Cross-Linguistic Perceptual Categorization of the Three Corner Vowels: Effects of Listener Language and Talker Age

Hyunju Chung  
Louisiana State University, USA

Benjamin Munson  
University of Minnesota, USA

Jan Edwards  
University of Maryland, USA

Abstract
The present study examined the center and size of naïve adult listeners’ vowel perceptual space (VPS) in relation to listener language (LL) and talker age (TA). Adult listeners of three different first languages, American English, Greek, and Korean, categorized and rated the goodness of different vowels produced by 2-year-olds and 5-year-olds and adult speakers of those languages, and speakers of Cantonese and Japanese. The center (i.e., mean first and second formant frequencies (F1 and F2)) and size (i.e., area in the F1/F2 space) of VPSs that were categorized either into /a/, /i/, or /u/ were calculated for each LL and TA group. All center and size calculations were weighted by the goodness rating of each stimulus. The F1 and F2 values of the vowel category (VC) centers differed significantly by LL and TA. These effects were qualitatively different for the three vowel categories: English listeners had different /a/ and /u/ centers than Greek and Korean listeners. The size of VPSs did not differ significantly by LL, but did differ by TA and VCs: Greek and Korean listeners had larger vowel spaces when perceiving vowels produced by 2-year-olds than by 5-year-olds or adults, and English listeners had larger vowel spaces for /a/ than /i/ or /u/. Findings indicate that vowel perceptual categories of listeners varied by the nature of their native vowel system, and were sensitive to TA.

Keywords
Cross-linguistic study, vowel perception, child speech

Corresponding author:  
Hyunju Chung, Department of Communication Sciences & Disorders, Louisiana State University, 81 Hatcher Hall, Field House Drive, Baton Rouge, LA 70803, USA.  
Email: hchung@lsu.edu
Introduction

The world’s languages differ widely in their vowel systems. Languages differ not only in the number of vowels that are lexically contrastive, but in the acoustic realization of vowels that are labeled with the same phonemic category. That is, each language uses the possible multidimensional articulatory and perceptual vowel spaces in a unique way. The mismatch between the “labeled” vowel phonemic categories and phonetic realization of vowels across languages has been well documented. For example, while Greek and Japanese have the same number of vowel phonemes (/i, a, e, o, u/), which are distributed similarly in the height and front–backness dimensions, the acoustic characteristics of a given Greek and Japanese vowel differ by their native vowel system. While Greek /u/ has consistently low second formant (F2) values, located toward the region of the acoustic vowel space that reflects a more back position of the tongue, the corresponding /u/ vowel of Japanese, which is often described as /ɯ/ (e.g., Keating & Huffman, 1984), has higher F2 values, located in the region that reflects a more fronted position of the tongue. In addition, although the same phonemic symbols are used to denote the low vowel /a/, Japanese /a/ has more consistently low F2 than the corresponding vowel in Greek (Chung et al., 2012).

Thus, the task of perceiving vowels of different languages is complex, as it involves an alignment of vowel systems that differ both in the number and acoustic characteristics of these elements. Moreover, factors such as talker dialect, sex, or age induce additional “lack of invariance” (Liberman et al., 1967) to the speech signal, making the mechanisms of vowel perception more complex. This issue of “lack of invariance” in the speech signal and listeners’ perception of these signals has been studied extensively in the literature, especially in the fields of second language acquisition and sociolinguistics. First, in the areas of second language acquisition, the Perceptual Assimilation Model (PAM) (e.g., Best, 1995; Best & Tyler, 2007) and the Speech Learning Model (SLM) (e.g., Flege, 1992, 1995) explain cross-linguistic speech perception with the concept of “perceptual assimilation”. That is, both PAM and SLM claim that when listeners perceive non-native sounds, which do not conform with the phonetic characteristics of the native sounds, they perceive sounds flexibly and assimilate unfamiliar, non-native sounds into their native phonological categories. Listeners, however, experience perceptual difficulty in discriminating native versus non-native sound contrasts as the degree of phonetic differences between native language and foreign language phones increases. This degree of difference has been represented by the goodness rating score of the perceived sounds (e.g., Flege et al., 1994; Nishi et al., 2008; Strange et al., 1998). One example is a series of experiments performed by Strange and colleagues. For example, Strange et al. (1998) showed that when asked to categorize and rate the goodness of non-native vowels (American English vowels), Japanese listeners rated non-native vowels that did not have similar spectral counterparts in their native vowel system as not good examples and showed inconsistency in their vowel categorization. AE vowels that had similar acoustic characteristics as the corresponding Japanese vowels, on the other hand, were categorized into the correct vowel categories (VCs) (/i, a, e, o, u/) of the Japanese native vowel system. That is, according to these models, listeners show flexibility in their perceptual boundaries and are able to assimilate variant vowels into their native phonemic categories.

Listener flexibility has also been documented in sociolinguistic work when perceiving vowels of different variants of a language (Cutler et al., 2004, 2005). For example, Shaw et al. (2018) showed that even despite phonetic variation present in vowels of different dialects of English (Australian English, London (UK), New Zealand, Yorkshire (UK), and Newcastle (UK)), listeners were able to categorize vowels to the same phonemic category of the intended vowels, suggesting that listeners show “perceptual flexibility” to the variation in vowels across dialects. In another
study, Kendall and Fridland (2012) showed that listeners’ perceptual boundaries varied by the nature of their dialectal patterns: listeners with the reversed tense–lax vowel pairs (/e/- /ɛ/) showed different perceptual boundaries than those without these dialectal patterns.

Sociolinguistic studies also showed an effect of multiple types of social and linguistic information (e.g., talker dialect and age) on vowel perception. For example, Niedzielski (1999) showed that listeners’ perception of vowels changed depending on whether listeners believed the vowels were produced by speakers from Detroit or Canada. In a series of studies Drager (2006, 2011) showed how the perception of talker age (TA) affects the categorization of the front vowels /æ/ and /ɛ/ in New Zealand English. In New Zealand English, these vowels are undergoing a sound change, such that the pronunciation of /æ/ in younger speakers resembles that of older speakers’ productions of /ɛ/, and the production of /ɛ/ in younger speakers resembles older speakers’ productions of /ɪ/. Drager (2006) showed that listeners’ perception of the New Zealand vowel was influenced by the perceived age of speakers. When asked to judge vowels that are on the continuum of /æ/ and /ɛ/, vowels were more likely to be judged as /æ/ when listeners believe these were produced by younger speakers. That is, listeners appeared to accommodate the sound change in progress: the very same token was categorized as /ɛ/ when the talker was thought to be old; and /æ/ when produced by a younger speaker. In addition, older listeners were more likely to judge vowels that are on the continuum of /æ/ and /ɛ/ as /æ/ when the stimuli were played along with a picture of younger speakers. This effect, however, was not found in younger listeners, which suggests that older listeners who themselves have documented the ongoing changes in New Zealand front vowels are more sensitive to this difference than younger listeners. Koops et al. (2008) also showed that listeners’ perception of vowels changes by the perceived TA. In their study, one of the ongoing changes in Southern vowels, unmerging of /ɪ/ and /ɛ/ before nasal consonants (as in pin and pen) among younger speakers was examined. The results showed that listeners tend to perceive vowels as more “merged” when they heard vowels along with the picture of older women than vowels coupled with a picture of a middle-aged woman, which demonstrates listeners’ perceptual bias that depends on perceived TA.

Research has also found an effect of perceived TA on listeners’ perception of sounds. For example, Schellinger (2008) examined how listeners’ perception of children’s speech is influenced by the age of a child as well as the category goodness of the children’s productions. They elicited ratings of children’s productions preceded by carrier phrases that varied in fundamental frequency (F0), and which either contained other speech errors, or were produced accurately. The combined influence of F0 and the presence of speech errors was intended to suggest the child’s age and overall developmental level. This study found that listeners’ judgment of children’s productions of fricatives /s/ and /θ/ changed by F0 (related to the perceived age), formant values, and the presence or absence of speech errors in the productions. The Munson et al. (2012) study, using the same set of stimuli used by Schellinger (2008), also found that listeners’ perception of sounds changes by the perceived TA, in that tokens were perceived to be correctly produced more often when listeners believed that those were produced by older children.

In summary, findings of previous vowel perception studies have shown (a) listeners’ sensitivity to the phonetic details of the native and non-native vowel stimuli (variances in the acoustic cue) as well as (b) listeners’ perceptual bias based on the perceived TA (social inferences). One of the recent models of speech perception proposed by Kleinschmidt and colleagues (Kleinschmidt & Jaeger, 2015; Kleinschmidt et al., 2018) combines these two issues in speech perception and proposed an ideal adapter model. This model claims that when listeners encounter unfamiliar sounds, they adjust their perception to overcome variability in the speech signal and recognize sounds based on their previous experience/knowledge. According to this theory, speech perception includes recognition, generalization, and distributional learning (Kleinschmidt & Jaeger, 2015, p.
Based on the above listed previous research on vowel perception, it was predicted that listeners’ perception of vowels would largely be constrained by phonetic characteristics of their native vowels. However, we expected to see flexibility in listeners’ perceptual categories, especially when perceiving vowels produced by younger talkers. Considering that vowel development is a gradual process that is not completed by a specific age (e.g., Chung et al., 2012), it was predicted that listeners would experience more difficulty recognizing vowels produced by younger speakers. Yet, listeners were predicted to adapt to these vowels based on their previous experiences and be able to recognize them, as suggested by the ideal adapter model. Thus, we predicted fewer cross-linguistic differences in vowel perceptual categories for vowels produced by younger children, as compared to vowels produced by older children, with the clearest differences on vowels produced by adult speakers. To test this hypothesis, listeners’ perception of vowels of all possible range of acoustic variation, five vowels (/i, e, a, o, u/) produced by monolingual speakers of three different age groups (2-year-olds, 5-year-olds, and adults) and five different first languages (American English, Cantonese, Greek, Japanese, and Korean) were included as a stimulus set. This was motivated by a desire to increase the variation in the stimuli that listeners were presented with, and not to test listeners’ perception of vowels produced by their same language versus other languages. Listeners were native speakers of American English, Greek, and Korean. To examine the relationship between listeners’ vowel perceptual categories and acoustic phonetic characteristics of vowel stimuli, the center and size of listeners’ vowel perceptual categories was examined. Additionally, the nature of listeners’ vowel perceptual subcategories was investigated by relating the acoustic properties of each vowel sound to goodness rating scale measures. In this study, the focus was placed on examining perceptual categories of the three “shared” vowel phonemes, /a/, /i/, and /u/, which are represented with the common phonemic symbols across the three languages. These three vowels were chosen to make comparisons of the center and size of the vowel perceptual space (VPS) across the three listener languages (LLs) with different number and organizations of vowels. Specifically, the current study aims to examine the flexibility or stability of listeners’ vowel perceptual system, how they adapt to variability in the speech signal induced by talker language and age. It is especially important to understand the nature of adult listeners’ VPS in relation to TA as it is directly related to the transcription-based accuracy judgment that is commonly used in clinical settings to assess speech of children with suspected speech sound disorders. It is also valuable in understanding how perception biases might affect the reliability and validity of transcriptions of children’s speech in research studies of phonetic development.

2 Methods

2.1 Participants

A total of 60 listeners, 20 in each of the three language groups (American English, Greek, and Korean) participated in this study. All listeners were monolingual speakers of each language and were between 18 and 30 years of age. All English-speaking listeners were recruited and tested in Columbus, Ohio, Greek-speaking listeners in Thessaloniki, Greece, and Korean-speaking listeners in Seoul, Korea. These were the same cities in which the vowel stimuli had been collected from native speakers of three of the five stimulus languages. Most of the listeners were undergraduate students in each country from the same city where the experiment was conducted, and had minimal experience with child speech. All listeners reported no history of speech, language, or hearing impairment.
2.2 Stimuli

The stimuli used in this experiment were selected from the *paidologos* corpus (Edwards & Beckman, 2008a, 2008b). The stimuli set included five vowels (/i, a, e, o, u/) in word-initial consonant vowel (CV) sequences excised from picture-prompted real-words and nonwords produced by 2- and 5-year-olds, and by adults of five first languages (American English, Cantonese, Greek, Japanese, and Korean). Each speaker’s dialectal background matched with that of listeners of each language (American English in Ohio; Greek in Thessaloniki, Greece; and Korean in Seoul, Korea).

All target vowels were preceded by either velar (/k/ or /ɡ/) or alveolar–palatal obstruents (/t/, /d/, /s/, or /ʃ/). For Japanese, both long and short allophones of /i/ and /u/ vowels were included, as vowel length does not influence vowel quality in Japanese (Keating & Huffman, 1984), or the assimilation of Japanese vowels to English VCs (Nishi et al., 2008). In this study, only CVs that were judged as correct by a trained native transcriber of each language were included. Numbers of vowel stimuli varied by language and age group because there were different numbers of vowels that were judged to be correct for the given VC.

For each target CV sequence, the beginning and end of the stimulus were determined as follows: the beginning was demarcated 10 milliseconds (ms) prior to the stop burst or frication. The end was set 10 ms after the location in the vowel where there was a dramatic reduction of F2 energy. Because the CV sequences were derived from naturally-produced utterances, their length varied depending on how each speaker produced the target words. The speech analysis program *Cool Edit* (version 1.2) (Johnson, 2000) was used to on-ramp and off-ramp the first and last 5 ms of the stimuli, using cosine-squared ramping to remove any potential artifacts that might have resulted when excising the stimuli from word-level utterances. A summary of the number of stimuli used in this study is provided in Table 1. In this study, the analysis was focused on measuring the size and location of perceptual categories of three corner vowels, /i/, /a/, and /u/.

2.3 Procedures

The perception experiment was programmed with *E-prime 1.0* experiment management software (Schneider et al., 2002). Listeners progressed through the experiment at their own pace with mouse clicks on the computer screen, hearing each stimulus over Sennheiser headphones (HD 280 professional). All participants were tested in a quiet room. Instructions were given visually on the computer screen in the listener’s native language. Listeners were told that the purpose of the study was to examine how people perceive vowels and that their task was to identify the vowels from the stimuli they heard. Listeners were not informed that they would be listening to vowels of different languages or of different age groups.

Each listener completed two tasks, a vowel categorization task followed by a goodness rating task. In the categorization task, listeners were asked to categorize vowels of each stimulus as one of their native VCs displayed on a computer screen. For English listeners, 11 keywords containing the 11 target monophthong vowels of English were given as response options using the Latin alphabet. For Greek listeners, the five vowels of Greek were given as response options using the Greek alphabet, and for Korean listeners, the seven vowels of Korean were given as response options using Hangul. Figure 1 illustrates these response options for the three languages. Keywords were used as response options only for English vowels because the English writing system, unlike the Greek and Korean systems, does not have a one-to-one grapheme-to-phoneme mapping.

For each stimulus, once the subject had finished responding to the categorization task, a goodness rating task followed immediately. For this second task, listeners were asked to judge the goodness rating of each VC by clicking on a line where one endpoint was labeled as “very good example” and the other endpoint was labeled as “very bad example” (or the appropriate translation in Greek...
The next stimulus was presented once the listener had finished responding to both the categorization and goodness rating tasks.

The entire set of CV sequences was presented in a single block. There were two stimuli blocks in this experiment, CV tokens and a token of vowel-only versions of the same stimuli. These stimuli were originally created to compare the perception of natural vowels to the perception of a set of synthetic vowels, as described in Plummer et al. (2013). In this study, the analysis focused on the first set of stimuli. These two blocks were counter-balanced and each block was preceded by one brief practice block. Each block consisted of three age blocks (adults, 5-year-olds, and 2-year-olds), which were presented in a random order for each listener. Each of the three age blocks included CV sequences produced by speakers of each age group in each of the five languages.

### 2.4 Acoustic analysis

First and second formant (F1 and F2) values were measured at the vowel temporal midpoint of each CV sequence using the LPC formant tracker in *Praat* (version 5.3.51) (Boersma & Weenink, 2006). The size of the analysis window was 25 ms and the dynamic range was set at 30 dB. First, vowel duration was measured from the vowel onset, the time at which the first clear glottal pulse was observed, to the vowel offset, the time at which the final glottal pulse began to fade out. Once the duration was calculated, the temporal midpoint was determined, and F1 and F2 values were extracted at this midpoint.

### 3 Results

#### 3.1 Acoustic characteristics

The acoustic characteristics of the vowel sounds (/i, a, e, o, u/) produced by speakers of five different language groups are illustrated in Figure 2 by TA group. To minimize the age-related differences in the size of the vocal tract on acoustic patterns of vowels (e.g., Vorperian & Kent, 2007), while still preserving linguistically and socially meaningful variation (Clopper, 2009), the raw F1

<table>
<thead>
<tr>
<th>Language</th>
<th>Age</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>2 years</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>English</td>
<td>2 years</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>Greek</td>
<td>2 years</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Japanese</td>
<td>2 years</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Korean</td>
<td>2 years</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>59</td>
<td>61</td>
<td>56</td>
<td>66</td>
<td>62</td>
<td>306</td>
</tr>
</tbody>
</table>
and F2 values of each vowel stimuli were recalculated using the z-score transformation following the recommendation of Lobanov (1971). In Figure 2, vowel stimuli were not separated by the phonemic categories of each vowel as the main aim of this paper was not to calculate the accuracy of vowel perceptual categorization, but to examine the size and location of adult listeners’ perceptual space for different vowel sounds. As can be seen in Figure 2, the acoustic characteristics of vowel stimuli used in this study cover a wide range of the F1 by F2 vowel acoustic space. Vowels produced by 2-year-olds cover a wider range of the normalized vowel acoustic space, especially in the F1 dimension, than those of 5-year-olds and adults, showing that 2-year-olds’ productions are more variable than those of 5-year-olds and adults.

3.2 Vowel categorization

Figure 3 shows the categorization results of vowels illustrated in Figure 2 by each LL group. The goodness ratings of each vowel token are represented by different sizes of the symbols. The tokens with larger symbols indicate judgments closer to the “very good example” and those with smaller symbols indicate the “very bad example” end of the scale.
Figure 2. Vowel stimuli produced by 2-year-olds (left), 5-year-olds (middle), and adult speakers (right) of five languages in z-transformed second formant by first formant vowel space. Talker languages are in gray scale, from black (Cantonese (c)) to light gray (Korean (k)) (American English (e), Greek (g), Japanese (j)).

Figure 3. Categorization results of vowels produced by 2-year-olds (“2”, black), 5-year-olds (“5”, gray), and adults (“a”, light gray) by three listener groups (from left to the right, English, Greek, and Korean listeners) for each of the /i/ (top), /u/ (middle), and /a/ (bottom) categories. The size of each symbol represents the goodness judgment for the given sound (larger symbols “very good example”; smaller symbols “very bad example”).
Visual inspection of the data presented in Figure 3 shows that English listeners had the most variable vowel perceptual patterns, followed by those of Greek and Korean listeners. Listeners also showed more variable vowel perceptual patterns when they perceived vowels produced by 2- and 5-year-olds than adults. Across the VCs, the perceptual pattern of /i/ was the most consistent across LL and TA groups. For /a/, variable perceptual patterns were observed, where a greater overlap with the /u/ category existed, especially for English listeners. For /u/, the perceptual spaces of Korean listeners were in general well-separated from /i/, while those of Greek and English listeners showed relatively more variable and overlapping patterns with /u/.

For the goodness ratings, English listeners’ ratings were overall lower than those of Greek and Korean listeners across the TA group and VCs. Across the TA, vowels produced by adults were given the highest scores, followed by those of the 5-year-olds, and 2-year-olds for Greek and Korean listeners. Across VCs, vowels that were categorized as /i/ overall showed the highest goodness rating across LL and TA groups, followed by /a/ and /u/. These patterns are summarized in Table 2.

To better quantify the above illustrated cross-linguistic vowel perceptual patterns across the TA and VCs, categorization results were weighted by a goodness rating scale. This was done by multiplying z-transformed F1 and F2 of each stimulus by goodness rating score for that stimulus from each listener. Then, the center and size of the VPS were calculated for each listener group, TA, and VC.

3.2.1 Center of the VPS. The center of the VPS of each listener (i.e., /i/ produced by two-year olds) was calculated by dividing the sum of the z-transformed formant values for all of the stimuli in a given category by the sum of the goodness ratings in that category (i.e., sum (z-transformed F1) / sum(goodness rating)). As can be observed in Figure 4, across the listener groups, English listeners showed the most overlapping VPSs, mainly between the vowels /a/ and /u/, while those of Greek and Korean listeners showed a relatively clear separation among the three vowels regardless of the TA.

The multivariate analyses of variance (MANOVAs), using Pillai’s trace tests were performed on the average F1 and F2 (z-score) as a function of LL (English, Greek, and Korean listener groups), TA (2-, 5-year-olds, and adults), and VC (/a/, /i/, and /u/) to determine if the VC center differs significantly by LL and TA, and if these effects are mediated by VC. The three-way interaction among

<table>
<thead>
<tr>
<th>Vowel category</th>
<th>TA</th>
<th>LL</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>Greek</td>
<td>Korean</td>
<td></td>
</tr>
<tr>
<td>/a/</td>
<td>2 years</td>
<td>0.53 (0.21)</td>
<td>0.65 (0.24)</td>
<td>0.63 (0.20)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>0.53 (0.23)</td>
<td>0.69 (0.22)</td>
<td>0.66 (0.19)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>0.51 (0.23)</td>
<td>0.73 (0.21)</td>
<td>0.70 (0.17)</td>
<td></td>
</tr>
<tr>
<td>/i/</td>
<td>2 years</td>
<td>0.62 (0.22)</td>
<td>0.59 (0.25)</td>
<td>0.59 (0.24)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>0.62 (0.23)</td>
<td>0.63 (0.24)</td>
<td>0.61 (0.21)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>0.65 (0.23)</td>
<td>0.67 (0.24)</td>
<td>0.63 (0.21)</td>
<td></td>
</tr>
<tr>
<td>/u/</td>
<td>2 years</td>
<td>0.53 (0.24)</td>
<td>0.54 (0.27)</td>
<td>0.55 (0.22)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 years</td>
<td>0.48 (0.24)</td>
<td>0.54 (0.26)</td>
<td>0.59 (0.22)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>0.51 (0.24)</td>
<td>0.59 (0.27)</td>
<td>0.61 (0.21)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The average goodness rating scores and standard deviation (in parentheses) for the vowels that were labeled either as /a/, /i/ or /u/ by three different listener language (LL) and talker age (TA) groups. The scale “1” represents “very good example” and “0” represents “very bad example.”
the three variables was also added to the model. To account for individual variability, each listener was entered as a random variable.

Results showed a significant effect of LL, TA, and VC on the average F1 and F2, LL: $F(2, 110) = 6.00, p < 0.001$; TA: $F(2, 220) = 25.31, p < 0.001$; VC: $F(2, 220) = 1701.7, p < 0.001$. The three-way interaction among the three variables was also significant, $F(8, 440) = 4.34, p < 0.001$.

The follow-up three MANOVAs for each of the three age groups also revealed a significant main effect of LL on vowel perceptual center, average F1 and F2 ($z$-score), in all three age groups, 2-year-olds: $F(4, 110) = 4.03, p = 0.004$; 5-year-olds: $F(4, 108) = 4.11, p = 0.004$; adults: $F(4, 110) = 4.03, p = 0.0004$, but with a larger effect for vowels produced by adult speakers than those of 2- and 5-year-olds, confirming our initial hypothesis.

A series of univariate analyses of variance separately for each of the two dependent variables (F1 and F2) was performed for each of the three TA groups to further examine the nature of the significant main effects. The results of these analyses are provided in the Online Appendix A and

**Figure 4.** The average weighted $z$-score transformed first formant and second formant (center of the perceptual category of vowel stimuli) that were labeled either as /a/, /i/, or /u/ produced by 2-year-olds (2 years), 5-year-olds (5 years), and adults (adults) for each of the three listener language groups.
the output of post-hoc Tukey tests in the Online Appendix B. Results showed that in all three age groups, F1 differed significantly by LL and VC. An interaction between LL and VC was also significant across the age groups, suggesting that the effect of LL on F1 differed significantly by VC. That is, F1 differed significantly by LL but only for a perceptual category of /a/, whose values were significantly lower for English listeners than those of Greek and Korean listeners. For F2, a significant effect of LL was found, but only when listeners were perceiving vowels produced by 5-year-olds and adults. The effect of VC and an interaction between the two variables were significant across the three TA groups. That is, the /a/ perceptual category of English listeners had significantly higher F2 than those of Greek and Korean listeners when perceiving vowels produced by 5-year-olds and adults. Also, F2 of /u/ English and Greek listeners’ perceptual categories were significantly higher than those of Korean listeners for vowels produced by 5-year-olds and adults. These patterns are illustrated in Figure 5.

3.2.2 Size of the VPS. The size of the VPS was examined by calculating the Euclidean distance (ED) between the weighted vowel perceptual center and the location of each of /a/, /i/, and /u/ in a vowel acoustic space. Figure 6 shows the weighted ED of each of the three vowels by TA and LL. These values are summarized in Table 3.

Linear mixed effects models were used to examine the effect of LL, TA, and VC on the size of the VPS. This analysis was performed using the lme() function from the nlme package (Pinheiro
et al., 2019) for R (R Development Core Team, 2018). The post-hoc analyses was performed using the `lsmeans()` function from the `emmeans` package (Lenth, 2016). The ED values were entered as a dependent variable, LL, TA, and VC as fixed factors, and each listener and perception of each stimuli were entered as random variables. Results showed no significant effect of LL, $F(2, 55) = 0.15, p = 0.862$, but significant effects of TA, $F(2, 8926) = 57.15, p < 0.0001$, and VC, $F(2, 8926) = 333.25, p = 0.0001$, on ED. A three-way interaction among the three variables was significant, $F(8, 8926) = 2.94, p = 0.0003$.

Subsequent analyses were performed to examine the main effect of TA and how this effect is mediated by the VC with three separate models for each of the three LL groups. The output of each of the three models is summarized in Table 4 and those of post-hoc Tukey tests are summarized in the Online Appendix C. The results showed that for English listeners, ED differed significantly only by VCs, with ED of /ɑ/ being the longest followed by those of /i/ and /u/ across the three age groups. This suggests that English listeners have a larger perceptual space for /ɑ/ than the /i/ and /u/ regardless of TA. For both Greek and Korean listeners, the effect of TA and VC, as well as an interaction between the two variables, were significant. As hypothesized, ED of vowels produced by younger children were longer than adults, suggesting a larger VPS for vowels produced by younger children than adults. This effect differed by VC, /ɑ/: 2-year-olds > 5-year-olds > adults; /i/: 2-year-olds > 5-year-olds; /u/: 5-year-olds > 2-year-olds; 5-year-olds > adults. Overall, listeners had the largest perceptual space for vowels produced by 2-year-olds, followed by those of 5-year-olds and adults, and for /ɑ/ than for /i/ and /u/.

Table 3. The mean Euclidean distance (ED) by listener language (LL) and three talker age (TA) groups (2- and 5-year-olds and adults) for each of the three vowel perceptual categories (/ɑ/, /i/ and /u/). See Figure 6 for standard deviations of the mean ED for each group.

<table>
<thead>
<tr>
<th>Age</th>
<th>/ɑ/</th>
<th>/i/</th>
<th>/u/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Greek</td>
<td>Korean</td>
</tr>
<tr>
<td>2 years</td>
<td>0.98</td>
<td>1.04</td>
<td>1.03</td>
</tr>
<tr>
<td>5 years</td>
<td>0.87</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>Adult</td>
<td>0.87</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Figure 6. Euclidean distance between the weighted vowel perceptual center and vowel stimuli that were labeled either as /ɑ/ (left column), /i/ (middle), or /u/ (right) across the three talker age groups (x-axis) and listener language group (English: black; Greek: gray; Korean: light gray).
Discussion

In this study, the effect of listeners’ native language and TA on identification of vowel sounds was examined. Listeners of three different native languages (English, Greek, and Korean) categorized and rated the goodness of vowels produced by talkers of three different age groups (2-year-olds, 5-year-olds, and adults) who were native speakers of five different languages (American English, Cantonese, Greek, Japanese, and Korean). The size and center of listeners’ perceptual categories of /a/, /i/, or /u/ varied by LL and TA, and these effects varied by VCs.

First, the vowel categorization results showed that (a) English listeners, (b) vowels produced by 2-year-olds, and (c) the vowel /a/ showed a wider and more variable perceptual space than did (a) Greek and Korean listeners, (b) vowels produced by 5-year-olds and adults, and (c) those of /i/ and /u/, respectively. As for the goodness ratings, overall, English listeners’ ratings were lower than those of Greek and Korean listeners across TA and VCs. Across TA groups, ratings of vowels produced by adults were the highest, followed by those of 5-year-olds and 2-year-olds. For VCs, on average, the ratings of /i/ were the highest followed by those of /a/ and /u/. This suggests that English listeners in general accept a wider range of vowels for the /a/ category than they do for the /i/ or /u/ categories, yet were less confident in their overall vowel perceptual judgment than Greek and Korean listeners.

As for the center of the VPS, cross-linguistic effects were found for /a/ and /u/. For /a/, the center of the perceptual space of English listeners had significantly lower F1 and a larger perceptual space in the F1 dimension than those of Greek and Korean listeners when perceiving vowels produced by talkers of all three age groups. For /u/, the center of the perceptual space of English listeners had significantly higher F2 than those of Greek and Korean listeners when perceiving vowels produced by 5-year-olds and adults. This cross-linguistic effect reflects the variable nature of these vowels in their native vowel system (e.g., Clopper & Pisoni, 2006; Harrington et al., 2008; Hillenbrand et al., 1995). English /u/ fronting is a well-known vowel shift, which started from Southern vowels but has spread across different regional dialects of American English (e.g., Labov et al., 2006). The perceptual center of English listeners reflects this fronted nature of /u/ that differs from those of Greek and Greek, which occupy a more back location of the vowel acoustic space as compared to English /u/ (Chung et al., 2012). This result shows that listeners of different languages impose different criteria when perceiving vowels to cope with the cross-linguistic differences in the acoustic signal. Overall, these findings are consistent with the initial hypothesis that

<table>
<thead>
<tr>
<th>LL group</th>
<th>Variables</th>
<th>Degrees of freedom</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>TA</td>
<td>2</td>
<td>2.25</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>VC</td>
<td>2</td>
<td>157.79</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>TA × VC</td>
<td>4</td>
<td>2.01</td>
<td>0.089</td>
</tr>
<tr>
<td>Greek</td>
<td>TA</td>
<td>2</td>
<td>37.38</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>VC</td>
<td>2</td>
<td>111.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>TA × VC</td>
<td>4</td>
<td>31.73</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Korean</td>
<td>TA</td>
<td>2</td>
<td>26.58</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>VC</td>
<td>2</td>
<td>89.50</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>TA × VC</td>
<td>4</td>
<td>19.39</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
predicted an influence of native vowel system on vowel perceptual categories. In addition, the fact that this cross-linguistic difference was observed only when listeners were perceiving vowels produced by 5-year-olds and adults serves as an additional evidence to support the claim that listeners are capable of detecting phonetic differences in vowel acoustic signals and adjusting their perceptual categories accordingly (perceptual flexibility): a stronger cross-linguistic effect for vowels produced by adult speakers as compared to those produced by the youngest age group.

As for the size of the VPS (represented by ED values), contrary to what was hypothesized, there was no significant effect of LL. However, ED values differed significantly by TA and VCs. Larger perceptual spaces were found when listeners perceived vowels produced by 2-year-olds, followed by those of 5-year-olds and adults, confirming our hypothesis. This effect was found mainly for the perceptual category of /a/. English listeners, however, showed no such TA effect, indicating that they maintained a similar size of the VPS regardless of TA. This could also mean that while Greek and Korean listeners become less strict with accuracy judgments when perceiving vowels produced by younger speakers, English listeners could be applying similar vowel perceptual criteria regardless of TA, given the more variable nature of the English vowel system, which has more overlapping acoustic patterns among VCs than those of Greek and Korean (Chung et al., 2012). That is, English listeners could be less prone to perceptual bias by TA during vowel perception. Instead, they could be applying overall less rigid vowel perceptual criteria to vowels of all TAs than Greek and Korean listeners.

The findings of this study have implications for the accuracy judgment of children’s vowels. Traditionally, vowels are considered to be mastered by the age of 3 (e.g., Irwin & Wong, 1983; Templin, 1957). Due to this perception that vowels are earlier acquired sounds (e.g., Davis & MacNeilage, 1990; Pollock, 2013; Stoel-Gammon & Herrington, 1990), not many standardized tests of articulation include vowels (Eisenberg & Hitchcock, 2010), although vowel errors are considered as one of the important diagnostic markers for childhood apraxia of speech (e.g., Davis et al., 2005; Lewis et al., 2004). In addition, evaluation of vowel accuracy and error patterns is not commonly included during the assessment of children with suspected speech sound disorders. The findings of the current study suggest that early mastery of vowel sounds could be partially a byproduct of listeners being more lenient and having larger VPSs when judging vowels produced by younger speakers. This could especially be the case for Greek and Korean listeners, who showed a larger perceptual space for perceiving vowels produced by younger children than those of older children or adults. For English, the overall larger size of the vowel perceptual boundaries regardless of TA could also have resulted in the reported early mastery of vowels. Overall, results of the present study showed how the vowel perceptual categories of listeners vary by the nature of their native vowel system and how listeners adjust their vowel perceptual category by TA. This is consistent with the PAM and the SLM models, which suggest that listeners assimilate unfamiliar sounds into vowel phonemic categories but are able to detect phonetic differences and show perceptual flexibility during vowel perception. This result is also in line with the ideal adapter model (Kleinschmidt & Jaeger, 2015; Kleinschmidt et al., 2018), which suggests that listeners can recognize talker language and age based on their previous experience with native vowels, and adjust their perception accordingly to overcome variations in the signal. Clinically, these findings reiterate the importance of matching the linguistic backgrounds of clinicians and child clients when judging the accuracy of their speech during assessment.

In this study, vowels presented in CV syllables excised from single word productions were used for understanding the relationship between listeners’ perception and acoustic properties of vowel sounds. Although the stimuli used in the current study were naturally-produced vowels, each stimulus was excised to generate a CV format, thus reducing the vowel durational differences across

In this study, vowels presented in CV syllables excised from single word productions were used for understanding the relationship between listeners’ perception and acoustic properties of vowel sounds. Although the stimuli used in the current study were naturally-produced vowels, each stimulus was excised to generate a CV format, thus reducing the vowel durational differences across
stimuli and limiting the range of vowel spectral changes, two of the variables that are known to contribute to the perception of vowel sounds (e.g., Hillenbrand et al., 1995; Nearey, 2013). Although listeners can reliably perceive phonetic differences in native versus non-native sounds even in shorter vowel stimuli (e.g., McCullough & Clopper, 2016), using whole words or even short phrases when performing vowel categorization and goodness judgment scale could have provided additional important acoustic cues to listeners and thus generated different results. Although the primary reason for using CV stimuli in the current study was to avoid any “lexical and semantic support” by having additional contexts besides the preceding consonants (Munson et al., 2012), a comparable future study should include whole words or short phrases as stimuli to provide a more complete picture of the effect of LL and TA on perception of vowel sounds. In addition, the current study included vowel stimuli produced only by three TA groups (2- and 5-year-olds and adults). Given that previous studies have reported age of acquisition of vowels between 24 to 36 months, inclusion of vowels produced by 3- and 4-year-olds could have captured gradual changes in listeners’ center and size of the VPS as children refine phonetic characteristics of each VC with age. Lastly, for the categorization task, listeners of the current study were given different number of choices depending on the nature of their native VCs. English listeners were given an 11-alternative forced choice, Greek listeners a 5-alternative forced choice, and Korean listeners a 7-alternative forced choice. Although this task was used to match the number of native vowel systems of each language, it should be acknowledged that the different number of choices could have influenced the categorization results. Despite these limitations, by using a large database that included vowel stimuli produced by talkers of five different languages and three age groups and listeners of three native languages, the current study showed evidence from cross-linguistic perception that listeners adjust their perception to recognize vowels that vary by TA.

Acknowledgements

We thank Mary E. Beckman for her valuable contributions to this experiment and to this project overall. We thank all listeners who participated in this study. We also thank Sokol Vako and Eun Jong Jong Kong for their insightful advice on the statistical analyses. We thank Meropi Kyriazi and Chanelle Mays for their work on data collection in Greece and in Columbus, Ohio, respectively. Parts of this work have been presented at the 159th Acoustical Society of America Meeting, Baltimore, MD 2010.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by National Institutes of Health Grant DC02932 to Jan Edwards and Mary E. Beckman, by National Science Foundation Grants BCS-0729140 to Jan Edwards, BCS-0729277 to Benjamin Munson, and BCS-0729306 to Mary E. Beckman, and by Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant P30-HD03352 to the Waisman Center.

ORCID iD

Hyunju Chung  
https://orcid.org/0000-0002-6657-7422

Supplemental material

Supplemental material for this article is available online.

Notes

1. For typographic simplicity, the symbol /a/ is used throughout the paper to represent the low back vowel in each language, even though /ɑ/ is likely more appropriate for some of the languages (i.e., American English, Japanese, and Korean).
The acoustic characteristics of these vowels, the cross-linguistic differences for each of the three age groups (2-, 5-year-olds, and adults), can be found in Figure 2 and Figure 3 of the companion paper, Chung et al. (2012).

References


