pISSN 2005-8063
eISSN 2586-5854
2022. 6. 30.
Vol.14 No.2
pp. 11-17

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2022.14.2.011

Check for updates

Speech recognition rates and acoustic analyses of English vowels produced by Korean students

Byunggon Yang*

Department of English Education, Pusan National University, Busan, Korea

Abstract

English vowels play an important role in verbal communication. However, Korean students tend to experience difficulty pronouncing a certain set of vowels despite extensive education in English. The aim of this study is to apply speech recognition software to evaluate Korean students' pronunciation of English vowels in minimal pair words and then to examine acoustic characteristics of the pairs in order to check their pronunciation problems. Thirty female Korean college students participated in the recording. Speech recognition rates were obtained to examine which English vowels were correctly pronounced. To compare and verify the recognition results, such acoustic analyses as the first and second formant trajectories and durations were also collected using Praat. The results showed an overall recognition rate of 54.7%. Some students incorrectly switched the tense and lax counterparts and produced the same vowel sounds for qualitatively different English vowels. From the acoustic analyses of the vowel formant trajectories, some of these vowel pairs were almost overlapped or exhibited slight acoustic differences at the majority of the measurement points. On the other hand, statistical analyses on the first formant trajectories of the three vowel pairs revealed significant differences throughout the measurement points, a finding that requires further investigation. Durational comparisons revealed a consistent pattern among the vowel pairs. The author concludes that speech recognition and analysis software can be useful to diagnose pronunciation problems of English-language learners.

Keywords: speech recognition, formant trajectory, duration, English vowel, Korean students

1. Introduction

Vowels are important linguistic units of daily conversation. When one pronounces a vowel sound in a word incorrectly, the conversation may stop unexpectedly and may require extra time to return to the point of divergence. Weckwerth (2022) mentioned that describing vowels is considered to be much more difficult than describing consonants because people cannot obtain much tactile feedback from the speech organs when producing vowels. He noted the two most basic traits of vowels in phonetics and phonology: vowel height (or openness) and tongue advancement (or position on the front-back dimension). He added that the second dimension did not involve a single prevailing term, for example, frontness, backness, and anteriority location. He described that height, tongue

* bgyang@pusan.ac.kr, Corresponding author

Received 27 April 2022; Revised 7 June 2022; Accepted 7 June 2022

ⓒ Copyright 2022 Korean Society of Speech Sciences. This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

advancement and lip rounding collectively formed the basis of vowel quality. Kennedy (2022) also mentioned that the phonology of vowels relied on placing abstract categories of height, backness, rounding, and length over gradient phonetic dimensions. He added that the phonological descriptions of tenseness in English phonology traditionally distinguish tense vowels from their lax counterparts according to their distribution in English words. In other words, lax vowels can occur only in closed syllables with a coda, while tense vowels do not have such restrictions based on phonotactics.

The acoustic differences between the tense and lax vowels include both frequency and temporal aspects. Lax vowels are more centralized in the vowel space and display diphthongal characteristics. Davenport & Hannahs (1998) noted that the front vowel [ɪ] was somewhat lower and more centralized than [i] in the vowel space. De Decker & Nycz (2012) acoustically examined the English vowel [æ] and found that it could be grouped with either tense or lax vowels depending on various speakers. The same classification was applied to the high back vowel pair [u, ʊ]: the former is long and peripheral, and the latter is short and centralized. In the lower front part of the vowel space, there are two vowels: a short mid-front vowel [ɛ] and a short low-front vowel [æ]. Lee & Rhee (2019) examined the relationship between vowel production and proficiency levels in read English texts and reported that higher-rated speakers could distinguish vowel contrasts better in the front vowel pairs /i, ɪ/ and /ɛ, æ/ but had difficulty distinguishing the central and back vowel pairs spectrally. The researchers measured the first two formant values at the middle point of the vowel segment and statistically compared them between the two level groups. They also noted that the lower-rated speakers used temporal cues less for the tense and lax vowel pairs.

The distinction of tenseness also depends on the temporal aspect of the vowel. Since the pronunciation of tense vowels requires greater effort and tension in the muscles of the vocal tract, they tend to have longer durations than lax vowels. Davenport & Hannahs (1998) classified the high front vowels into the long monophthong [iː] and the short monophthong [ɪ]. In addition, a tense vowel is reported to be longer than a lax vowel of a similar height if the adjacent syllable environment is equal. Yang (1996) noted that the average duration of the vowel [æ] produced by ten male Americans was 126 ms on average, while that of the vowel [ɛ] was 132 ms in the hVd context. In females, the duration of vowel [æ] was shorter than that of the vowel [ɛ] by 2 ms. The vowel [æ] is generally classified as a lax vowel, but it is treated as an exception to display that it is not shorter than the low-tense vowel [ɒ] (Nearey, 2006).

To date, few studies have explored both speech recognition and acoustic analyses of English vowels produced by Korean speakers. Yang (2010) reported that Korean students had difficulty producing English high tense and lax vowel pairs. On the other hand, the American students produced tense and lax pairs much more distinctly than their Korean counterparts did from the formant trajectories measured at six equidistant points. Korean students may make errors partly because they think the vowels are similar to those in Korean and partly because they do not acquire correct vowel targets for a given word. The main purpose of this study was to apply speech recognition software to Korean students' word production to obtain correctly recognized rates and to compare formant trajectories and durations of English vowel pairs. The results may provide pedagogical implications in the teaching and evaluation of English vowel pronunciation.

2. Method

2.1. Participants and Speech Stimuli

Thirty female Korean students participated in the recordings. They were divided into two groups: 13 undergraduate students and 17 graduate students who were taking English phonetics courses. They produced a list of 40 words twice each and recorded them on their own mobile phones. The task was given as a part of home assignments for the courses. The speech stimuli were as follows: *did, it, steal, axe, cooed, leave, deed, sad, guess, set, full, could, bed, who'd, said, eat, stewed, sat, live, left, hood, pull, it, bad, did, should, sheep, fool, gas, ship, heel, shooed, hill, laughed, still, pool, X, stood, ship,* and *it*. The first two words at the beginning and the last two words at the end of the list were discarded in the analysis. Those four peripheral dummy words were included to facilitate the participants' adaptation to the start of the recording and to avoid lengthening them at the end of the list. Thus, the total number of recorded words was 2,160 words (30 participants×36 words×two repetitions). The duration of each recorded file was approximately two to three minutes per participant.

2.2. Data Collection and Analysis Procedures

Data collection was performed in two steps. First, the speech recognition rate was obtained using a Microsoft Word menu "Dictate" in Office365 on an iMac. Second, the first and second formant trajectories and vowel durations were collected using Praat (v.6.2; Boersma & Weenink, 2021). Statistical analyses were conducted on the data using R (R Core Team, 2021).

The speech recognition rate was collected by playing each sound file on the iMac and by counting the number of correctly recognized words of the list. The recognition rate was obtained if either one of the two productions corresponded to the vowel of the target word. The lenient criterion was adopted to consider human mistakes and machine errors. Human beings may say a wrong word on the first attempt but may correct it on the second try and continue their conversation without any interruption. In addition, computer recognition was found to be generally consistent but yielded gibberish outputs for a list of unrelated words in a row. Thus, the author reset the system several times in case it produced obvious error words continuously. "Dictate" was developed by Microsoft Garage group applying recent artificial intelligence and machine learning technology to speech recognition. According to an evaluative study on the speech recognition of live lecture conditions, accuracy rates amounted to greater than 95% by Google and Microsoft comapanies (Millett, 2021). They added the importance of good audio and clear speech for better results. Previous studies reported that the reliability of Google speech recognition was also very high for English sentences by native speakers (Yang, 2017; Yang, 2020), but the recognition tool was inaccessible for this research.

The formant and duration measurements were performed by creating a few Praat scripts to secure valid and reliable data. A folder handling script opened all the sound files of the participants and placed them onto the object window. Then, an object script initialized the participant's name and opened the selected sound file on the View & Edit window. An editor window script prompted the author to choose the vowel segment for an analysis and moved the start and end of selection to the nearest zero crossing (Yang, 2009).

The duration and average values of the first and second formants of each vowel of the word list were measured from every sound segment of 25.6 ms at ten time points, which were assigned by dividing the total duration periodically. Those values were appended to the result file on the computer. The author visually checked energy bands on the spectrograms and corrected obvious errors of formant measurements in light of the average values. Formant jumps or drops were found at the initial and final measurement points of a waveform. In some cases, the formant number settings were tweaked to best match the formant trajectory to a smoothly moving contour. Boxplots of formant trajectories at each measurement point of the word were plotted to display a general formant distribution by R. Generalized additive mixed models (GAMMs) were applied to the first formant data, which reflect the degree of jaw opening (Pickett, 1980; Sóskuthy, 2017; van Rij, 2015; Wood, 2006). The formant trajectories were measured to compare the nonlinear vowel characteristics throughout the vowel segments (Yang, 2010).

3. Results and Discussion

3.1. Speech Recognition Rates

Table 1 lists a statistical summary of the speech recognition rates of thirty Korean female students.

Table 1. Statistics of correctly recognized words by a total count and by vowel height

n	Correctly recognized words	Higher vowels [i, u, ɛ]	Lower vowels [ɪ, ʊ, æ]
1,080	591 (54.7%)	290 (53.7%)	301 (55.7%)

As shown in Table 1, the total number of correctly recognized words was 591 out of 1,080, amounting to 54.7%, just over half of the whole list of words. These rates fall just above a chance level. The actual recognition rate would be lower than those in the table if we assigned one point to each and every correct pronunciation of all the words. As described in the previous section, the recognition was applied leniently by counting the number of words when either one of the two productions of a given word was correct.

The division of higher and lower vowels based on the height of the tongue inside the oral cavity yielded comparable recognition rates, 53.7% and 55.7%, respectively. From the results, we notice that the Korean students had difficulty distinguishing these vowels. We will comment on the results of the traditional division of tense–lax categories in the following discussion of individual vowels.

Table 2. Statistics of correctly recognized words grouped by the vowel and word in descending order of frequency

Word	[i]	Word	[ɪ]	Word	[u]	Word	[ʊ]	Word	[ɛ]	Word	[æ]
leave	29	*ship*	24	*stewed*	18	*full*	24	*set*	28	*sad*	29
deed	24	*it*	22	*shooed*	12	*pull*	21	*left*	26	*bad*	28
eat	20	*still*	13	*who'd*	10	*could*	18	*guess*	22	*axe*	16
sheep	21	*hill*	12	*cooed*	10	*stood*	14	*X*	17	*laughed*	15
heel	16	*did*	12	*pool*	9	*hood*	12	*said*	5	*gas*	14
steal	14	*live*	2	*fool*	7	*should*	12	*bed*	2	*sat*	13
Sum	124	Sum	85	Sum	66	Sum	101	Sum	100	Sum	115
%	68.9	%	47.2	%	36.7	%	56.1	%	55.6	%	63.9

Sum indicates the total number of correct words while % is a percentage.

Table 2 lists the number of correctly recognized words grouped by the vowel and word in descending order of frequency. The sum of the tense vowel [i] presents the highest speech recognition rate with 68.9%, followed by the lower front vowel [æ] with 63.9%. The tense back vowel [u] records the lowest recognition rate with 36.7%. The first four columns list the traditional tense–lax vowel pairs. If we combine the tense vowels [i, u] into a group, the average recognition rate is 53%. The lax counterparts [ɪ, ʊ] recorded 52% recognition. The difference was very small. Both the tense and lax vowel productions must have been difficult to distinguish for the Korean students. The words 'live' and 'bed' recorded the lowest recognition rate. A majority of the students produced the word 'live' with a wider jaw opening to produce the vowel sound of the word 'leave'. On the other hand, the low rate for the word 'bed' seems to be related to an overly conscious attitude toward the pair. Classroom teaching and training in these vowel sounds to emphasize the importance of opening the jaw further to achieve the correct vowel sound [æ] must have pushed them to produce the lax vowel like the tense vowel. The highest score, 29 out of 30, was recorded in the two words 'leave' and 'sad'. Interestingly, twenty-eight students produced the two words 'bad' and 'set' correctly. However, the recognition in the word 'sat' dropped to 13 counts despite their capacity to produce both vowel sounds correctly. We can conclude that the students could produce both vowels, but they must have not acquired the vowel in their lexicon. The author proposes that the students anchor their acquisition firmly on their foreign language system so that they can produce any given word pairs correctly in daily conversation. The individual choice of a student's target vowel immediately after its pronunciation might be useful to clarify the production of the correct vowel sound. Moreover, several tense and lax vowel pairs may have to be included in an evaluation sheet to correctly measure a student's authentic capacity to distinguish the pair or not. Further studies would be desirable to determine whether a prompt of a phonetic vowel symbol for a given word might be helpful for better recognition after teaching students intelligible articulatory gestures.

3.2. Vowel Formant Trajectories

Figures 1 and 2 illustrate boxplots of the first and second formant frequencies of two tense and lax vowel pairs [i, ɪ] and [u, ʊ]. Figure 3 displays a boxplot of those formant values of the vowel pair [ɛ, æ]. These values were collected from all thirty Korean students at ten measurement points. As seen in the figure, the lower bound for the first formant values of the tense and lax vowel pair [i, ɪ] is approximately 200 Hz, and the upper bound is approximately 800 Hz. The interquartile ranges at the measurement points are wider in the later part of the vowel segments. This trend may reflect the coarticulation effects of the following codas of various target words on the formant trajectories. Yang (2009) showed a coarticulation effect of the vowel trajectory on the alveolar coda [d], whose locus converged toward 1,800 Hz (Delattre et al., 1955). Generally, the median first formant values of the lax vowel [ɪ] denoted by dotted lines are higher than those of the tense counterpart denoted by thick lines. The first formant tends to reflect the degree of jaw opening (Pickett, 1980); thus, the students must have attempted to distinguish the vowel pairs controlling the jaw. On the other hand, the median second formant values of the lax vowel [ɪ] are lower than those of the tense counterpart, which indicates a further backing gesture of the tongue for the lax vowel. Outlier points appear mostly over the upper bound of the first formant values, while they stay mostly under the lower bound of the second formant.

The vowel pair [u, ʊ] in Figure 2 indicates a similar trend. The median first formant values of the lax vowel [ʊ] are slightly higher than those of the tense counterpart. However, the median second formant values of the lax vowel [ʊ] are higher than those of the tense vowel [u]. The second formant values at the first and fourth measurement points almost overlap. These slight differences might not be sufficient to signal listeners to distinguish the two vowels from each other. Here, again, outlier points appear mostly over the upper bound of the first formant values, while they mostly appear under the lower bound of the second formant. In Section 3.1., we reported a low recognition rate for the back vowels, i.e., 36.7% for the vowel [u]. However, the lax vowel [ʊ] recorded a recognition rate of 52%. It would be desirable to investigate the unequal recognition rates of the tense–lax vowel pair to explain the mismatch.

Figure 3 displays a boxplot of the formant trajectories of the vowel pair [ɛ, æ]. The median first formant values of the vowel [æ] are slightly higher than those of the vowel [ɛ], while the median second formant values of the two vowels almost overlap, except for the slightly lower values at the later part of the segment. In the figure, outlier points spread both over and under the outer fences of the first and second formants. Those outlier points might be dependent on the vowel height of individual speakers and on the various onsets and codas of the recorded words.

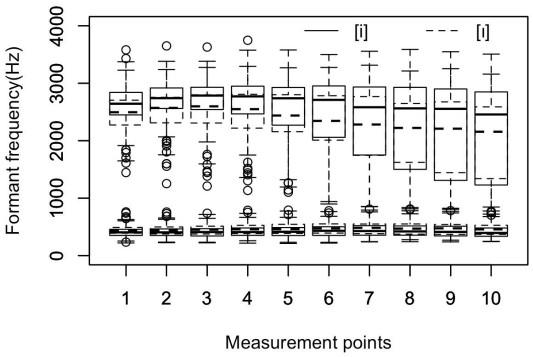

Figure 1. A boxplot of the first and second formant frequencies of two tense and lax vowel pairs [i, ɪ]

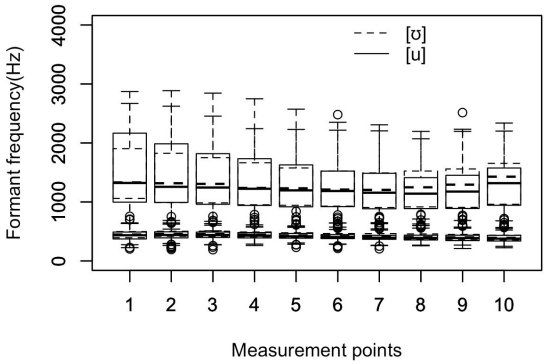

Figure 2. A boxplot of the first and second formant frequencies of two tense and lax vowel pairs [u, ʊ]

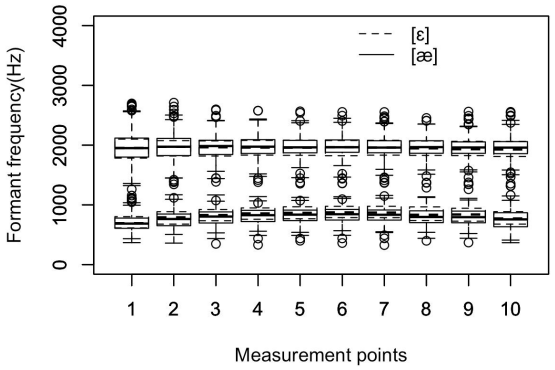

Figure 3. A boxplot of the first and second formant frequencies of two tense and lax vowel pairs [ɛ, æ]

Table 3. First and second median formant values of six vowels at ten measurement points

Formant	Vowel	1	2	3	4	5	6	7	8	9	10
F1	i	404	404	409	410	414	431	427	420	412	394
	ɪ	436	439	453	459	461	474	483	469	474	456
	u	435	439	439	428	417	404	392	388	387	377
	ʊ	447	458	454	447	432	419	407	403	400	390
	ɛ	686	763	804	827	836	843	834	797	794	757
	æ	693	792	829	857	867	875	871	834	839	772
F2	i	2,644	2,741	2,786	2,772	2,739	2,711	2,583	2,563	2,555	2,457
	ɪ	2,496	2,571	2,599	2,550	2,439	2,345	2,282	2,223	2,210	2,157
	u	1,322	1,256	1,244	1,223	1,196	1,185	1,156	1,140	1,176	1,318
	ʊ	1,331	1,317	1,306	1,237	1,231	1,210	1,204	1,248	1,293	1,428
	ɛ	1,950	1,973	1,985	1,974	1,963	1,965	1,959	1,967	1,961	1,959
	æ	1,953	1,973	1,965	1,959	1,958	1,966	1,960	1,946	1,932	1,927

Table 3 lists the median formant values of the six vowels at ten measurement points. The median values may better represent formant trajectories because they are less affected by extreme values. This paper used the analysis results from Praat after correcting prominent errors in the final dataset, but it would be desirable to check and adjust any inappropriate formant values exhaustively considering the vowel spectra (Yang, 1990; Yang, 1996). From the table, one can note that the first formant values of the lax vowels of [ɪ, ʊ] are higher than those of the tense vowels [i, u] throughout the measurement points, as seen in Figure 1. Numerically, the average first formant difference across the ten measurement points amounted to 47.9 Hz for the front vowel pair, while it was 15.1 Hz for the back vowel pair. The average first formant difference for the front low vowel pair was 28.8 Hz. All these marginal acoustic differences might be related to the low word recognition rates of the students. The average second formant differences across the ten measurement points are 267.9 Hz for the high front vowel pair, 58.9 Hz for the high back vowel pair, and 12.7 Hz for the high low vowel pair. Here, we will focus on the first formant difference because the major qualitative vowel differences are related to the jaw opening gesture, and the first formant tends to reflect the articulatory gestures of speakers (Pickett, 1980). Yang (1996) reported that the average first formant values of the vowels [i, ɪ] produced by ten American male speakers were 286 Hz and 409 Hz, respectively. The difference amounts to 123 Hz. The values for ten American female speakers were 390 Hz and 466 Hz, a difference of 76 Hz. In Table 3, the first formant values of the vowel pair [i, ɪ] at the third measurement point are 409 Hz and 453 Hz, respectively. The third measurement point may be comparable to the previous measurement point at one-third of the total vowel duration in Yang

(1996). The difference amounts to 44 Hz. For the front vowel pair [ɛ, æ], the average difference was 28.8 Hz. Here, again, the small acoustic difference in the Korean students' data might not be sufficient to be recognized as two distinct vowels. In English, the formant difference between the two vowels in the first formant is reported to be much wider (Yang, 1996). For example, the average first formant values of the vowel [ɛ, æ] produced by American male speakers were 531 Hz and 687 Hz, respectively, with a difference of 156 Hz. These values for the American female speakers were 631 Hz and 825 Hz, respectively, a difference of 194 Hz. On the other hand, the difference in Table 3 amounts to 25 Hz at the third measurement point, with 804 Hz and 829 Hz for the vowel pair [ɛ, æ]. Finally, the median first formant values of the vowels [u, ʊ] at the third measurement point were 439 Hz and 454 Hz, respectively, a difference of 15 Hz. The difference appears almost negligible when we consider that the average first formant values of the vowels [u, ʊ] produced by the American male speakers in the previous study were 333 Hz and 446 Hz, respectively, a difference of 113 Hz. Moreover, the values for the female speakers were 417 Hz and 491 Hz, a difference of 74 Hz. Yang & Whalen (2015) also reported clear contrasts between the tense and lax vowel pairs on a vowel space in which eighteen American male and female participants distinguished those vowels in a clear speaking style. All these acoustic median differences we have observed thus far indicate that the Korean students' productions were not sufficient to be recognized as separate vowels. Thus, one can say that these students produced the two vowel pairs [u, ʊ] and [ɛ, æ] almost identically.

Here, we attempted to test for statistical significance between the formant trajectories of the three vowel pairs using GAMMS after normalizing the formant data individually. The scale function in R was used to normalize the individual raw data of the first formant values by obtaining the z-score (Yang, 2019).

A partial statistical summary of the vowel pair [i, ɪ] is as follows:

Family: gaussian
Link function: identity
Formula: scaled ~ vowelordered+s(point, k=9)+
s(point, by=vowelordered, k=9)
Parametric coefficients:

	Estimate	*Std. Error*	*t-value*	*Pr (>|t|)*
(Intercept)	*3.31*	*.01*	*296.01*	*<2e-16**
vowelorderedɪ	*.196*	*.016*	*12.34*	*<2e-16**

Approximate significance of smooth terms:

	edf	*Ref.df*	*F-value*	*p-value*
s(point)	*3.5*	*4.3*	*11.6*	*<2e-16**
s(point):vowelorderedI	*1.00*	*1.00*	*.01*	*.91*

R-sq.(adj)= .057 Deviance explained=5.84%
GCV=.22 Scale est.=.22 n=3,500 **p<.05.*

From the partial summary above, the first formant of the tense vowel [i] is significantly different from that of the lax vowel [ɪ]. The deviance explained from the *R*-squared adjusted values for the vowel pair [i, ɪ] was 5.84%. The other two pairs yielded highly significant differences as well. The deviance explained for the vowel pair [u, ʊ] amounted to 11.1%, while that for the vowel pair [ɛ, æ] was 21.5%. Here, we display only the smoothed plots of the three pairs in Figure 4. All three estimated difference plots illustrated significant differences throughout the ten measurement points, which can be recognized from the smoothed plots. The significant results among the three pairs may be related to the large amount of collected data, and further investigation is needed to match the speech recognition results appropriately.

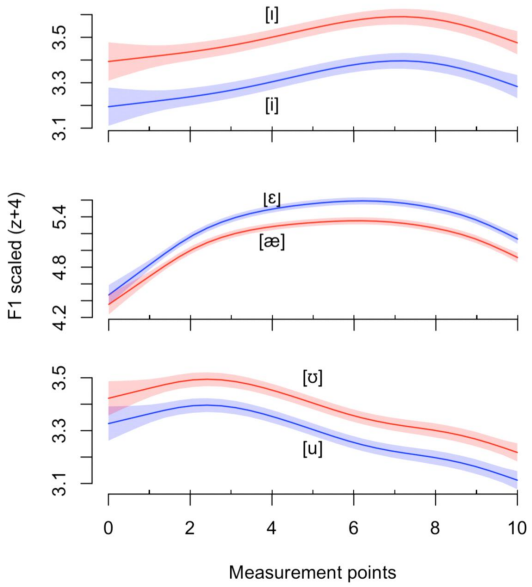

Figure 4. Smoothed plots of the three vowel pairs at ten measurement points by generalized additive mixed model.Y-axis indicates the first formant values scaled individually into z-scores with 4 added.

To explain the mismatch of the low speech recognition and the significant statistical result, we will briefly review two relevant articles (Yang, 2006; Yang & Whalen, 2015). Yang (2006) synthesized vowel stimuli increasing or decreasing formant values of a given vowel and asked twenty-seven American and Korean listeners to decide whether the original vowel is different from the varied vowel. On average, the American female listeners produced 115 Hz as an acceptable range of the first formant of the same vowel and 244 Hz for the second formant. The acceptable range for the first formant of the vowel [i] of the American female listeners started from a lower F1 value of 205 Hz to a higher F1 value of 320 Hz, while that for the vowel [ɪ] indicated a range from 383 Hz to 473 Hz. The high back vowel pair [u, ʊ] ranged from 260 Hz to 389 Hz for the tense vowel [u] and from 395 Hz to 523 Hz for the lax counterpart [ʊ]. The front mid-vowel [ɛ] ranged from 505 Hz to 617 Hz, while the front low vowel [æ] was recognized within the range of 576 to 726 Hz. Moreover, Yang & Whalen (2015) performed perception experiment using three sets of the synthetic stimuli: the base set, and two additional sets scaled up or down by 15% of the first and second formant frequency values of the base set. Eighteen male and female participants perceived the synthetic stimulus sets with clearly separated points on the vowel space, especially tense and lax vowel pairs (Yang & Whalen, 2015). Referring to these results, one can say that the current acoustic difference between the first and second formants might not be sufficient to be distinguishable in the recognition despite the statistically significant differences of GAMMS. This might explain the low recognition rate in Table 1. Since the vocal tract anatomy of both Korean and American females varies, further studies with appropriate speaker normalization might

be necessary to resolve the mismatch.

3.3. Vowel Durations

This section briefly addresses vowel durations to examine any interesting pattern in the measured data. Table 4 lists descriptive statistics of the six vowels in milliseconds. The vowel [u] displays the longest duration, followed by the vowel [ɛ]. The range of differences in the medians and means ranged from 6 ms to 20 ms. The mode may better define the temporal characteristics of each vowel, avoiding the influence of extreme outliers. The *SD*s are wider for the higher vowels than for the lower vowels [ɛ, æ].

Table 4. Basic statistics of the durations in milliseconds of the six vowels of English words produced by Korean students

Vowels	i	ɪ	u	ʊ	ɛ	æ
Median	296	223	306	281	169	202
Mean	284	233	315	287	189	222
SD	127	119	101	99	87	95

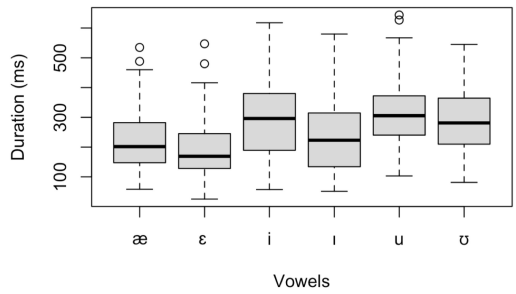

Figure 5. A boxplot of vowel durations of English words produced by Korean students

Figure 5 illustrates a boxplot of vowel durations of English words produced by thirty female Korean students. In the figure, the durations of the lax vowels [ɪ, ʊ] are generally shorter than those of the tense counterparts [i, u]. The low lax vowel [æ] is longer than the mid-front vowel [ɛ]. Since the boxplot shows the median values, we obtained a 73 ms difference between [i, ɪ] and a 25 ms difference between [u] and [ʊ]. The difference amounts to 33 ms for the mid- and low vowels [ɛ, æ]. Here the relationship between the tense and lax high vowels appeared quite consistent. We can conclude that Korean students were quite consistent in realizing temporal aspects of tense and lax vowel pairs although several participants produced mid- and low vowels with a different pattern.

Finally, we will consider a temporal organization of the participants from a boxplot of durations of the vowel pair [ɛ, æ] in twelve words in Figure 6. From the figure, one can note that the duration of the lower vowel [æ] appears longer than that of the higher vowel [ɛ]. In addition, the vowel followed by a voiced coda records a longer duration than the vowel with a voiceless coda. The bandwidth between the upper and lower bounds varied. These differences might be derived from various contexts, specifically from the adjacent onsets and codas of the target words. These phonotactic factors might have influenced the participants' pronunciation of the vowels along with the peripheral articulatory gestures. Further studies considering various syllable structures would explain certain temporal patterns of the participants.

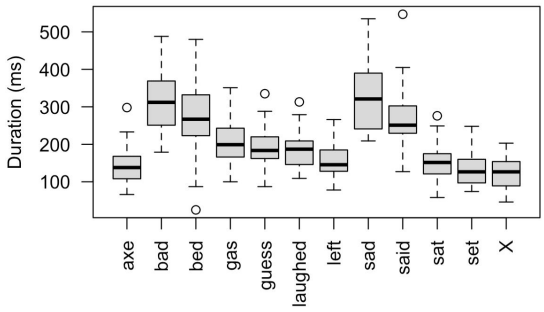

Figure 6. A boxplot of durations of the vowel pair [ɛ, æ] in twelve English words produced by Korean students

4. Summary and Conclusion

The English vowels form important segmental sounds of words in daily communication. This study collected thirty Korean female students' pronunciations of thirty-six English words twice each and evaluated their production using speech recognition software followed by acoustic analyses to diagnose their pronunciation problems. Speech recognition rates were obtained to count correctly recognized words out of the total number of words. Formant trajectories and durations were also collected using Praat. The results showed that the overall recognition rate was 54.7%. Various rates were reported depending on the vowel pairs. Some students produced the same vowel for the qualitatively different tense and lax counterparts. From the acoustic analyses of the vowel formant trajectories, some of these vowel pairs were almost overlapped or had slight acoustic differences at the majority of ten measurement points. Median values clearly showed this trend, but statistical tests revealed significant differences in the moving formant trajectories. Finally, durational comparisons revealed consistent patterns: a longer duration for the high tense vowel and for a voiced coda. The author concludes that speech recognition and analysis software can be useful to diagnose pronunciation problems of English learners but that caution should be exercised in the interpretation of statistical results. Further studies would be desirable to train students to acquire intelligible vowel pronunciations and to evaluate them to establish an appropriate teaching strategy.

References

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer (version 6.2) [Computer program]. Retrieved from http://www.praat.org/

Davenport, M., & Hannahs, S. J. (1998). *Introducing phonetics and phonology.* London, UK: Hodder Arnold.

De Decker, P. M., & Nycz, J. R. (2012). Are tense [æ]s really tense? The mapping between articulation and acoustics. *Lingua*, *122*(7), 810-821.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, *27*(4), 769-773.

Kennedy, R. (2022). The phonetics/phonology interface. In R. A. Knight, & J. Setter (Eds.), *The Cambridge handbook of phonetics* (pp. 682-706). Cambridge, UK: Cambridge University Press.

Lee, S., & Rhee, S. C. (2019). The relationship between vowel

production and proficiency levels in L2 English produced by Korean EFL learners. *Phonetics and Speech Sciences*, *11*(2), 1-13.
Millett, P. (2021). Accuracy of speech-to-text captioning for students who are deaf or hard of hearing. *Journal of Educational*, *Pediatric & (Re)Habilitative Audiology*, *25*, 1-13.
Nearey, T. (2006). English vowels. Linguistics 205 course notes of practical phonetics. Retrived from https://sites.ualberta.ca/~tnearey/Ling205/ Week4/EnglishVowelsNarrow4Up.pdf
Pickett, J. M. (1980). *The sounds of speech communication: A primer of acoustic phonetics and speech perception (Perspectives in Audiology Series)*. Baltimore, MD: University Park Press.
R Core Team. (2021). R: A language and environment for statistical computing (version 4.1.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. Retrieved from https://arxiv.org/abs/1703.05339v1
van Rij, J. (2015). Overview of GAMM analysis of time series data. Retrieved from https://jacolienvanrij.com/Tutorials/GAMM.html
Weckwerth, J. (2022). Vowels. In R. A. Knight, & J. Setter (Eds.), *The Cambridge handbook of phonetics* (pp. 40-64). Cambridge, UK: Cambridge University Press.
Wood, S. N. (2006). *Generalised additive models: An introduction with R.* Boca Raton, FL: CRC Press.
Yang, B. (1990). *Development of vowel normalization procedures: English and Korean* (Doctoral dissertation). The University of Texas, Austin, TX.
Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics*, *24*(2), 245-261.
Yang, B. (2006). Discrimination of synthesized English vowels by American and Korean listeners. *Speech Sciences*, *13*(1), 7-27.
Yang, B. (2009). Formant trajectories of English vowels produced by American males. *Phonetics and Speech Sciences*, *1*(3), 65-72.
Yang, B. (2010). Formant trajectories of English high tense and lax vowels produced by Korean and American speakers. *Korean Journal of Linguistics*, *35*(2), 407-423.
Yang, B. (2017). Google speech recognition of an English paragraph produced by college students in clear or casual speech styles. *Phonetics and Speech Sciences*, *9*(4), 43-50.
Yang, B. (2019). A comparison of normalized formant trajectories of English vowels produced by American men and women. *Phonetics and Speech Sciences*, *11*(1), 1-8.
Yang, B. (2020). An evaluation of Korean students' pronunciation of an English passage by a speech recognition application and two human raters. *Phonetics and Speech Sciences*, *12*(4), 19-25.
Yang, B. (2022). Measuring vowels. In R. A. Knight, & J. Setter (Eds.), *The Cambridge handbook of phonetics* (pp. 261-284). Cambridge, UK: Cambridge University Press.
Yang, B., & Whalen, D. H. (2015). Perception and production of English vowels by American males and females. *Australian Journal of Linguistics*, *35*(2), 121-141.

• **Byunggon Yang,** Corresponding author
Professor, Dept. of English Education
Pusan National University
30 Changjundong, Keumjunggu, Busan 46241, Korea
Tel: +82-51-510-2619
Email: bgyang@pusan.ac.kr
Fields of interest: Phonetics, Phonology