A Critical Study on the Accuracy and Reliability of an Automated Analysis on H1-H2 for Korean Phonation Types*

Mi-Ryoung Kim
(Korea Soongsil Cyber University)

Kim, Mi-Ryoung. (2016). A Critical Study on the Accuracy and Reliability of an Automated Analysis on H1-H2 for Korean Phonation Types. *The Linguistic Association of Korea Journal, 24*(4), 103-127. This study examined the amplitude difference between the first harmonic and the second harmonic (H1-H2 or spectral tilt) for Korean phonation types (initial stops) and critically evaluated the accuracy and reliability of an automated analysis based on a PRAAT script-based algorithm (Remijsen, 2014). Speech data were collected from seven speakers of native Seoul Korean. The output was compared to those of the manual outcomes. The results showed that, from both the automated and "by-hand" method, H1-H2 was highest (i.e., positive) for lax stops, higher for aspirated stops, and lowest (i.e., negative) for tense stops, showing a pattern of tense < aspirated < lax stops at vowel onset. However, the automated results neither corresponded to the manual outputs nor were fully reliable. Among individual outcomes, serious errors were observed for some speakers' outcomes in that the H1-H2 ranges for the same aspirated stops were -32 dB to 1 dB and the error rate was above 70%. The results indicate that, including Remijsen's (2014) Praat script-based algorithm employed in this study, any automated analysis on acoustic parameters cannot fully be reliable but must be always accompanied by hand-corrections to enhance the accuracy and reliability.

Key Words: H1-H2, automated, "by-hand" measurement, Korean initial stops

[*] An earlier version of this paper was presented at the 5th International Conference on Phonology and Morphology (Kim and Kang, 2014), and those results appear in the conference proceedings. I thank Jae-Koo Kang and Paul Olejarcuk for their technical assistances on the experiment. I also thank three anonymous reviewers for their insightful comments. All the remaining errors are mine.

1. Introduction

Numerous researchers have described phonation (stops) contrasts across languages (Ladefoged, 1971, 1983; Ladefoged and Maddieson, 1996; Gordon and Ladefoged, 2001; Keating et al., 2011). Ladefoged (1971) suggested that there might be a continuum of phonation types, defined in terms of the aperture between the arytenoid cartilages, ranging from voiceless (further apart), through breathy voiced, to regular, modal voicing, and then on through creaky voice to glottal closure (closest together). This continuum is depicted schematically in (1) (cited from Gordon and Ladefoged, 2001: 1).

(1) Continuum of phonation types

	Most open	◀	—	▶	**Most closed**
Phonation type	Voiceless	Breathy	Modal	Creaky	Glottal closure

The majority of languages employ two points along the phonation continuum in making contrasts: voiced and voiceless sounds. This contrast is particularly common among stop consonants and is exploited in a number of languages, such as English, Japanese, Arabic and Russian. However, Korean is reported to distinguish three types of voiceless stops, so-called lax (or lenis), tense (or fortis), and aspirated stops (cf. Kim, 2000; Kim and Duanmu, 2004; Kim, 2014 for a regular voice-voiceless hypothesis of Korean stops). The different types of stops are associated with different glottal configurations in the production of not only the stops themselves but also the onset of the following vowel. In other words, the voice quality of the vowel is influenced by the preceding consonant. Previous studies show that the voice quality of the vowel is similar to a breathy voice after the lax stop and to a laryngealized or 'pressed' voice after the tense stop (Abberton, 1972; Han, 1998).

Among various acoustic features to make a distinction among phonation types, the amplitude difference between the first harmonic and the second harmonic (H1-H2 or spectral tilt), the degree to which intensity drops off as frequency increases, has been suggested as one of the major acoustic cues measuring breathiness of the speech. It is frequently used to distinguish between breathy and modal voicing. Stevens states, "a greater or positive H1-H2 indicates breathiness of the vowel and a small or negative H1-H2 indicates "pressed" (or creaky) voicing quality (1999: 86)." Thus, a greater H1-H2 would indicate breathiness of the vowel and a smaller or negative H1-H2 would indicate "pressed" voicing quality. For Korean stops, vowels following tense stops had a significantly lower (or negative) H1-H2 value, or more creakier phonation, than vowels following either aspirated or lax stops. In contrast, vowels following aspirated or lax stops had a significantly higher (or positive) H1-H2 value, or breathier phonation, than vowels following tense stops. This has been supported by a number of studies (Ahn, 1999; Kim et al., 2002; Cho et al., 2002; Kang and Guion, 2008; Kim, 2014) as follows:

Ahn (1999) applied a normalized H1-H2 measure, which adjusted for formant frequency effects, to Korean stops. His findings were generally consistent with Kagaya's (1974) aperture data: the normalized H1-H2 difference at vowel onset was larger (i.e., positive) for aspirated and lax stops than for tense stops (i.e., negative). Cho et al. (2002) reports, "all four Seoul speakers make a clear three-way distinction among stops, showing a pattern of tense < aspirated < lax in H1-H2 (negative H1-H2 for tense but positive H1-H2 for aspirated and lax stops)". Similarly, Kim (2014) reports that H1-H2 is greater (i.e., positive) for aspirated and lax stops than for tense (i.e., negative) stops. However, she finds that 11 speakers out of 13 show a H1-H2 merger between lax and aspirated stops, showing a pattern of tense < aspirated = lax stops. While H1-H2 differences distinguish Korean tense stops from aspirated and lax stop, they cannot differentiate aspirated stops from lax stops. Along with a partial or complete merger of voice onset time (VOT), H1-H2 differences between aspirated and lax stops are being neutralized or merged (Kim, 2014). Since Kim (2000) and Silva (2006), numerous studies have reported that Korean stops are undergoing a sound change in that, between aspirated and lax tops due to the fact that there is a partial or complete VOT merger but there is a fundamental frequency (f0) or tonal difference (Kim, 2008, 2011, 2012a, b, c, 2013, 2014, 2015; Kim and Duanmu, 2004: Kang and Guion, 2008; Oh, 2011).

The aforementioned studies on H1-H2 are outcomes mainly from the "by-hand" measurements. In the past decades, phonetic measurements have heavily relied on human judgment, which requires significant labor and makes even large laboratory experiments onerous. They are all done by experimenters' "by-hand" measurements. When data are enormously big, measuring H1 and H2 separately and calculating H1-H2 by hand is a time-consuming and never-ending job. For the convenience and efficiency on large copra, numerous laboratory studies have heavily relied on automated measurements using script-based algorithms. Since introducing automation in Praat, a sound analysis program (Boersma and Weenink, 2013), a large amount of speech data, both from laboratory and naturalistic settings, are becoming increasingly available and easy to construct, and promise to change the questions researchers can ask about human speech production. This promise depends on the development of accurate algorithms to quicken and replace manual measurement, which becomes infeasible for large corpora. However, reliability of this automated script-based measurement has not been clearly discussed yet, nor the comparison between automated and "by-hand" measurement has been made. It is still questionable on whether we can fully rely on automated measurement without any hand-corrections. Shue et al. (2011) directly compare the automated Praat results with the "by-hand" results and show similar outcomes between the two methods. However, since their main concern is to introduce Voicesauce[1], neither detailed discussion on the automated method nor possible errors of automated voiced measurements was provided. Thus, this study aims at evaluating the reliability and accuracy of the automated measurement method, compared to the "by-hand" measurement method, by examining H1-H2 for the three Korean phonation types.

2. Methods

2.1 Participants

Participants' age, gender, dialect, and speech rate, which may have some influences on H1-H2, were controlled. Seven speakers of native Seoul Korean

[1] VoiceSauce is an application, which provides automated voice measurements over audio recordings. It is available in Matlab and freestanding PC versions for free download from http://www.ee.ucla.edu/~spapl/voicesauce/.

participated in this study. There were all females. Their mean age was 25 years and the individuals ranged from 20 to 31 years. Since they were raised and educated in Seoul (i.e., the capital city of Korea), they all spoke the standard Seoul dialect. None of the speakers had any history of speech pathology or phonetic training.

2.2 Speech materials and procedure

Eighteen monosyllabic words were balanced across the three phonation (stop) types (lax, aspirated, and tense), and the three places of articulation (labial, alveolar, and velar) followed by a vowel /a/ context and/or a stop. Thus, the syllable types of the target words were either CV or CVC, where the final consonant C was either [t] or [k] (unreleased stop). All words were real, as presented Table 1.

TABLE 1. Speech materials

	Labial	Alveolar	Velar
Aspirated	/p^hat/'red bean' /p^ha/'to dig'	/t^hat/ 'blame' /t^ha/'to get in'	/k^hat/ 'stop' /k^ha/'car'
Lax	/pat/ 'field' /pa/ 'to see'	/tat/ 'anchor' /ta/ 'all'	/kat/ 'cap' /ka/ 'to go'
tense	/p*ak/ 'head' /p*a/ 'to grind'	/t*ak/ 'precisely' /t*a/ 'to pick'	/k*ak/ 'croak' /k*a/ 'to peel'

Note: /p, t, k/ represents for lax or lenis, /p^h, t^h, k^h/ for aspirated and /p*, t*, k*/ for tense or fortis).

Recordings were made in a quiet office or in a sound-attenuated room directly into a Samsung SENS NT900XC4C-A78 laptop computer. The recordings were digitized at a sampling rate of 22,050 Hz. The target words were recorded in the frame sentence [igə____hasɛjo] "Say this ____" in English. They were presented in Hangeul (the writing system of Korean) three times in a random order. The stimuli were automatically popped up at a 3-second interval for each sentence using a PowerPoint slide show (one sentence per slide). This was able to control speakers' speech rate and tempo. Prior to recording, a short period of familiarization on words and sentences was given. A total of 378 tokens (18 words x 7 speakers x 3 repetitions) were obtained. They were measured automatically and manually. All measurements were done by Jae-Koo Kang, an assistant, and confirmed by the researcher. All utterances were recorded and analyzed using Praat 5.3.47, a speech analysis program (Boersma and Weenink, 2013).

2.3 Acoustic measurements on H1-H2

For a comparative purpose, energy values (dB) for the first (H1) and second (H2) harmonics were taken at vowel onset and midpoint. The onset of the vowel was defined as the first and periodic pulse of a vocalic waveform that shows features typical of a vowel. The midpoint of the vowel was defined as a half-distance between the onset and the offset of the vowel. For each position, the differences between H1 and H2 were obtained by two different methods: (i) by the "by-hand" method, using FFT spectra with a 25 ms window (40Hz bandwidth) and (ii) by the automated method, using a Praat script-based algorithm. Unfortunately, Praat cannot reliably tell where one word starts and where another ends. Neither can it find the specific segment we are looking for, nor identify the vowel in the word. As such, we often need to segment sound files with that information when using any sort of automated measurements. In Praat, this is done by creating TextGrid annotations in a TextGrid file, which is saved separately from the sound itself. TextGrid annotations are composed of different tiers which mark either intervals or specific points within the sound file. Note that annotating sound files is not automatic but purely by-hand. With annotating files, therefore, we cannot say that there is an exclusively automatic way to measure any acoustic parameters (see Styler, 2015 for TextGrid annotation).

In the present study, for all sound files, a segmental tier was created to annotate or label the target words into consonants (C) and vowels (V) on a tier according to the phonation type (e.g., tense=t*, lax=t, aspirated=t^h), the place of articulation (i.e., labials, alveolars, velars), and the coda type (CV or CVt/CVk). The aperiodic "consonantal portion" consisted of a stop closure, the release burst, plus any aspiration. The periodic "vowel portion" consisted of the vowel; vowel onset was taken to be the first periodic higher-amplitude pulse after consonant release/aspiration. Segmentation was done at the zero crossing in the waveform displays, but was verified with simultaneous spectrographic displays. Figure1 gives representative spectrograms and waveforms illustrating labeled segmentation.

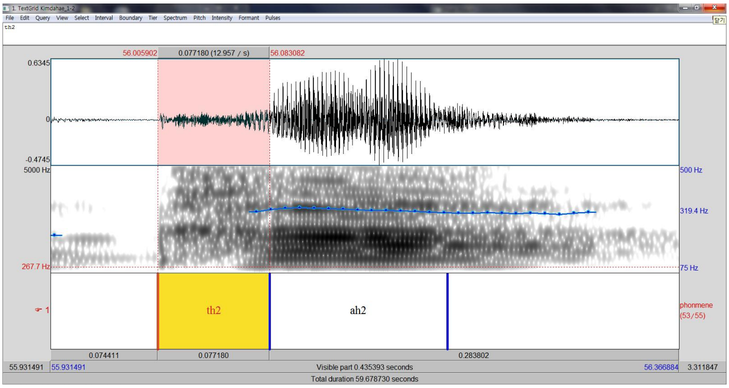

FIGURE 1. Spectrographic and waveform displays for one of aspirated alveolar stop tokens. Three vertical lines indicate, from left to right, consonant onset, consonant offset/vowel onset, and vowel offset. A TextGrid annotation indicates "tha2" and "ah2" on a segment or phoneme tier for the second token of the word tha 'to ride'.

In Figure 1, the Praat TextGrid Editor Window shows the wave form of the sound (at the top), a broadband spectrogram showing the spectral energy of the sound over time (in the middle), and a segmental tier (at the bottom) representing a consonantal and vocalic portion pointed with lines on a tier. Once all of the files are TextGridded, we will be a much better position to start measuring data automatically or manually. Thus, this step is one of the most important procedures and it takes long time to finish annotating sound files. Once finish annotating, H1 and H2 can be obtained for each labeled interval on a tier. The "by-hand" and automated methods are discussed in section 2.3.1 and 2.3.2, respectively.

2.3.1 Manual (or "by-hand") measurements

The "by-hand" measurements were made in Praat, from FFT spectra made with a 21 Hz bandwidth and a 40 ms Hamming window, positioned immediately after vowel onset, so covering about the first third of the vowels. To obtain H1 and H2 each manually, the general procedures written by Styler (2015: 23) are as follows: Firstly, to take a spectral slice properly, set Window Length to "0.025" (effectively producing a narrow-band spectrogram) and Window Shape to "Hamming. In order to get specific details about the frequencies and individual harmonics in a sound at a given moment in time, spectrogram setting should be adjusted properly because examining a narrowband spectrogram alone do not provide sufficient information. Secondly, to find the amplitude and frequency of a given point in the spectrum, click the point and read off the amplitude (on the left) and the frequency (at the top), as shown in Figure 2.

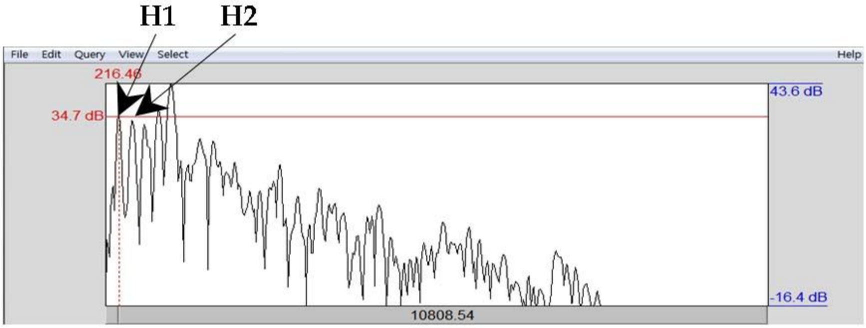

FIGURE 2. A spectral slice Editor window

In Figure 2, the amplitudes of the first and second harmonics were manually marked and logged using a cursor, as in several previous studies of voice quality in the literature. This method is one of standard practices. According to Styler (2015), "getting harmonic frequency and amplitude in a spectrum Editor window in Figure 2 is fairly straightforward, as clicking anywhere within the spectrum editor window will give you the frequency and amplitude measurements at the cursor." However, determining H1 from FFT spectra is not always straightforward. There sometimes needs a researcher's subjective judgement on which peak is considered as the first one or H1. As seen in Figure 2, there are a couple of peaks before the H1 peak, which are relatively very low, it is ok to ignore them. However, it is sometimes hard to determine whether it is ignorable or not. Therefore, we can say that this method is not absolutely straight but rather needs a researcher's judgement. Owing to this, the manual measurements were done twice not only an assistant but also the researcher. Note that many of the files could not be analyzed by this method because the FFT did not show a clear harmonic structure. In addition, measuring H1-H2 using FFT spectra with this manual method takes long time and much effort. Some problems with the "by-hand" measurement can be easily and efficiently overcome by the automated measurements, allowing us to find the highest point on a given harmonic without clicking guesswork[2].

2.3.2 Automated measurements

Under the automated methods, almost all acoustic parameters such as VOT, f0, intensity, vowel duration, and so on can be easily and accurately measured, using a Praat script-based algorithm (Boersma and Weenink, 2013; Remijsen, 2014; Styler, 2015; Kim, 2014). Praat scripting is a wonderful tool for situations where we find ourselves repetitively doing the same tasks over and over again, and with increased sophistication, we can have Praat make simple decisions based on its measurements and changes. Simply put, Praat scripting is a way to use the computer as a co-pilot, handling the boring, repetitive tasks independently and allowing us to do our job more efficiently. A Praat script is literally just a text file with a series of commands to Praat. Thus, a script can be either created all by ourselves or cited from the web resources. For automatic speech analyses, numerous Praat scripts have already been written and circulated on the internet (UCLA Praat script resources, 2016). However, many scripts are not absolutely perfect but may result in some fatal errors. Errors might mainly be due to the fact that artificical intelligence has no brain to think.

For the present study, H1-H2 values were obtained using a scripted-based algorithm originally written by Chad Vicenik (Remijsen, 2014) and modified by the researcher and Paul Olejarcuk[3]. The script employed in this study is one of

[2] One anonymous reviewer commented that it might be worthy of discussing what the reasons of the errors from automated measurement on H1-H2 can be. As you see FFT spectra in Figure 2, the researcher can manually ignore abnormal peaks before the H1 peak while the praat algorithm cannot. This might be one of the main reasons why unreliable H1-H2 values are frequently obtained from automated measurements.

the widely-used scripted algorithms when researchers measure voice quality. Modifications were not great but trifle to run the script properly for the current speech data. For example, sound files and textGrid directory were renamed and the commands of chunks were newly designated from default to six. Depending on the number of chunks, the values on acoustic parameter were obtained at the position. Due to this command, the algorithm produced six H1-H2 measurements over a vowel portion. Among six measurement points, only the first point values were taken for the onset of the vowel and the fourth point values were taken for the midpoint of the vowel. Once the scripted-based algorithm is working properly, just run and wait for the values. After running a script successfully, H1-H2 values were automatically created as a textfile.

The automated Praat measures are neither pitch-synchronous nor corrected for formants. With the script, a file cannot be analyzed or is discarded by the script. There will be no measurement if Praat cannot detect an F0 and all three formants. This indicates that errors associated with the algorithm may often occur.

2.4. Statistical Analysis

The differences between H1 and H2 were statistically tested using repeated measures analysis of variance (ANOVA) in the context of a general linear model (GLM; SPSS/PASW, 2012). Repeated measures ANOVAs include "between" subjects effects (i.e., seven individuals and two measurement types) and "within" subject effects (i.e., three phonation types – aspirated, lax, and tense and three place types – labials, alveolars, velars). Acoustic correlates such as H1 and H2 and their measurement points (i.e., vowel onset and midpoint) were dependent variables. Their main and interaction effects were statistically analyzed at a 0.05 significance level.

Post hoc Tukey HSD multiple tests were also run to answer the questions: (i) for H1-H2 on each position (vowel onset and midpoint), whether any differences between the automated and manual measurement were significant, (ii) for any H1-H2 on each position, whether any differences in pairs among the phonation

[3] Paul Olejarcuk was a PhD student in the field of Phonetics at the University of Oregon at the time of the research.

and place type were significant and (iii) for H1-H2 on each position, whether any differences in pairs among the phonation and place type were significant. For statistical analysis, the effects of most central interest – H1-H2 differences between the two methods and among the three phonation types at each point – are mainly discussed because they are the main concerns of the present paper. Statistical outcomes that are not interested in the current study are omitted for discussion.

3. Results

In this section, the manual and automated H1-H2 results for three Korean phonation types are presented at vowel onset and midpoint for the pooled and individual data respectively. The "by-hand" results are given first and they are compared with previous findings in section 3.1. The automated results are given in section 3.2. Their output was directly compared to the manual results in section 3.3. The comparisons were further discussed in detail for one speaker, who showed a big discrepancy between the two measurements.

3.1 Manual (or "by-hand) H1-H2

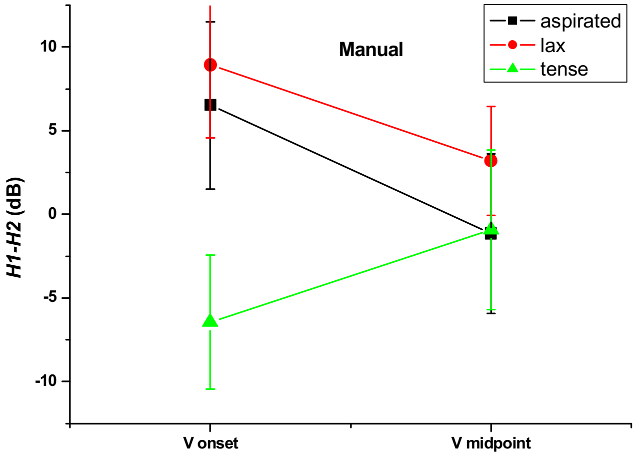

FIGURE 3. Pooled results: H1–H2 (dB) (±1 SD) values measured manually at vowel onset and midpoint for lax, aspirated, and tense stops. Data were averaged across seven Seoul Korean female speakers (SD = Standard Deviation).

Figure 3 gives mean H1-H2 values measured by hand at vowel onset and midpoint for each phonation type for the pooled data. Results of repeated measures ANOVAs showed that, at vowel onset, there was a main effect of phonation type (F(2, 357) = 428.547, $p < 0.0001$) in that H1-H2 was highest (8.933 dB) for lax stops, higher (6.529 dB) for aspirated stops, and lowest (-6.437 dB) for tense stops, as shown in Figure 3. Post hoc Tukey HSD multiple comparisons revealed that stops significantly differed from each other, showing a pattern of tense < aspirated < lax in H1-H2. Among the three phonation types, the differences between tense and aspirated or lax stops were much greater than those between aspirated and lax counterparts. The statistical results indicate that, at vowel onset, H1-H2 can still distinguish the three phonation types in Korean, as reported in previous findings (Cho et al., 2002; Kim et al., 2002; Kang and Guion, 2008; cf. Ahn, 1999).

At vowel midpoint, there was also a main effect of phonation type (F(2, 357) = 40.331, $p < 0.001$) in that H1-H2 was higher (3.204 dB) for lax stops than aspirated (-1.143 dB) and tense (-0.908 dB) stops. Post hoc Tukey HSD multiple comparisons revealed that lax stops differed significantly from either aspirated or tense stops ($p < 0.001$) whereas tense stops did not differ from aspirated stops ($p > 0.05$), showing a pattern of tense = aspirated < lax. The results indicate that, at vowel midpoint, H1-H2 cannot distinguish the three phonation types, showing a merger between aspirated and tense stops. The H1-H2 differences at vowel onset were expected to disappear in the middle of a vowel because the voice quality becomes to be normal. Although the differences become smaller, the voice quality of the vowel (i.e., breathiness) after the lax stop still maintained till midpoint.

Figure 4 gives the breakdown of individual speaker's results according to the phonation type.

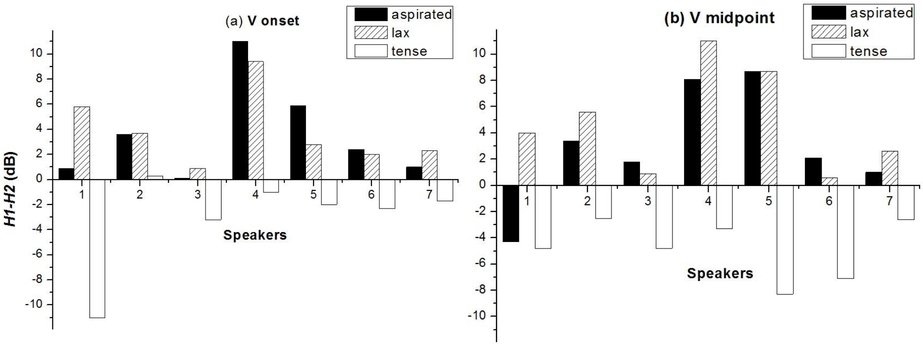

FIGURE 4. Individual results: mean H1-H2 (dB) values measured manually at (a) vowel onset and (b) midpoint for lax, aspirated, and tense stops.

At vowel onset, there was a main effect of speakers on H1-H2 (F(6, 357) = 30.443, p < 0.0001) in that H1-H2s significantly differed from each other among speakers. Post hoc Tukey HSD multiple comparisons revealed that some speakers showed a corresponding pattern but others did not. For example, speaker 1, 3, and 5 had a statistical pattern of tense < aspirated < lax, while speaker 2, 4, 6, and 7 had a pattern of tense < aspirated = lax (p < 0.001). At vowel midpoint, there was also a main effect of speakers on H1-H2 (F(6, 357) = 42.559, p < 0.0001) in that H1-H2 was different from each other among speakers. Similar to the vowel onset, the statistical groupings of the three phonation types at the midpoint were also different among speakers.

There were also interspeaker variations for the presence or absence of breathiness following lax stops. For speaker 1, 2, 4, and 5, breathiness persists throughout the vowel whereas, for speaker 6, it does not. In order to compare the breathy voice with the modal voice, Figure 5 gives representative spectrograms and waveforms illustrating the (a) modal and (b) breathy voice produced by two different speakers. As seen in Figure 5(b), spectrographic and waveform displays for the breathy [h] are characterized by higher formants and a fair amount of noisy energy which contributes a relatively jagged appearance to the waveform and diminishes the clarity of individual pitch pulses, as consistent with findings reported in the literature (Ladefoged, 1971, 1983; Gordon and Ladefoged, 2001). In Figure 5(a), in comparison, the modal voice after /t/ is not marked by this turbulence and has relatively well-defined pitch pulses. One of the more salient features differentiating modal and breathy voice in the spectrograms is the visually well-defined stop-to-vowel transition characteristic of the modal voice but not the breathy voice ([h] at about 35 ms after the lax stop /t/).

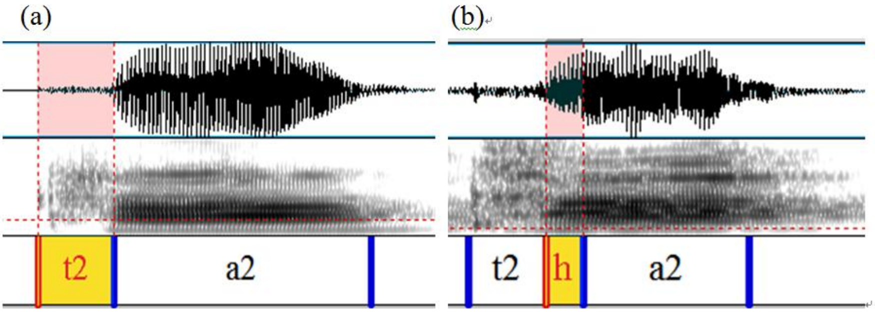

FIGURE 5. Spectrographic and waveform displays of (a) modal and (b) breathy stops for the same Korean word /ta/ 'all' uttered by the two different speakers.

3.2 Automated H1-H2

For the automated pooled data, Figure 6 shows mean H1-H2 values at the onset and the midpoint of the vowel for each phonation type.

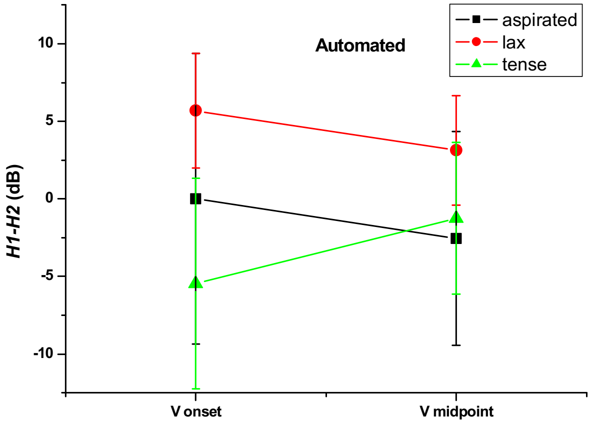

FIGURE 6. Pooled results: H1-H2 (dB) (±1 SD) values measured automatically at vowel onset and midpoint for lax, aspirated, and tense stops.

Results of repeated measures ANOVAs showed that, at vowel onset, there was a main effect of phonation type (F(2, 357) = 49.075, p < 0.0001) in that H1-H2 was greatest (5.6824 dB) for lax stops, intermediate (0.016249 dB) for aspirated stops, and smallest (-5.4557 dB) for tense stops. There was neither place effect nor interaction between place and phonation type (p > 0.05). Post hoc Tukey HSD multiple comparisons revealed that stops significantly differed from each other, showing a pattern of tense < aspirated < lax in H1-H2 (p < 0.0001). At vowel midpoint, there was also a main effect of phonation type (F(2, 357) = 29.544, p < 0.001) in that H1-H2 was greatest (3.1512 dB) for lax stops, intermediate (-1.247 dB) for tense stops, and smallest (-2.557 dB) for aspirated stops. Post hoc Tukey HSD multiple comparisons revealed that lax stops differed significantly from either aspirated or tense stops (p< 0.001) whereas tense stops did not differ from aspirated stops (p > 0.05), showing a pattern of tense = aspirated < lax. Individual H1-H2 differences for each phonation type are illustrated in Figure 7.

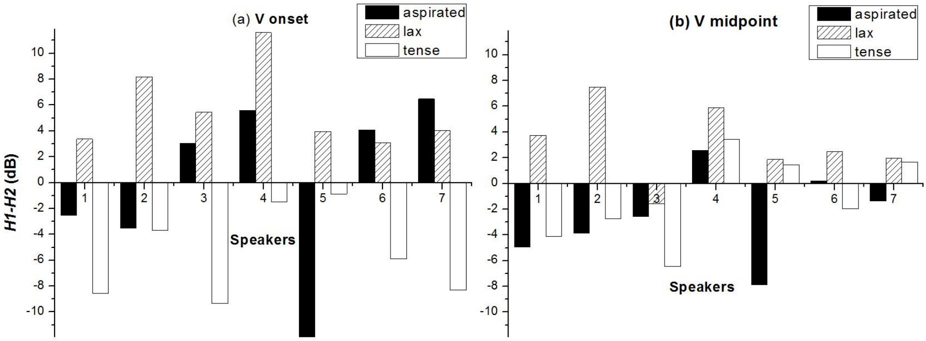

FIGURE 7. Individual results: mean H1–H2 (dB) values measured automatically at (a) vowel onset and (b) midpoint for lax, aspirated, and tense stops.

At vowel onset, there was a main effect of speakers on H1-H2 (F(6, 357) = 20.223, p < 0.0001) in that H1-H2 significantly differed from each other. As shown in Figure 4, the patterns among the three phonation types are very different across speakers. All speakers but 2 and 6 had three statistical H1-H2 groupings whereas Speaker 3, 4, 6, and 7 had two (tense < aspirated = lax) (p < 0.001). Speaker 2 and 6 showed a merger between aspirated and tenses or lax stops. Similarly, at vowel midpoint, there was a main effect of speakers on H1-H2 (F(6, 357) = 22.449, p < 0.0001). Most of speakers had two to three statistical H1-H2 groupings but their patterns in size were quite different. Big discrepancies among speakers were found for speaker 5's aspirated stop.

3.3 Automated vs. "by-hand" H1-H2

For direct comparison, the automated and "by-hand" results are together re-illustrated with the bar graphs in Figure 8. The automated and "by-hand" H1-H2 differences for each phonation type are presented at vowel onset and midpoint measured.

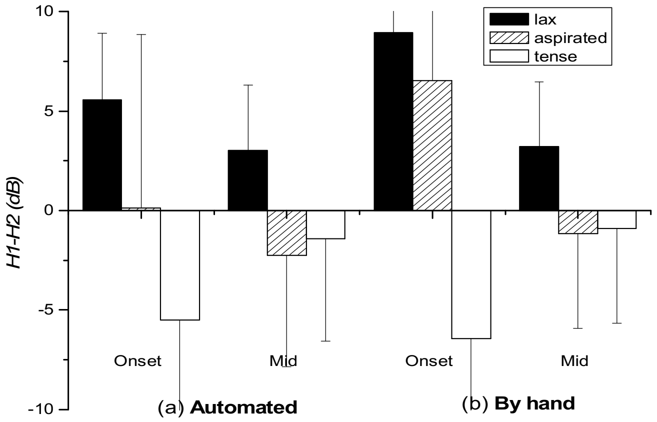

FIGURE 8. Mean H1-H2 (dB) values (±1 SD) measured (a) automatically and (b) "by-hand" at the onset and midpoint for vowels after the three phonation types. Colored bars distinguish the categories of stops. Error bars represent standard deviations.

Overall, the two methods were very similar in terms of statistical outcomes and their size. For the two methods, there was a main effect of phonation type at vowel onset (p > 0.0001) in that H1-H2 was greatest (5.570 dB vs. 8.933 dB) for lax stops, intermediate (0.0129 dB vs. 6.529 dB) for aspirated stops, and smallest (-5.504 dB vs. -5.971 dB) for tense stops. Post hoc Tukey HSD multiple comparisons revealed that, for each method, stops significantly differed from each other, showing a pattern of tense < aspirated < lax in H1-H2 (p < 0.0001).

The pattern corresponds to the midpoint, as presented in Table 2. Table 2 summarizes the statistical grouping of the three stops according to size.

TABLE 2. Statistical subsets of the three stops according to size

	Onset	Midpoint
Automated	tense<aspirated<lax	aspirated<tense<lax
By-hand	tense<aspirated<lax	aspirated<tense<lax

The two methods were statistically not different from each other. Results of repeated measures ANOVAs showed that, at midpoint, there was no main effect of method (F(1, 754) = 2.644, p > 0.05) in that automated H1-H2 was not significantly different from "by-hand" H1-H2. At vowel onset, however, there was a main effect of method (F(1, 754) = 24.854, p < 0.0001) in that automated H1-H2 was significantly different from "by-hand" H1-H2. The automated measurements show greater mean differences between lax and aspirated stops than those of the "by-hand" automated measurements. The automated measurements also show greater within-stop variability than those of the "by-hand" method. Among the three phonation types, the aspirated stop was the one that showed the greatest variability. This is presented in Table 3.

TABLE 3. Mean and the ranges of aspirated stops at vowel onset

	mean		mim.~max	
Speaker	Auto	By-hand	Auto	By-hand
1	−2.85	0.23	−4.9~1.0	−3.6~4.1
2	−3.36	1.85	−6.9~2.4	−11.6~8.6
3	3.08	9.93	−0.5~6.2	6.4~13.5
4	4.63	9.6	1.6~9.2	4.7~16.6
5	**−10.1**	**4.6**	**−25.7~7.3**	**−1.2~13.6**
6	3.52	9.3	−3.0~9.1	6.2~12.5
7	5.94	10.2	**−29~10.5**	**6.4~17.4**
Mean	0.13	6.53	−29~10.5	−11.6~17.4

Vowels following aspirated or lax stops have significantly higher (or positive) H1-H2 values, or breathier phonation, than vowels following tense stops (Ahn, 1999; Kim et al., 2002; Cho et al., 2002; Kim, 2014). That indicates that, if vowels following aspirated stops have highly negative H1-H2 values, they would be considered as errors. From the automated methods, negative H1-H2 values are observed for speaker 1, 2, and 5. Their errors are relatively higher, compared to the "by-hand" output. The great variability can be observed from the automated method, as presented in Table 3.

For speaker 5, mean H1-H2 for the aspirated stop is -10.1 dB and the range is from -25.7 dB to 7.3 dB. Fourteen out of 18 words were produced with negative H1-H2 for the aspirated stop. For speaker 7, the range of H1-H2 is from -29 dB to 10.5 dB. However, only one out of 19 was produced with negative H1-H2. The error rate of speaker 5 and 7 was 77% and 5.6%, respectively. Although the error rate was very low for speaker 7, -29 dB was enough to lower overall mean H1-H2 for aspirated stops and produced unreliable results. Let's consider speaker 5's results in more detail who showed the most discrepancy among seven speakers. Figure 9 shows manual and automated H1-H2 at (a) vowel onset and (b) midpoint for the three phonation types. The big discrepancy between the two methods is noticeable both at vowel onset and midpoint, as seen in Figure 10.

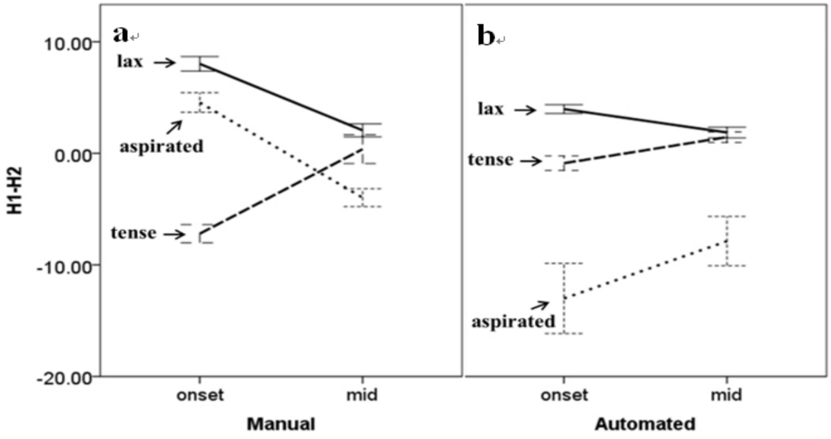

FIGURE 9 (a) Manual vs. (b) Automated H1-H2 values (±1 SEM) at vowel onset and midpoint according to lax, aspirated, and tense stops for speaker 5 (SEM = Standard Errors of Mean).

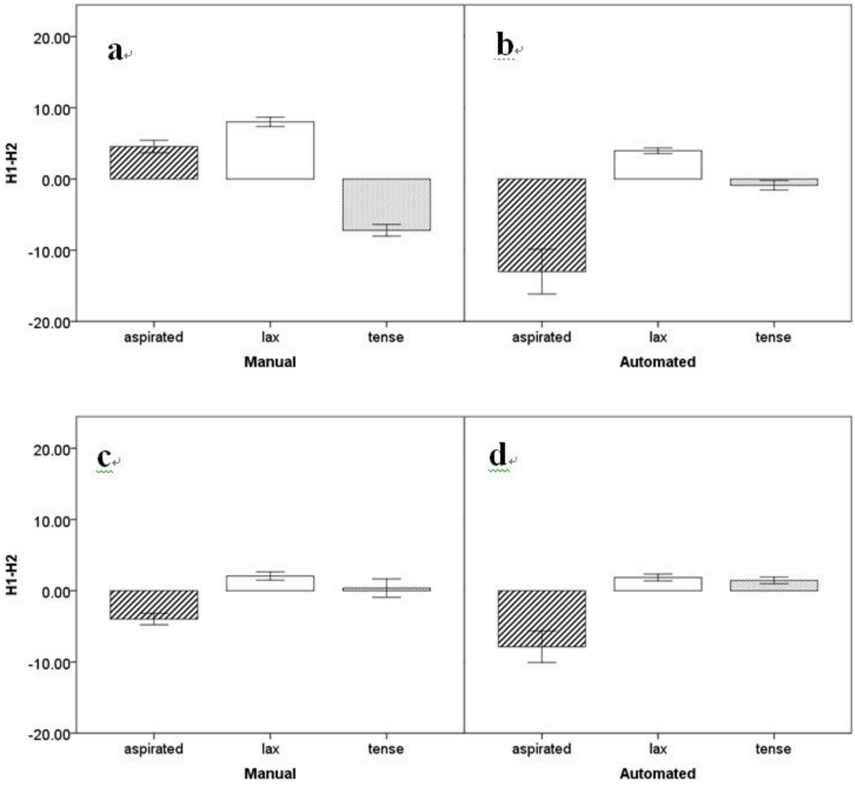

FIGURE 10. Manual and automated H1-H2 at vowel onset (a, b) and vowel midpoint in (c, d).

From some speakers' unusual results on H1-H2, it is clear that the automated methods have serious errors because they unconsciously measure abnormal peaks without taking any consideration (see footnote 2). It is possible to change the mean of the pooled data overall, and they are hard to be adjusted without hand corrections. Compared to automated errors, by-hand errors were acceptable not to change the mean of the pooled data.

4. Summary and Discussion

This study investigated H1-H2 to evaluate the accuracy and reliability of the automated method using a Praat script-based algorithm (Remijsen, 2014). For the three Korean phonation types, H1-H2 was measured at vowel onset and midpoint. The outputs of the automated method were compared with those of the "by-hand" method. The "by-hand" results on H1-H2 are summarized with major findings, compared with previous findings as follows: Firstly, for the pooled data, H1-H2 can statistically distinguish the three phonation types in that H1-H2 was highest (8.9 dB) for the lax stop, higher (6.5 dB) for the aspirated stop, and lowest (-6.4 dB) for the tense stop. The pattern of tense < aspirated < lax stop corresponds to early findings (Cho et al., 2002; Kim et al., 2002; Kang and Guion, 2008; Kim, 2014; cf. Ahn, 1999). As expected, H1-H2 values for tense stops are almost always negative and the lowest, indicating that they carry some creakiness with small glottal constriction. In contrast, vowels following lax stops carry some breathy voice because positive H1-H2 values at vowel onset maintained till midpoint. Secondly, similar to VOT, H1-H2 also showed a merger between aspirated and lax stops. Considering the statistical outcomes of the individual data (see Figure 4), five out of seven female Seoul speakers showed a H1-H2 merger, showing a pattern of tense < lax = aspirated. The H1-H2 merger between aspirated and lax stops corresponds to the VOT merger, reported in early findings (Silva, 2006; Kim, 2014). The results suggest that H1-H2 is another indicator to show that Korean stops are undergoing a sound change along with VOT and f0.

From the automated results, only significant findings are summarized as follows: Firstly, at vowel onset, the automated results did not correspond to the "by hand" counterparts. Although H1-H2 was highest for lax stops, intermediate for aspirated stops, and lowest for tense stops, showing the similar pattern of tense < aspirated < lax stops as that of the manual output, their overall H1-H2 values across the phonation types were statistically greater than the "by hand" results. At vowel midpoint, however, the automated results were not different from the manual counterparts, indicating that the automated methods might be more reliable and accurate in the middle of the vowel than in the beginning of the vowel. This can be predictable because the differences in voice quality are due to different consonantal properties, they disappear in the middle of the vowel.

Secondly, the automated results were neither consistent with physical properties nor supported well by previous findings. Among the three phonation types, the aspirated stop showed the biggest discrepancy between the automated and manual results on H1-H2. At vowel onset, the manual H1-H2 values were positive (+4.5 dB) whereas the automated H1-H2 values were not (0 dB). Considering that there is a wide glottal opening, it is expected for aspirated stops to have positive H1-H2 values. In addition, it is natural that aspirated stops are patterned with lax stops because of long voicing lag (VOT). In contrast to aspirated stops, tense stops show the opposite pattern: -6.43 dB under the manual results but -5.46 dB under the automated results. In addition, automated H1-H2 for the smallest value of the aspirated stop (-25.7 dB) is even much lower for tense stop (-5.13 dB). From early findings (Cho et al., 2002; Ahn, 1999, Kim et al., 2002; Kang and Guion, 2006; Kim, 2014), tense stops are expected to produce with creaky voicing whereas aspirated stops are expected to produce with breathy voicing at vowel onset. Taking the early findings into consideration, the automated results for both aspirated and tense stops do not provide absolutely correct and reliable values. Comparing the results at vowel midpoint with those at vowel onset, the reliability of the automated measures seems slightly higher and more stable in that, unlike vowel onset, the automated measures show the similar pattern with the manual measures across the three categories: H1-H2 is greatest for lax stops, intermediate for tense stops, and smallest (negative for the automated measures but positive for the manual measures) for aspirated stops. However, H1-H2 for aspirated stops is much greater from the manual measures than from the automated ones. Note that, even at midpoint, H1-H2 was found to be negative for aspirated stops from the automated measurement. This again casts doubt on the strong creakiness of aspirated stops. In addition, it is natural for all three phonation types to be close together since the voice quality difference becomes smaller at the midpoint.

Thirdly, the errors from the automated measures were so big that they could make the overall results change. For example, due to speaker 5's errors, the overall results of the aspirated stops became low or negative (Figure 9). Serious errors might be due to the fact that the script-based algorithm randomly takes H1s which need to be ignored as abnormally lower peaks (see footnote 2). Thus, the automated measurements bring more questions into the reliability of the method. Comparing the present results with early findings on H1-H2, the automated results can be only reliable when accompanied by hand corrections because they fully rely on the script-based algorithm without any caution.

Overall, the automated results seem to be very similar to the manual results in the pool data, as discussed in Shue et al. (2011). However, taking the individual data into a deep consideration, they are remarkably different from the manual results. The differences are the greatest at vowel onset for aspirated stops. The "by-hand" measurements show smaller mean differences across the three categories than the automated measurements. The automated measurements also show much greater within category variability than the manual measurements. The greater variability is due to greater variability of both H1 and H2 separately. The results indicate that the reliability and accuracy of the automated measurements are questionable. As described in introduction and method, the automated measurement in this study is based on Remijsen's Praat script retrieved online in 2014. For future studies, it may or may not be possible to simply fix the inaccuracy with some corrections in the script.

The current results has three implications as follows: Firstly, regardless of the measurement type, the researcher him- or herself must be fully responsible for any kinds of errors. With large amount of data, it is always possible to have errors. Even for the "by hand" method, there could be subjective errors to make a decision picking up H1. Secondly, due to the incorrectness and unreliability of the automated method, it always needs to be accompanied by hand corrections. Thirdly, the fact that H1-H2 cannot distinguish aspirated from lax stops supports that Korean is undergoing sound change in terms of H1-H2.

It is highly expected that more and more studies will rely on the automated method to measure acoustic parameters to take care of a large amount of speech data quickly and accurately. This is due to the fact that automation have lots of advantage by not only reducing the labor time of a researcher but also removing the researchers' errors. However, the large use of automation heavily depends on the development of accurate algorithms to quicken and replace manual method. The present study will serve a substantial reference for future research bearing not only on the issues of the accuracy and reliability of the automated method, but also on the issues of the three phonation types in Korean.

References

Abberton, E. (1972). Some la ryngographic data for Korean stops. *Journal of International Phonetic Association*, 2, 67-78.

Ahn H.-K. (1999). *Post-release phonatory processes in English and Korean: Acoustic correlates and implications for Korean phonology*. Unpublished doctoral dissertation. University of Texas at Austin.

Boersma, P. and Weenink, D. (2013). *Praat: Doing phonetics by computer* (Version 5.3.47) [computer program]. Retrieved April 23, from http://www.fon.hum.uva.nl/praat/.

Cho, T., Jun, S.-A., and Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, 30, 193-228.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29, 383-406.

Han, N. (1998). A comparative acoustic study of Korean by native Korean children and Korean-American children. Unpublished Master dissertation. UCLA.

Kagaya, R. (1974). A fiberscopic and acoustical study of the Korean stops, affricates, and fricatives. *Journal of Phonetics*, 2, 161-180.

Kang, K., and Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of Acoustical Society of America*, *124*(6), 3909-3917.

Keating, P., Esposito, C., Garellek, S. K., and Kuang, J. (2011). Phonation contrasts across languages. *UCLA Working Papers in Phonetics*, 108, 188-202.

Kim, M.-R. (2000). *Segmental and tonal interactions in English and Korean: A phonetic and phonological study*. Unpublished doctoral dissertation, The University of Michigan.

Kim, M.-R. (2008). Lax stops in Korean revisited: VOT neutralization. *Studies in Phonetics, Phonology and Morphology*, *14*(2), 3-20.

Kim, M.-R. (2011). The relationship between cross-language phonetic influences and L2 proficiency in terms of VOT. *Speech Sciences*, *3*(3), 3-11.

Kim, M.-R. (2012a). Tonogenesis in Korean: Some recent speculations on the sound change. *Korea Journal of Linguistics*, *37*(2), 243-283.

Kim, M.-R. (2012b). The [voice] system of Korean stops revisited with special reference to the aspirated-lax merger. *Studies in Phonetics, Phonology and*

Morphology, 18(2), 211-243.

Kim, M.-R. (2012c). L1-L2 transfer in VOT and f0 production by Korean English learners: L1 sound change and L2 stop production. *Speech Sciences, 4*(3), 31-41.

Kim, M.-R. (2013). Tonogenesis in contemporary Korean with special reference to the onset-tone interaction and the loss of a consonantal opposition. Paper presented at the 21st international congress on Acoustics and 165th meeting of the acoustical society of America, June 2-7, Montreal, Canada.

Kim, M.-R. (2014). Ongoing sound change in the stop system of Korea: A three-to two-way categorization. *Studies in Phonetics, Phonology, and Morphology,* 14(2), 185-203.

Kim, M.-R. (2015). A study of L1 and L2 influences on the speech of Korean-English bilinguals: With special reference to VOT and F0 (written in Korean). *Speech Sciences,* 7(3), 13-26.

Kim, M.-R., Beddor, P., and Horrocks, J. (2002). The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics, 30*(1), 77-100.

Kim, M.-R., and Duanmu, S. (2004). Tense and lax stops in Korean. *Journal of East Asian Linguistics,* 13, 59-104.

Kim, M.-R., and Kang, J. (2014). A comparative analysis of automated and manual measurement on H1-H2 for three Korea stops. In *Proceedings of the 5th International Conference on Phonology and Morphology.* 259-262.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics.* Chicago: University of Chicago.

Ladefoged, P. (1983). The linguistic use of different phonation types. In D. Bless & Abbs (Eds.) *Vocal fold physiology: Contemporary research and clinical issues* (pp. 351-360). San Diego: College Hill Press.

Ladefoged, P., and Maddieson, I. (1996). *The sounds of the world's languages.* Oxford: Blackwell Publishers.

Oh. E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics,* 39, 59-67.

Remijsen, B. (2014). Praat script resources: phonation measurements written by Chad Vicenik under analysis of sounds using text grids. Retrieved May 25, 2014, from http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, 23, 287-308.

Shue, Y.-L., Keating, P., Vicenik C., and Yu K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of ICPhs XVII International Congress of Phonetics Sciences*, 1846-1849.

SPSS/PASW statistics. (2012). IBM SPSS Statistics for Windows. Version 21.0. Armonk, NY: IBM Corp.

Stevens, K. N. (1999). *Acoustic Phonetics*. Cambridge: MIT Press.

Styler, W. (2015). Using Praat for Linguistic Research (version 1.6). Retrieved July 5th, 2015, from http://savethevowels.org/praat/.

UCLA Praat Script Resources. (2016). Retrieved September 02, 2016, from http://www.linguistics.ucla.edu/facitilites/acoustic/praat.html.

Mi-Ryoung Kim
Department of Practical English
Korea Soongsil Cyber University
Jongno Biz-Well 23, Samil-daero 30-gil, Jongno-Gu
Seoul 110-340, Korea
Phone: 82-2-708-7845
Email: kmrg@mail.kcu.ac

Received on September 28, 2016
Revised version received on December 21, 2016
Accepted on December 30, 2016