)	387	1158	387	1132	387	1158
	o (/oo/)	346	1005	345	994	347	1108
F	i (/ii/)	394	2292	395	2308	394	2292
	e (/ee/)	437	2133	442	2138	413	2062
	a (/aa/)	627	1520	637	1531	547	1556
	i (/xx/)	429	1690	433	1694	415	1698
	ə (/vv/)	403	1344	399	1348	383	1425
	u (/uu/)	442	1263	443	1232	442	1263
	o (/oo/)	379	1112	379	1092	364	1181

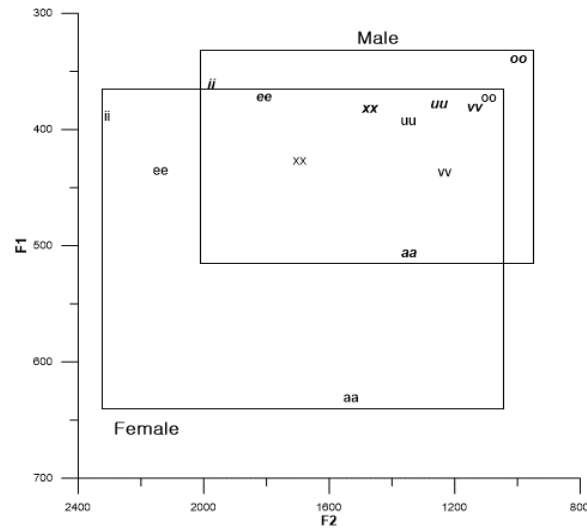


Figure 2. Males' and females' acoustic static vowel charts with average F1 and F2 frequencies of signals of each vowel type sampled at 50% of the vowel duration (italicized symbols for males' vowels)

Despite the fact that spectral properties of males' vowels are acoustically quite different from those of females', listeners were never influenced by these acoustic gaps in vowel perception. In addition, spectral properties of all monophthong signals were quite unstable along the temporal domain. Figure 3 shows that average F1 and F2 frequencies of all genders' signals of each vowel type at 20% and 80% of the vowel duration were plotted as vowel vectors to see temporal vowel quality change.

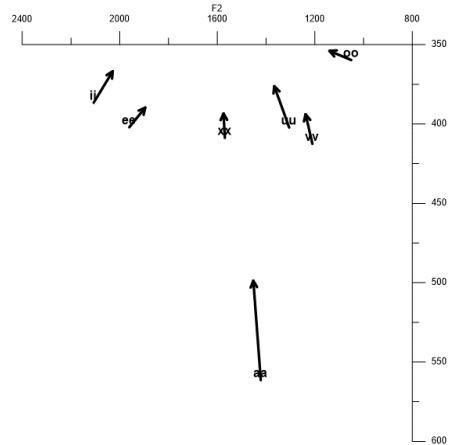


Figure 3. Acoustic vowel vector chart with average F1 and F2 frequencies of all genders' signals of each vowel type sampled at 20% and 80% of the vowel duration (The vector arrow shows the temporal vowel quality change) (Hong 2019)

When compared to the static vowel chart of all genders in Figure 3, the spectral properties of all monophthong vowels never remained stable across the vowel duration. The spectral changes across the vowel duration cannot be captured by the static spectral model. Then the question is whether listeners pick up such spectral changes when they perceive vowels. For this purpose, simple static and dynamic spectral models were built to verify which models can categorize vowel signals better.

4.2 Simple static or dynamic spectral model classification performance

Among all 7 simple static or dynamic spectral models (with no non-spectral parameters involved), dynamic spectral models identified vowel signals as intended vowels far better than static ones. F123_205080 (56% correct) performs the best, being followed by F123_2050 (54.9%) and F123_5080 (53.9%). F123_2080 (52.6%) performs better than the static F123_20 (50.8%), F123_50 (51.4%) and F123_80 (42.8%), as shown in Figure 4.

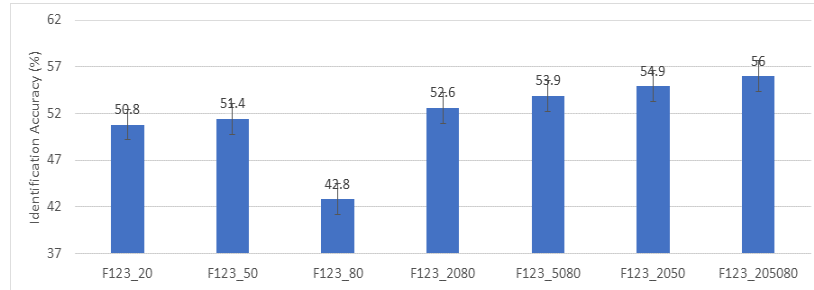


Figure 4. Classification accuracies of dynamic or static spectral models

The classification accuracies ranged from 56% for dynamic F123_205080 to 42.8% for static F123_80⁵. It seems that the more spectral properties in the temporal domain were referred to by simple spectral models, the better accuracies they turned out. Namely, Korean vowels in spontaneous speech may be better identified with spectral properties sampled at multi-sections of vowel duration than with spectral properties at one section of vowel duration.

Hillenbrand et al. (1995) demonstrated, using a discriminant classifier for identification of English vowel signals in the hVd syllable, that dynamic F123_2080 and F123_205080 models equally performed quite well to the listeners' perception level. Hillenbrand et al. (1995) further demonstrated that spectral properties at more than 3 sections of the vowel duration did not necessarily enhance vowel identification performance for vowel identification in the hVd syllable in AE.

However, the accuracies of all seven simple spectral models for Korean vowel identification were rather disappointing, showing that purely acoustic spectral properties, dynamic or static, alone cannot explain the perceptual characteristics of Korean monophthongs in spontaneous speech, requiring more cues for better vowel identification.

⁵ The poor performance of F123_80 will be handled in a separate study (in preparation), since some complication is involved in the spectral characteristics of Korean vowel signals in spontaneous speech.

5. Discussion

5.1 Classification performance of the full spectral models fitted to all the parameters in Group P or MP

When Group P and Group MP parameters were fitted to by the 7 simple spectral models, the resulting 14 full models' classification accuracies enhanced drastically, as illustrated in Figure 5.

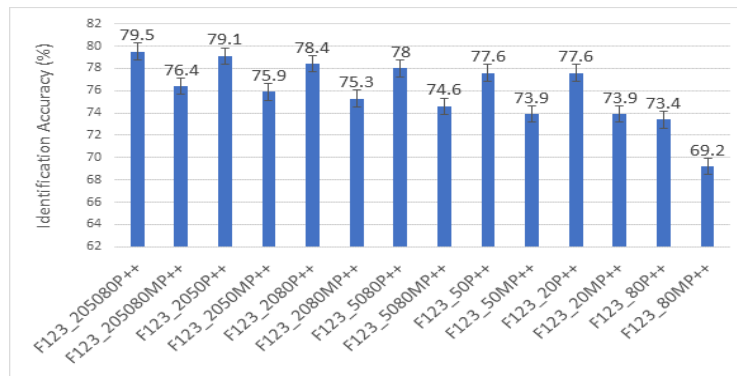


Figure 5. Classification accuracies of the 14 full spectral models fitted to the parameters in Group P or Group MP (“F123_205080P++” refers to “F123_205080” fitted to the parameters of Group P while “F123_205080MP++” to “F123_205080” fitted to the parameters of Group MP)

The full models fitted to the parameters of Group P performed far better than the corresponding full models fitted to those of Group MP. This indicates that the flanking phone identities (prePho and nextPho) were perceptually more influential on spectral vowel identification than manners and places of flanking phones (prePhoMan, prePhoPlace, nextPhoMan and nextPhoPlace) or the combined place and manner of flanking phones (“prePhoMan + prePhoPlace” and “nextPhoMan + nextPhoPlace”).

Among the seven full models fitted to Group P, dynamic full models (F123_205080P++ (79.5%), followed by F123_2050P++ (79.1%) and F123_2080P++ (78.4%)) showed far better performance than static full models (F123_50P++ (77.6%) and F123_20P++ (77.6%), followed by F123_80P++

(73.4%)).

These results said that vowels were identified the best when spectral properties were referred to simultaneously at three positions at 20%, 50%, and 80% of the vowel duration than two positions at 20% and 50%, at 50% and 80%, or at 20% and 80%, or one position at 20%, 50%, or 80%. Namely, dynamic models performed better than static models.

Researchers tried to resolve the gap between acoustic and perceptual characteristics of vowels through speaker normalization based on static spectral properties. Johnson 2008 and references therein proposed that the steady-state central acoustic values of F1, F2 and F3 were perceptually modulated by other acoustic or non-spectral cues like F0 or gender. On the other hand, the VISC approach demonstrated that dynamic spectral properties alone explained between-speakers variation based on vowel signals in the hVd syllable better than the speaker normalization approach (Hillenbrand 2013 and references therein).

However, the present modeling results showed that VISC alone could not explain perceptual characteristics of spectral variation of vowels, as simple spectral models (the best performance: 56% for F123_205080) suffered from disappointingly poorer identification accuracies than full models (the best performance: 79.5% for F123_205080P++ among the full models fitted to Group P and 76.4% for F123_205080MP++ among the full models fitted to Group MP).

The present modeling results targeting on identification of the vowel signals in spontaneous speech suggested that not just speaker normalization through perceptual modulation of non-spectral cues but the direct perceptual influence of non-spectral cues on spectral properties of vowels should be considered for the identification of vowels in spontaneous speech coupled with VISC. This further suggested that when listeners perceive vowels, they refer to non-spectral cues as well as dynamic spectral properties along the temporal dimension.

Since a lot of non-spectral cues may affect spectral vowel identification, the relative perceptual contributions of individual non-spectral cues to model classification may vary. In the next two subsections, the relative individual perceptual contributions of non-spectral cues to spectral vowel identification is to be studied with full spectral models fitted to the cue parameters in Group P in 5.2 and with full spectral models fitted to the cue parameters in Group MP in 5.3.

5.2 The perceptual contributions of non-spectral cues in Group P to full spectral models' vowel identification

It was assumed that the perceptual contribution of a parameter to a spectral model's identification of vowels, was the model's identification performance difference between before and after the removal of a target parameter from the parameter group (Group P in this subsection) which the full spectral model was fitted to. Table 8 illustrates the relative perceptual contributions of non-spectral parameters in Group P to spectral identification of vowels across the seven dynamic and static full spectral models

Table 8. Relative perceptual contributions of non-spectral cues (%) in Group P on static and dynamic full spectral models in vowel classification

Rank	F123_50P++ (77.6)	F123_20P++ (77.6)	F123_80P++ (73.4)	F123_2050P++ (79.1)	F123_5080P++ (78)	F123_2080P++ (78.4)	F123_205080P++ (79.5)
1	prePho-(12.6)*	prePho-(13.7)	prePho-(13.6)	prePho-(11.6)	prePho-(11.9)	prePho-(11.8)	prePho-(11.6)
2	nextPho-(8)	nextPho-(7.1)	nextPho-(10.4)	nextPho-(6.7)	nextPho-(7)	nextPho-(6.6)	nextPho-(6.5)
3	locSylInWord-(2.9)	locSylInWord-(2.9)	locSylInWord-(3.3)	locSylInWord-(2.8)	locSylInWord-(2.8)	locSylInWord-(2.6)	locSylInWord-(2.7)
4	dur-(0.6)	dur-(1)	dur-(1)	gender-(0.5)	dur-(0.7)	dur-(0.4)	dur-(0.8)
5	gender-(0.5)	gender-(0.6)	F0-(0.4)	dur-(0.5)	gender-(0.6)	gender-(0.3)	gender-(0.7)
6	spkAge-(0.3)	spkAge-(0.4)	gender-(0.3)	F0-(0.4)	spkAge-(0.3)	F0-(0.2)	spkAge-(0.5)
7	F0-(0.3)	F0-(0.3)	spkAge-(0.2)	spkAge-(0.3)	F0-(0.2)	spkAge-(0.1)	F0-(0.5)
8	spRate-(0.1)	spRate-(0.2)	spRate-(0)	spRate-(0.1)	spRate-(0.1)	spRate-(0)	spRate-(0.2)
9	locSylInUtt-(0.1)	locSylInUtt-(0.2)	locSylInUtt-(0)	locSylInUtt-(0)	locSylInUtt-(0)	locSylInUtt-(0)	locSylInUtt-(0.2)

*“prePho-(12.6)” refers to the perceptual influence of prePho on the full model F123_50P++. It means that if prePho was removed from Group P which F123_50P++ was fitted to, the identification accuracy went down by 12.6%.

The preceding phone identity (prePho) topped Group P parameters in its perceptual contribution to all seven full spectral models (accounting for 11.6~13.7% of vowel identification accuracies across all full spectral models), followed by the next phone identity (nextPho: 6.5~10.4%). This indicates that the preceding phone identity was perceptually more contributive than the following phone identity.

The perceptual contribution of locSyllInWord was the next in perceptual contribution, accounting for 2.6~3.3%, which were far lower than contributions of prePho and nextPho. However, locSyllInUtt accounted for meager 0.2%. All the other parameters such as duration, gender, spkAge, F0, and spkRate, showed extremely low perceptual contributions less than or equal to 1%, which was quite unexpected, constituting strong counter evidence to previous studies which said that these parameters were essential elements for speaker normalization.

5.3 The perceptual contributions of non-spectral cues in Group MP to full spectral models' vowel identification

The perceptual contributions of individual non-spectral parameters in Group MP to which the seven full spectral models were fitted to, are shown in Table 9.

Table 9. Relative perceptual contributions (%) of non-spectral cues in Group MP to static and dynamic full spectral models in vowel classification

Rank	F123_50M P++ (73.9)	F123_20M P++ (73.9)	F123_80M P++ (69.2)	F123_2050 MP++ (75.9)	F123_5080 MP++ (75.6)	F123_2080 MP++ (75.3)	F123_2050 80MP++ (76.4)
1	prePhoMan Place- *(10.3)	prePhoMan Place- (11.4)	prePhoMan Place- (11.7)	prePhoMan Place-(9.7)	prePhoMan Place-(9.7)	prePhoMan Place-(9.9)	prePhoMan Place-(9.6)
2	nextPhoMa nPlace- *(6.3)	nextPhoMa nPlace- (5.8)	nextPhoMa nPlace- (8.4)	nextPhoMa nPlace- (5.1)	nextPhoMa nPlace- (5.4)	nextPhoMa nPlace- (5.2)	nextPhoMa nPlace- (6.3)
3	prePhoMan -(4.1)	prePhoMan -(4.2)	prePhoMan -(4.9)	prePhoMan -(3.7)	prePhoMan -(3.9)	prePhoMan -(3.7)	prePhoMan -(3.8)
4	prePhoPlac e-(2.7)	prePhoPlac e-(3.1)	nextPhoMa n-(3.5)	prePhoPlac e-(2.4)	prePhoPlac e-(2.6)	prePhoPlac e-(2.5)	prePhoPlac e-(2.6)
5	nextPhoMa n-(2.7)	locSyllInW ord-(2.9)	prePhoPlac e-(3.3)	locSyllInWo rd-(2.3)	locSyllInW ord-(2.4)	locSyllInW ord-(2.5)	nextPhoMa n-(2.5)
6	locSyllInWo rd-(2.7)	nextPhoMa n-(2.8)	locSyllInW ord-(3.3)	nextPhoMa n-(2.2)	nextPhoMa n-(2.3)	nextPhoMa n-(2.2)	locSyllInW ord-(2.5)
7	nextPhoPla ce-(1.8)	nextPhoPla ce-(1.6)	nextPhoPla ce-(2.6)	nextPhoPla ce-(1.1)	nextPhoPla ce-(1.2)	nextPhoPla ce-(1.3)	nextPhoPla ce-(1.3)
8	dur-(1)	dur-(1.1)	dur-(1.4)	dur-(0.9)	dur-(0.7)	dur-(0.9)	dur-(1)
9	gender- (0.6)	gender- (0.6)	gender- (0.7)	gender- (0.7)	F0-(0.4)	gender- (0.5)	gender- (0.9)
10	spkAge- (0.5)	spkAge- (0.5)	spkAge- (0.6)	spkAge- (0.4)	gender- (0.3)	spkAge- (0.4)	F0-(0.7)

11	F0-(0.4)	F0-(0.4)	F0-(0.5)	F0-(0.4)	spkAge-(0.2)	F0-(0.4)	spkAge-(0.6)
12	locSyllInUtt-(0.2)	spRate-(0.2)	locSyllInUtt-(0.3)	locSyllInUtt-(0.1)	spRate-(0.1)	locSyllInUtt-(0.1)	spRate-(0.2)
13	spRate-(0.1)	locSyllInUtt-(0.1)	spRate-(0.2)	spRate-(0)	locSyllInUtt-(0.1)	spRate-(0)	locSyllInUtt-(0.2)
***“prePhoManPlace-” refers to the simultaneous removal of prePhoMan and prePhoPlace. ***“nextPhoManPlace-” refers to the simultaneous removal of nextPhoMan and nextPhoPlace.							

In all full spectral models, the combined parameters like prePhoManPlace (prePhoMan + prePhoPlace) and nextPhoManPlace (nextPhoMan + nextPhoPlace) contributed to spectral model identification accuracies of full spectral models' by 9.6~11.7% and 5.1~8.4%, respectively. These combined parameters in Group MP partly reflected the perceptual roles of the flanking phones in spectral identification of vowels, though not as contributive as the flanking phone identities in Group P (prePho 11.6~13.7%, nextPho 6.5~10.4%). It was also observed that the combination of manner and place of the preceding phone was perceptually more contributive than that of the following phone.

As for the contributions of single parameters, PrePhoMan (3.7~4.9%) and prePhoPlace (2.4~3.3%) were two of the most contributive cues, followed by locSyllInWord (2.3~3.3%), nextPhoMan (2.2~3.5%), and nextPhoPlace (1.1~2.6%). However, all the other parameters contributed far less: 0.7~1.4% for dur, 0.3~0.9% for gender, 0.2~0.6% for spkAge, 0.4~0.7% for F0, 0.1~0.3% for locSyllInUtt, 0~0.2% for spRate. These results also showed that non-spectral parameters except for manner and place of the flanking phones and locSyllInWord, exerted far less contributions to spectral vowel identification than argued in the literature.

5.4 Further discussion

In the literature, manner and place cues of the flanking consonants were reported to affect vowel formant transitions, constituting important cues in vowel perception (Fant 1960, Stevens and House 1963, Stevens and House 1963, Browman and Goldstein 1990, Hillenbrand et al. 2000). However, there have been almost no reports in the literature as to what extent manner and place of the flanking phones affected vowel transitions or vowel perception relative to other non-spectral cues.

The present study demonstrated that the perceptual contributions of manner and place of flanking phones on spectral vowel identification (i.e., coarticulation effects)

were far larger than the other non-spectral cues. Their contributions were maximized when both cues are fed simultaneously into full spectral models. However, manner and place of the flanking phones showed relatively rather limited perceptual contributions when compared to contributions of spectral properties and flanking phone identities. The flanking phone identities exerted more contribution than the combined manner and place. And manner of the flanking phones was perceptually more contributive than place.

The combined manner and place of the preceding phone were far more contributive than those of the following phone. The identity, manner, and place cues of the preceding phone were perceptually more contributive than those of the following phone. Though less contributive than identity, manner and place of the flanking phones, the syllable position of vowels in words also contributed to spectral identification of vowels more than the rest of the parameters in Group A and B including syllable positions of vowels in utterances.

As noted in the literature, F0 was an important variable for spectral perception of vowels (Miller 1953, Fujisaki and Kawashima 1968, Slawson 1968). And gender was also reported to affected vowel perception considerably as males and females are different in vocal tract size and shape (Fant 1966, Nordström 1977, Goldstein 1980, Traunmüller 1984). However, the present results showed that perceptual influence of F0 ($\leq 0.5\%$ for Group P, $0.7\leq$ for Group MP) and gender ($\leq 0.7\%$ for Group P, ≤ 0.9 for Group MP) on spectral identification of vowels is quite little or extremely limited in all full spectral models, which is rather unexpected. It means that vowel identification rates of all full spectral models stayed rather robustly high without referring to F0 and/or gender information. Duration ($\leq 1\%$ for Group P, ≤ 1.4 for Group MP), spkAge ($\leq 0.5\%$ for Group P, ≤ 0.6 for Group MP), spRate ($\leq 0.2\%$ for Group P, ≤ 0.2 for Group MP), and locSyllInUtt ($\leq 0.2\%$ for Group P, ≤ 0.3 for Group MP) exerted far less or almost no perceptual influence, strongly suggesting that the amount of perceptual contributions of them on vowel perception was overrated in the literature.

6. Conclusion

The NN pattern recognition modeling of spectral properties of Korean monophthong signals in spontaneous speech, showed that full dynamic spectral models

demonstrated far better vowel identification performance than full static ones. And both full dynamic and static spectral models showed far better vowel identification performance when the flanking phone identities were referred to than when manner and place of the flanking phones were referred to. However, these coarticulation-related cues showed more perceptual contributions to spectral vowel identification than the other non-spectral cue parameters such as speakers' gender, speakers' age, speaking rate, F0, vowel duration, and syllable positions of vowels in words and utterances.

And the identity, place and manner of the preceding phone (e.g. coarticulation information with preceding phones) were perceptually more contributive to spectral vowel identification than those of the following phone (e.g. coarticulation with following phones). Syllable location of the target vowel within a word was also perceptually contributive, though not as contributive as flanking phone identities and flanking phones' place and manner cues. However, syllable location of the target vowel within an utterance, duration, speaking rate, gender, and F0 showed rather little or almost no perceptual contribution to spectral model classification, far less than those cues related to flanking phones.

Though the full dynamic spectral models sampled at 20%, 50%, and 80% of the vowel duration showed considerably high vowel identification accuracy (up to 79.5% with Group P, up to 76.4% with Group MP), the model identification performance was unfortunately not satisfyingly comparable to human listeners' perception.

The present study is directly compared to the modeling study in Hong (2019), where 5 simple spectral models (F123_50, F123_2050, F123_5080, F123_2080, and F123_205080) were further fitted to one of the same non-spectral cue parameters as in this study for the individual perceptual contribution of a non-spectral parameter to spectral vowel identification. Hong's (2019) approach is different from the present study in that Hong's (2019) used simple spectral models to which one additional non-spectral parameter was further fed, to see the perceptual contribution of the non-spectral parameter to simple spectral models. This means that Hong's modeling approach ignored the possible interaction, if any, among non-spectral parameters in full spectral models' vowel identification, which the present study considered important.

Hong (2019) and the present study agreed in the observations that identities of flanking phones were the most contributive to spectral vowel identification, followed by manner and place of flanking phones and that properties of the preceding phone

were perceptually more contributive than those of the following phone. And all the other parameters (except for locSyllInWord) were not so contributive as argued in the literature. On the other hand, locSyllInWord exerted a little bit more perceptual contribution in Hong (2019) than in the present paper. However, it was far less contributive both in Hong (2019) and the present study than coarticulation-related cues like identities, place and manner of flanking phones.

The target vowels in the present study were restricted to seven Korean monophthongs. Further study is needed to see how the modeling results turn out when all Korean vowels (seven monophthongs plus ten diphthongs) are to be analyzed at the same time.

REFERENCES

- ASSMANN, PETER F. and GEOFFREY S. MORRISON. 2013. Introduction. In MORRISON, GEOFFREY S. and PETER F. ASSMANN (eds.), *Vowel Inherent Spectral Change*, 1-6. Berlin Heidelberg: Springer-Verlag.
- BOERSMA, PAUL and DAVID WEENINK. 2018. Praat: Doing phonetics by computer [computer program]. Version 6.0.37. Retrieved 14 March 2018 from <http://www.praat.org>.
- BOTTOU, LEON. 2010. Stochastic gradient descent (v.2). <https://leon.bottou.org/projects/sgd>.
- BROWMAN, CATHERINE P. and LOUIS M. GOLDSTEIN. 1990. Tiers in articulatory phonology with some implications for casual speech. In JOHN KINGSTON and MARY E. BECKMAN (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 341-376. Cambridge: CUP.
- BYRD, DAN. 1994. Relations of sex and dialect to reduction. *Speech Communication* 15, 39-54.
- 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.
- ENGSTRAND, OLLE. 1988. Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America* 83,

1863-1875.

- FANT, GUNNAR. 1960. *Acoustic Theory of Speech Production*. Gravenhage: Mouton.
- _____. 1966. A note on vocal tract size factors and non-uniform F-pattern scalings. *Laboratory Quarterly Progress and Status Report* 4, 22-30.
- FUJISAKI, HIROYA and TAKAKO KAWASHIMA. 1968. The influence of various factors on the identification and discrimination of synthetic speech sounds. Paper Presented at the 6th International Congress on Acoustics, August 1968, Tokyo.
- GAY, THOMAS. 1978. Effect of speaking rate on vowel formant movements. *Journal of Acoustical Society of America* 63, 223-230.
- GOLDSTEIN, URSULA. 1980. *An Articulatory Model for the Vocal Tracts of Growing Children*. PhD dissertation. MIT, Cambridge, MA.
- HENTON, CAROLINE. 1992. The abnormality of male speech. In WOLF, GEORGE (ed.), *New Departures in Linguistics*, 27-55. New York: Garland Publishing Co.
- HILLENBRAND, JAMES M. 2013. Static and dynamic approaches to vowel perception. In MORRISON, GEOFFREY S. and PETER F. ASSMANN (eds.), *Vowel Inherent Spectral Change*, 9-30. Berlin-Heidelberg: Springer Verlag.
- HILLENBRAND, JAMES M., LAURA A. GETTY, MICHAEL J. CLARK and KIMBERLEE WHEELER. 1995. Acoustic characteristics of American English vowels. *Journal of Acoustical Society of America* 97, 3099-3111.
- HILLENBRAND, JAMES M., MICHAEL J. CLARK and TERRANCE M. NEAREY. 2000. Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America* 109, 748-763.
- HIRATA, YUKARI and KIMIKO TSUKADA. 2004. The effects of speaking rates and vowel duration on formant movements in Japanese. In WARREN, WILLIS and SANG-HOON PARK (eds.), *Proceedings of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*, 73-85.
- HONG, SOONHYUN. 2019. Perceptual influence of non-spectral cues on spectral properties of Korean vowel signals in Seoul Corpus: A neural network modeling study. *Studies in Modern Grammar* 102, 135-164. The Society of Modern Grammar.
- KINGMA, DIEDERIK and JIMMY BA. 2014. Adam: A method for stochastic optimization. ArXiv Preprint ArXiv:1412.6980.
- JOHNSON, KEITH. 1990. The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America* 88, 642-654.

- _____. 2008. Speaker normalization in speech perception. In DAVID B. PISONI and ROBERT E. REMEZ (eds.). *The Handbook of Speech Perception*, 363-389. Malden: Blackwell Publishing.
- LECUN, YANN, LEON BOTTOU, GENEVIEVE B. ORR and KLAUS-ROBERT MÜLLER. 1996. Efficient backprop. In *Proceedings of Neural Networks: Tricks of the Trade*, 9-50.
- LEHISTE, ILSE and DAVID MELTZER. 1973. Vowel and speaker identification in natural and synthetic speech. *Language and Speech* 16, 356-364.
- LEHISTE, ILSE and GORDON E. PETERSON. 1961. Transitions, glides, and diphthongs. *Journal of Acoustical Society of America* 33, 268-277.
- LINDBLOM, BJORN. 1963. Spectrographic study of vowel reduction. *Journal of Acoustical Society of America* 35, 1773-1781.
- LINDBLOM, BJORN and MICHAEL STUDDERT-KENNEDY. 1967. On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America* 42, 830-843.
- MILLER, ROGER L. 1953. Auditory tests with synthetic vowels. *Journal of Acoustical Society of America* 25, 114-121.
- MOON, SEUNG-JAE and BJORN LINDBLOM. 1994. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of Acoustical Society of America* 96, 40-55.
- NEAREY, TERRANCE M. and PETER ASSMANN. 1986. Modeling the role of vowel inherent spectral change in vowel identification. *Journal of Acoustical Society of America* 80, 1297-1308.
- NG, ANDREW. 2015. Unsupervised feature learning and deep learning. Lecture note. <https://www.csee.umbc.edu/courses/graduate/678/fall14/visionaudio.pdf>.
- NORDSTRÖM, PER-ERIK. 1977. Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics* 5, 81-92.
- PETERSON, GORDON E. and HAROLD E. BARNEY. 1952. Control methods used in a study of vowels. *Journal of Acoustical Society of America* 24, 175-184.
- RUMELHART, DAVID E., GEOFFREY E. HINTON and RONALD J. WILLIAMS. 1986. Learning representations by back-propagating errors. *Nature* 323, 533-536.
- SLAWSON, A. WAYNE. 1968. Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America* 43, 87-101.

- STEVENS, KENNETH N. and ARTHUR S. HOUSE. 1963. Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research* 6, 111-128.
- STRANGE, WINIFRED, JAMES J. JENKINS and THOMAS L. JOHNSON. 1983. Dynamic specification of coarticulated vowels. *Journal of Acoustical Society of America* 74, 695-705.
- TRAUNMÜLLER, HARTMUT. 1984. Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. *Speech Communication* 3, 49-61.
- VAN SON, R. J. J. H. and LOUIS C. W. POLS. 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. *Journal of the Acoustical Society of America* 92, 121-127.
- WRIGHT, RICHARD, STEFAN FRISCH and DAVID B. PISONI. 1997. Research on spoken language processing: Speech perception. *Progress Report No. 21* (1996-1997). Indiana University.
- YOON, TAE-JIN. 2019. Two-layer neural network-based vowel classification experiments using formant trajectory. *Studies in Phonetics, Phonology and Morphology* 25.1, 95-112. The Phonology-Morphology Circle of Korea.
- YUN, WEONHEE, KYUCHUL YOON, SUNWOO Park, JUHEE LEE, SUNGMOON CHO, DUCKSOO KANG, KOONHYUK BYUN, HYESEUNG HAHN and JUNG SUN KIM. 2015. The Korean corpus of spontaneous speech. *Phonetics and Speech Sciences* 7.2, 103-109. The Korean Society of Speech Sciences.

Soonhyun Hong (Professor)
 Department of English Language and Literature
 Inha University
 100 Inharo, Michuholgu
 Incheon 22212, Republic of Korea
 e-mail: shong@inha.ac.kr

Received: February 4, 2020
 Revised: April 20, 2020
 Accepted: April 27, 2020