Australian Journal of Linguistics, 2015
Vol. 35, No. 2, 121–141, http://dx.doi.org/10.1080/07268602.2015.1004998

Routledge
Taylor & Francis Group

Perception and Production of English Vowels by American Males and Females*

1959

BYUNGGON YANG AND D. H. WHALEN

Pusan National University; The City University of New York, Haskins Laboratories and Yale University

(*Accepted 25 June 2014*)

The aims of this study are to explore the link between the perception and production of vowel sounds and to make a minor contribution to the debate about the hyperspace effect. The sample of 18 American English speakers (male and female) identified ideal American English vowels from sets of synthetic vowels, and the same participants produced those vowels in a clear speaking style. The formant values of the produced vowels were measured and compared with those of perceived vowels on the vowel space. The results show that a vast majority of the perceived vowel spaces of the male and female groups were not significantly different, whereas the produced vowel spaces of the two groups were significantly different. Additionally, in the male group, the perceived vowel space was larger than the produced vowel space, whereas the opposite phenomenon was observed in the equivalent vowel space of the female group. The perception of vowels in this study, therefore, appears to reference a speaker who is not necessarily the same as the listener. Thus, the hyperspace effect based on a simple comparison between perceived and produced vowel spaces should be reconsidered.

Keywords: Perception; Production; American English; Vowel; Formant; Synthesis; Method of Adjustment

1. Introduction

When male and female participants listen to the speech of other people with various vocal tract sizes, do they show a pattern of perception similar to their own production

*We would like to thank Douglas Honorof, Carol Fowler, Arthur Abramson, Edward Flemming, Michael Carey and anonymous reviewers for their comments and suggestions on the manuscript. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST, Basic Research Promotion Fund) [NRF-2011-013-A00032].

or a different pattern? The aims of this study are to explore any link between perception and production and to invoke further discussion regarding the proposed hyperspace effect (Johnson *et al.* 1993). To that end, this study compares the larger female-produced vowel space and the smaller male-produced vowel space in light of the perception of an ideal vowel space. The produced vowel space of a speaker can be formed by obtaining the first two formant frequency values. Formants are acoustic correlates of the vowel qualities and are dependent on a speaker's vocal tract shape and length (Fant 1970; Pickett 1987). The formant frequencies are inversely proportional to the length of the vocal tract. A shorter vocal tract yields higher formant values. Generally, female vocal tracts are shorter than those of males, and the American female vowel space is thus much larger than that of the American male (Peterson & Barney 1952; Hillenbrand *et al.* 1995; Yang 1990, 1996).

Two questions may arise here. First, how do speakers with different vocal tract lengths manage to categorize acoustically different tokens as 'the same vowel'? Second, is a theory based on a simple comparison between the perception and production of various speakers consistent with the observed data? The first question has been explored in a speaker normalization context, which refers to a procedure to factor out non-linguistic characteristics such as vocal tract size and gender (Ladefoged & Broadbent 1957; Traunmüller 1988). Dialectal, sociolectal, idiolectal and phono-stylistic differences also account for the acoustic variation among male and female speakers (Traunmüller 1988). A listener seems to apply his or her own phonological system to identify the same vowel among acoustically different productions of male and female speakers. Various algorithms have been proposed to determine a vowel's identity based on acoustic features, but to date there is no decisive approach (for a review, see Flynn 2011; Yang 1996). The second question has been addressed in a debate between Johnson *et al.* (1993) and Whalen *et al.* (2004a). Thus, the present study was conducted to provide a minor contribution to an existing debate and a general exploration of the link between perception and production.

This study provides a replication and extension of the Johnson *et al.* (1993) and Whalen *et al.* (2004a) studies to explore whether speech perception reflects a listener's own production. A major extension is achieved by employing three sets of synthetic stimuli that simulate a base speaker and two additional speakers with shorter or longer vocal tracts. Johnson *et al.* (1993) proposed the hyperspace effect, in which the vowel tokens chosen for the MOA (Method of Adjustment) task were more extreme than those produced by the participants. In the MOA procedure, the participants chose the most exemplary sound for a given target vowel after listening to synthesized vowel stimuli created by using various combinations of formant values. The adjustment refers to how participants adjust their final choice by repeatedly clicking on a set of stimuli with a slight frequency shift to find the best exemplar. Johnson *et al.* (1993) adopted the MOA and found that boundaries of internal prototypes of vowel targets differed from boundaries of produced vowel targets (Samuel 1982; Repp & Liberman 1987). Comparing the formant characteristics of vowels selected using this procedure to those of vowels produced by 10 female and four male participants of diverse linguistic backgrounds, Johnson *et al.* (1993) observed that the perceived vowel space was much larger in both the F1 and F2 dimensions than the produced space. This mismatch led them to hypothesize that the speakers must have phonetic targets on a hyperarticulated vowel space at the first stage of vowel production and then reduce it to produce them at the second stage. Johnson *et al.* (1993) proposed that the perceived vowel space is a marked expansion of the produced space irrespective of the male and female listeners' vocal tract size.

Whalen *et al.* (2004a) argue that Johnson *et al.*'s (1993) findings were attributable to methodological artefacts. In particular, Whalen *et al.* (2004b) note that Johnson *et al.*'s (1993) synthetic stimuli were created to represent a male speaker with appropriate f0 values, but the perceived vowel spaces appeared more extreme than those for a male speaker. They also noted that Johnson *et al.*'s (1993) male and female participants provided similar perceptions, which they attributed to perceptual normalization by listeners with different vocal tract sizes. Thus, Whalen *et al.* (2004b) identified a problem in comparing the perceptual vowel space of one speaker with the production vowel space of many speakers averaged over different vocal tracts. By extension, it may stand to reason that the vowel spaces produced by a group of female participants with shorter vocal tracts, for instance, would be larger than the areas that they would report perceptually for a male speaker with a longer vocal tract. Johnson *et al.* (2004) claimed that Whalen *et al.* (2004a, 2004b) had misunderstood their point. Johnson *et al.* (1993) intended to define hyperspace as the hyperarticulated vowel space of a speaker and noted that the synthetic stimuli might represent a speaker and, within that speaker, listeners might have chosen the hyperarticulated vowel space for that speaker. Thus, Johnson *et al.* (1993) claimed that they were justified in plotting perceived and produced vowel spaces together to directly illustrate their hyperspace effect because their production data matched the range chosen by the listeners. The range matching in Johnson *et al.* (1993) may indicate vowel space expansion from the hyperarticulated data to the perceptual vowel space peripherally and proportionately in four directions (up, down, left and right) within a single plane. However, the data in Whalen *et al.* (2004a) illustrate both hyperarticulation and hypoarticulation (with expansion only in certain directions). Based on their experimental data, Whalen *et al.* (2004a) claimed that the hyperspace hypothesis might have methodological artefacts.

The debate about hyperspace between Johnson *et al.* (1993) and Whalen *et al.* (2004a, 2004b) remains open, and one of the aims of this paper is to present experimental results related to this controversial issue. If the synthetic stimuli used by Johnson *et al.* (1993) were consistent with the production of an average male, a comparison of the perceived and produced vowel spaces of female speakers might demonstrate the opposite effect—vowel space reduction—because female formant frequencies are generally higher than those of males. In this respect, Whalen *et al.* (2004b) suggested that a range of synthetic voices suggestive of different vocal tracts should be employed for further study. One of the departure points of the present study is to examine how the perceived vowel spaces would differ when synthetic stimuli mirroring between-speaker differences in vocal tract length are presented to participants in the MOA task.

This paper consists of perception and production experiments. The perception experiment replicates the synthetic stimulus set of Whalen *et al.* (2004a) with two additional modifications to the synthetic stimuli that specifically scale up or down their first four formants uniformly. The perceived vowel space of male and female listeners might be expected to change accordingly. In the production experiment, f0 and the first three formant values were collected from the same participants in a clear speaking style. A clear speaking style is expected to induce participants to produce somewhat larger vowel spaces to approximate the exemplary vowel space in the MOA task. The perception experiment was performed first and was followed by the production experiment. At the end of the paper, the perceived and produced vowel spaces of male and female speakers will be compared to explore a link between speech perception and production and to discuss whether the hyperspace effect is applicable in this context.

2. Perception Experiment

The purpose of the perception experiment was to observe how participants respond to synthetic stimulus sets—representing three speakers with different vocal tracts—and to obtain the exemplary vowel spaces of the participants. For this experiment, the first author replicated the parameters of the synthetic stimuli used in the previous study (Whalen *et al.* 2004a; the 'base synthetic stimulus set') and created two additional sets of synthetic stimuli to simulate speakers with shorter or longer vocal tracts. The three sets were presented to the American English speakers. The participants were asked to choose the best exemplar for each target vowel presented on a computer screen (see Figure 2 below) and to rate the naturalness of their selection (Johnson *et al.* 1993; Whalen *et al.* 2004a).

2.1. Participants

The sample in the perception experiment consisted of 18 American English speakers (nine males and nine females); the participants were students at Yale University or researchers at Haskins Laboratories without any reported defects in vision, hearing or reading. They were paid for their participation in the two experiments. In general, most of the participants were speakers of the eastern American English dialect. Three male participants were born and raised until the age of 14 in New York, two grew up in New Jersey or Connecticut; the rest grew up in Missouri, Ohio, Louisiana or Texas. Their average age was 26.2 years (range: 18–38). The participants were surveyed regarding the *awed–odd* distinction (Hillenbrand *et al.* 1995); seven male participants had an *awed–odd* distinction in their individual dialects, whereas the other two did not. The average age of the nine females was 26 (range: 20–31). Six of the females were born and had spent most of their early years in Connecticut or Rhode Island, New Jersey or New York; three females grew up in California, Chicago or Ohio. Only one female participant responded that she did not distinguish the *awed–odd* pair in her speech.

2.2. Stimuli

Three sets of synthetic stimuli were created. The first base synthetic stimuli were generated by a software synthesizer in Praat (Boersma & Weenink 2013). A script was developed to create a KlattGrid (see Klatt and Klatt (1990) for a detailed description of parameters) with four formants. The beginning F1 was set to 250 Hz, whereas that of F2 was set to 800 Hz. The maximum F1 was 1000 Hz, whereas that of F2 was 2914 Hz. Two mathematical functions of Praat were used to convert Hertz to Bark or the reverse. Bark transforms the acoustic formant values into auditory ones (Sharf 1970; Zwicker & Terhardt 1980). The step size for F1 was set to 0.42 bark, whereas that of F2 was 0.39 bark, following Whalen *et al.* (2004a). Separate regression formulae proposed by Nearey (1989) were employed to generate F3 from the formant values of F1 and F2. F4 was set to 3500 Hz, if F3 was below that value. Otherwise, F4 was determined by adding 300 Hz to F3. Formant bandwidths were added to the Klatt grid that matched those used by Whalen *et al.* (2004a): 75 Hz for F1, 100 Hz for F2, 150 Hz for F3 and 200 Hz for F4.

Stimulus duration was generated by the same formula as in the previous papers, i.e. duration (in ms)=191.754+0.121×F1−0.00347×F2. An extra duration of 20 ms was added: 5 ms for sufficient vowel duration and 15 ms specifically for gradual amplitude offset within which amplitude values were linearly faded down to zero to reduce the chance of participants hearing an abrupt stimulus offset. Next, the f0 tier was extracted from the grid object and an f0 value was added at the time point of 0.1 s, which replaced the original KlattGrid. The net effect of this manipulation was monotony from the beginning to the end of the vowel duration. Thus, 330 stimuli were created with f0s set at 110 Hz, based on an average f0 value of American English vowels (Yang 1996). Following this, the Klatt voicing amplitude tier was extracted to add 80 dB at the time point of 0.1 s, which led to 80 dB throughout the vowel duration.

The second up-scaled set was created by increasing all four formant values of the base synthetic stimulus set by 15%, fixing f0 at 230 Hz, which is appropriate for a smaller female vocal tract. For the third down-scaled set, each of the four formant values of the base synthetic stimulus set was decreased by 15%, fixing f0 at 110 Hz, which is appropriate for the larger vocal tract of a male speaker. In the present study, synthetic stimuli were created to simulate the relationship between male and female formant values by uniformly scaling up or down the first four formant values of the base stimulus set used in the previous study (Whalen *et al.* 2004a). Specifically, the formant values of the synthetic stimuli were shifted up or down by approximating the average vocal tract ratio of male and female speakers reported in Yang (1996), where the American female formant frequencies were approximately 14% higher than those of the male speakers. Such scaling is expected to yield correspondingly distinct perceptual responses from participants. Moreover, Barreda and Nearey (2012) reported that f0 affected vowel quality mainly indirectly when a participant presumed the identity of the speakers. Their results showed a strong correlation between sex and f0 and a significant relationship between f0 and vowel quality.

The starting formant values of all three sets were fixed at 250 Hz for F1 and 800 Hz for F2 by finding the difference between the F1 and F2 values of the first stimulus of the base set and those of the up- and down-scaled sets and shifting all the stimuli linearly to the same origin by adding 37 Hz for F1 and 120 Hz for F2 to those of the down-scaled set or by subtracting 38 Hz for F1 and 120 Hz for F2 from those of the up-scaled set. Figure 1 illustrates the boundary data points of the three sets. The three sets were the base, up-scaled and down-scaled synthetic stimulus sets, respectively; likewise, the perceptual responses of the participants to the sets will be called base, up-scaled or down-scaled perception data.

2.3. Procedure

The experiment lasted approximately 40–50 minutes for each participant. The participants were tested individually in a sound isolation room at Haskins Laboratories. Stimuli were presented binaurally over Sennheiser PC330 headphones from a Samsung laptop computer (SENS RF711). The sound volume was adjusted to a participant-specific comfort level.

Figure 2 shows a screen capture of the MOA experiment. The current and final trial numbers of the experiment were displayed on the top right corner of each stimulus page to inform the participants of their progress.

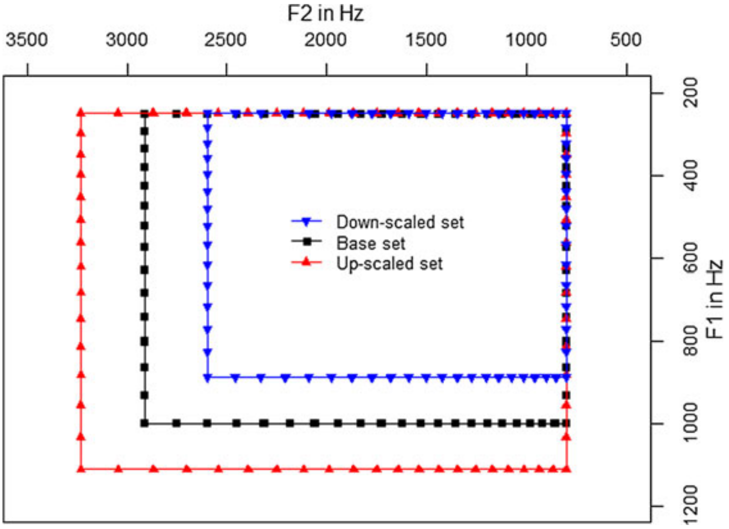
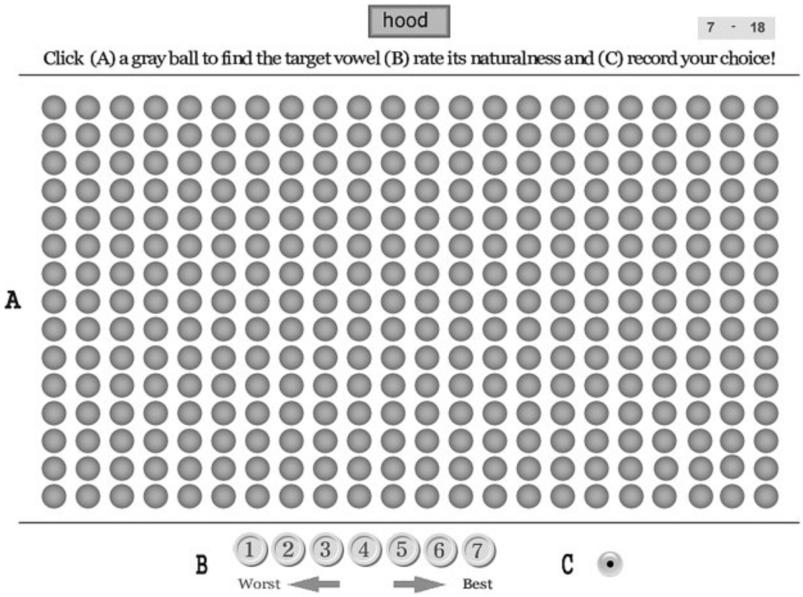


Figure 1 The boundary data (22×15) on the vowel space of F2 by F1 of the base synthetic stimulus set (Base set) and modified sets by scaling all four formant values of the base set up (Up-scaled set) or down (Down-scaled set) by 15%


hood

7 - 18

Click (A) a gray ball to find the target vowel (B) rate its naturalness and (C) record your choice!

Figure 2 A screen capture of the method of adjustment and naturalness rating on a target vowel in the word *hood* in the perception experiment. The participant's task is to find the best exemplar sound and rate its naturalness after listening to associated sounds with grey balls

Before the experiment, the first author gave the participants detailed instructions on the computer screen and asked them to focus on the quality of the vowel in their decision and not to try to memorize vowel positions in the grid. Although Johnson *et al.* (1993) reported that varying the instructions did not affect the outcome significantly, specific instructions seemed to be important to lead participants to a careful decision and to reduce confounding factors caused by individual assumptions about the best exemplar (Smiljanić & Bradlow 2009).

The participants sat before the notebook computer and listened to the stimuli binaurally over their headphones. Their task was to select the best exemplar of a given target vowel from the grid of 330 stimuli. First, they would see one word at the top of the screen, which was randomly chosen from nine English monophthongs in Whalen *et al.* (2004a): *heed*, *hid*, *head*, *had*, *hud*, *odd*, *awed*, *hood* and *who'd*. The participants' task was to find the vowel token they considered the best exemplar of the vowel in that word. In finding the best exemplar, they could click the grey balls as many times as desired to listen to the associated sound file before they decided on the best candidate. Each grey ball flickered once after a mouse click to visually signal the participant's selection, and the F1 and F2 values of the last clicked sound were retained in the computer memory for the final data. Next, the participants were instructed to provide a naturalness rating for the token they chose using the scale buttons below. The naturalness rating and the target vowel and formant values were automatically recorded in the notebook.

To encourage the participants to choose the best exemplar not by remembering its grid location but by clicking several grey balls and gradually approaching the final candidate, distant target vowels from the vowel space (i.e. the front and back vowels) were presented alternately (Whalen *et al.* 2004a). Moreover, two grids with opposite orientations on the F2 axis in the vowel space were provided in alternating trials. The first grid had F1 and F2 values arranged in the order of the acoustical vowel space, in which F1 increased downward on the vertical axis whereas F2 increased leftward on the horizontal axis. The vowel space was presented visually with F2 on the horizontal axis and F1 on the vertical; both had an origin near the top right corner such that the visual space followed the orientation of the vowel chart currently used by the International Phonetic Association. In the second grid, the F2 values increased rightward, while the F1 values remained the same.

To help the participants achieve consistency in their decisions, one of three photographs (two of male students and one of a female student of average height taken from Google images on the Internet) was presented briefly before each MOA set on the computer screen. For the base synthetic stimuli, a picture of a male student was presented whereas a picture of a female student was shown for the up-scaled synthetic stimuli. For the down-scaled synthetic stimuli, a picture of a tall male student was given. The height of the student was easily discernible in the surrounding objects and environment along with direct verbal explanations by the experimenter. Johnson *et al.* (1999) reported an effect of apparent gender or instructions to imagine a male or female speaker on the vowel categorization. They found that vowel boundaries changed as a function of the gender of the visually presented speaker.

2.4. Data Collection and Analysis

Each response from the participants was collected and stored in a text file. Next, the perception results were sorted according to vowels and formant frequency values, and the F1 and F2 values of the synthetic stimuli of each target vowel were separately multiplied by each of the two naturalness ratings collected from each participant, added together, and divided by the sum of the two naturalness ratings, following Whalen *et al.* (2004a). This procedure reduces individual variations in the measurements that may arise due to a single response by using a weighted average. In this way, 972 formant values were obtained in the perception experiment (9 vowels × 3 sets × 18 participants × 2 formants). Statistical analyses were conducted on the vowel perception data of the base, up-scaled and down-scaled synthetic stimulus sets using SPSS (v.20) and *R* (v.2.14.0). Because the authors do not assume that the perception data were drawn from a given probability distribution, non-parametric tests were conducted on the data to compare male and female differences in perception with respect to the three synthetic sets. Descriptive analyses were made mostly by obtaining the mean and standard deviation of each target word from the three synthetic sets using Microsoft Excel.

2.5. Results and Discussion

Tables 1 and 2 list the average formant values of the male and female responses to the perceptual tasks with respect to the three synthetic stimulus sets reflecting their naturalness ratings. Figures 3 and 4 illustrate the perceived vowel spaces of the two sex groups according to the three synthetic sets. To examine perceptual difference, the acoustic units were transformed into the bark scale.

In Figures 3 and 4 two visible trends are clear: a similar vowel perception on the same synthetic stimulus set and a gradual increase from the down-scaled set through the base set to the up-scaled set.

First, the perceived vowel points of the male and female groups converged to the given stimulus set. A Mann-Whitney *U* test was conducted between the male and female perception data of the base set in bark using SPSS. For F1, there were no significant differences in the majority of the comparisons, except for the cases of the three vowels [ɪ, ʌ, ɑ] at the alpha level $p<0.05$ ($n=18$, Mann-Whitney $U=14$, $p=0.019$ for [ɪ]; $n=18$, Mann-Whitney $U=12.5$, $p=0.011$ for [ʌ]; $n=18$, Mann-Whitney $U=13$, $p=0.014$ for [ɑ]). None of the comparisons between the male and female F2 values in bark were statistically significant. Another Mann-Whitney *U* test was conducted between the male and female perception data of the up-scaled set in bark. The results were not statistically significant for any comparison between the male and female participants' vowels in F1 and F2. Moreover, no statistically significant differences were observed for any of the comparisons between the male and female participants' vowel data of the down-scaled set in F1. For F2, none of the statistical comparisons were significant, except those of the two vowels [ɔ, ɑ] ($n=18$, Mann-Whitney $U=16$, $p=0.031$ for [ɔ]; $n=18$, Mann-Whitney $U=3.5$, $p=0.000$ for [ɑ]). Although we did not attempt to control the dialects of the participants, it might be necessary to examine the outliers and test the data again after screening participants with different dialects

Table 1 Average formant values reflecting naturalness ratings of the male results of the method of adjustment tasks, with f0 set at 110 Hz for both the down-scaled and base data and with f0 set at 230 Hz for the up-scaled data

Scale	Down		Base		Up	
Vowel	F1	F2	F1	F2	F1	F2
i	298 (18)	2424 (156)	309 (24)	2713 (132)	312 (44)	2892 (199)
ɪ	435 (35)	2311 (133)	480 (44)	2444 (231)	524 (49)	2636 (219)
ɛ	655 (77)	2138 (165)	646 (105)	2365 (262)	754 (114)	2592 (310)
æ	783 (57)	1990 (208)	874 (77)	2188 (302)	962 (82)	2381 (304)
u	344 (47)	980 (116)	337 (31)	1015 (157)	358 (64)	974 (144)
ʊ	479 (73)	1221 (189)	454 (29)	1304 (182)	550 (73)	1459 (243)
ʌ	617 (55)	1231 (101)	660 (38)	1294 (115)	756 (41)	1465 (159)
ɔ	734 (57)	1057 (143)	781 (59)	1058 (97)	854 (102)	1117 (225)
ɑ	793 (48)	1160 (128)	841 (57)	1180 (105)	953 (72)	1198 (185)
Average	571 (52)	1612 (149)	598 (51)	1729 (176)	669 (71)	1857 (221)

Note: The last row lists the average formant values of the nine vowels. SD is given in parentheses.

Table 2 Average formant values reflecting naturalness ratings of the female results of the method of adjustment tasks with f0 set at 110 Hz for both the down-scaled and base data and with f0 set at 230 Hz for the up-scaled data

Scale	Down		Base		Up	
Vowel	F1	F2	F1	F2	F1	F2
I	288 (22)	2357 (112)	301 (33)	2591 (129)	354 (63)	2939 (177)
ɪ	425 (60)	2174 (244)	420 (36)	2312 (264)	554 (92)	2617 (238)
ɛ	634 (78)	2064 (266)	653 (38)	2239 (217)	755 (84)	2558 (278)
æ	808 (82)	2084 (256)	884 (107)	2176 (216)	1001 (104)	2545 (274)
u	305 (30)	901 (123)	337 (37)	1006 (86)	360 (46)	1053 (179)
ʊ	448 (35)	1096 (86)	425 (36)	1199 (116)	533 (63)	1419 (237)
ʌ	576 (84)	1117 (123)	585 (52)	1229 (86)	728 (54)	1376 (185)
ɔ	694 (33)	870 (46)	738 (53)	984 (98)	866 (84)	1068 (142)
ɑ	780 (54)	1019 (105)	758 (97)	1134 (97)	969 (80)	1236 (105)
Average	551 (53)	1520 (151)	567 (54)	1652 (145)	680 (75)	1868 (202)

Note: The last row lists the average formant values of the nine vowels. SD is given in parentheses.

from the data when a short perceptual distance between the two vowels [ɔ, ɑ] can be observed (Hillenbrand *et al.* 1995). Dialect differences are an issue, as noted by Whalen *et al.* (2004a). Although speakers' dialects may have fewer known effects on perception, they will certainly affect production and may create issues regarding comparisons of perception and production for the same speakers. Certain participants may have applied different phonological systems from the other participants to the perception of the synthetic stimulus sets. It is not reasonable to assume that

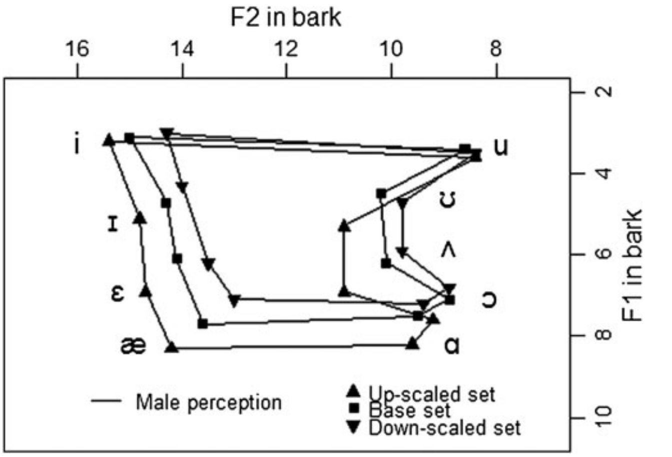


Figure 3 Vowel chart (F2×F1) showing the method of adjustment results of the male perception data. The units are in bark. Each space was connected peripherally by a line. 'Base set' indicates the perception data on the base synthetic stimulus set. 'Down-scaled set' and 'Up-scaled set' refer to the perceptions of the synthetic stimuli that were scaled down or up by 15% from all the formant values of the base synthetic stimulus set, respectively. The vowel sounds are noted near the data points

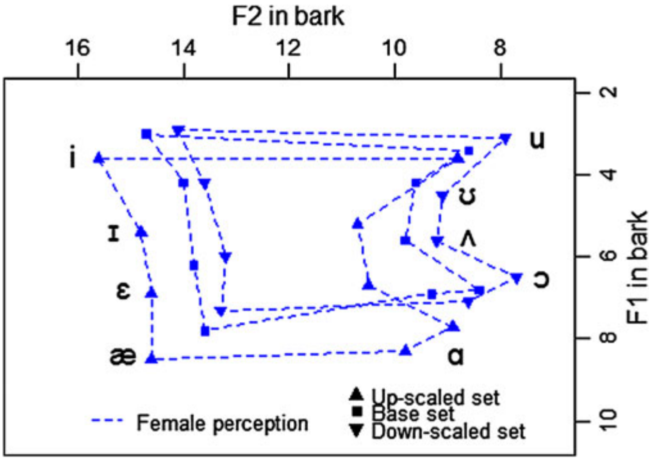


Figure 4 Vowel chart (F2×F1) showing method of adjustment results of the female perception data. The units are in bark. Each space was connected peripherally by a line. 'Base set' indicates the perception data on the base synthetic stimulus set. 'Down-scaled set' and 'Up-scaled set' refer to the perceptions of the synthetic stimuli that were scaled down or up by 15% from all the formant values of the base synthetic stimulus set, respectively. The vowel sounds are noted near the data points

the different dialects spoken in the areas where the current participants came from will have identical vowel phonetics (Thomas 2001; Clopper *et al.* 2005). These considerations were shown in Jacewicz *et al.* (2007: 1468) which tested regional varieties of American English (central Ohio, south-central Wisconsin and western North Carolina) and were unable to conclusively determine that dialect had not affected their data. A similar conclusion was reached by Jacewicz *et al.* (2010: 849). This idiolectal issue might have contributed to the few significantly different cases. However, the authors will not pursue this issue further below and conclude that the perception of the male and female groups tends to show a similar trend within the same synthetic stimulus set.

Second, the participants' perception of the synthetic stimulus sets varied according to the modification rate, which makes sense when considering that the additional synthetic stimulus sets were prepared by scaling down or up by 15% from the base stimulus set. Specifically, the perceived vowel spaces of the male and female groups on the up-scaled set were larger than those on the base set, as shown in Figures 3 and 4. The expansion occurred in the direction of higher formant values in F1 and F2. Conversely, the perceived vowel spaces of the two groups on the down-scaled set were smaller than those on the base set. Tables 1 and 2 support these observations numerically because the average formant values of the down-scaled set on the bottom row increase gradually through the base set and further to the up-scaled set (for male F1, from 571 Hz through 598 Hz to 669 Hz; for female F1, from 551 Hz through 567 Hz to 680 Hz; for male F2, from 1612 Hz through 1729 Hz to 1857 Hz; for female F2, from 1520 Hz through 1652 Hz to 1868 Hz).

To define proportionate relations among the three sets, regression analyses were conducted on all the male and female data of the 972 F1 and F2 values of the down-scaled, base and up-scaled sets using *R*. The perception data of the base set were used as independent variables, and the down-scaled and up-scaled perception data served as dependent variables. A slope of 0.881 (*df*=160, r^2=0.968, *p*=0.00) and an intercept of 66 (*p*=0.00) were obtained from the comparison between the male down-scaled and base sets. The formant values of the down-scaled data were systematically lower than those of the base set. On average, the down-scaled data show a 12% shift from the base set because the intercept is relatively small compared with the formant values in F1 and F2. Note that the base set was reduced by 15% to create the down-scaled set with the same f0. The regression analysis between the male up-scaled and base sets resulted in a slope of 1.054 (*df*=160, r^2=0.951, *p*=0.00) and an intercept of 39 (*p*=0.17). That slope indicates a 5.4% increase in the up-scaled data, which is less than the 15% modification rate. Conversely, the comparison between the female down-scaled and base sets yielded a slope of 0.905 (*df*=160, r^2=0.966, *p*=0.00) and an intercept of 31 (*p*=0.08), in which the down-scaled set is an approximately −9.5% reduction from the base set. The male and female listeners must have tuned to the synthetic vocal tract regardless of having the same f0 in both the base and down-scaled sets. Another comparison between the female up-scaled and base sets resulted in a slope of 1.115 (*df*=160, r^2=0.966, *p*=0.00) and an intercept of 37 (*p*=0.09). Once again, it is notable that the female perception increased by 11.5% for the up-scaled set. The exact cause of the discrepancy between the modification rate and the actual perception rate is unknown in both the male and female groups. Certain individual characteristics of the participants may have contributed to the results; this phenomenon requires further study.

Here, the perception of the three sets of the synthetic stimuli by the male and female participants appears to vary according to the modification rate. This result supports the notion that vowel perception was made independent of the listeners' own vocal tract sizes. Here, the listeners shifted their ideal vowel formants with both changes in the modification rates of the synthetic sets presented to them and the relationship among the four formant values. This was true for the down-scaled space, although f0 was not changed from the base to the down-scaled set. Because the two synthetic stimulus sets were created to simulate speakers with shorter or longer vocal tracts through a uniform scaling from the base synthetic stimulus set, the listeners' perception must have shifted from the base perception space to the down-scaled or up-scaled ranges. In other words, they adjusted their perception to what seemed to be the vocal characteristics of the 'speaker' they were listening to.

3. Production Experiment

In this experiment, production data from the same participants in the perception experiment were collected to compare their perceived with their produced vowel spaces and to discuss later whether their productions are related to their perceptions.

The f0 and the first three formant values of the target vowels were obtained in a clear speaking style.

3.1. Participants

The same 18 American speakers participated in the production experiment.

3.2. Stimuli

For the production experiment, the same nine English monophthongs in the perception experiment were used: *heed*, *hid*, *head*, *had*, *hud*, *odd*, *awed*, *hood* and *who'd*. Each word was printed twice on a page with specific instructions for a clear speaking style. Eighteen words (two instances of the nine words above) appeared in the list alternating front–back or high–low dimensions of the vocal tract. Practice tokens of three words (*heed*, *odd*, *who'd*) were added to the list but not included in the final acoustic analyses. (Participants in a pilot study were observed to produce stably after a few initial words.) The total number of words in the list was 21.

3.3. Procedure

In the production task, the participants recorded the list of English words in a clear speaking style. The first author made all the recordings in a sound isolation room at Haskins Laboratories. The experiment lasted approximately 10–20 minutes per participant, excluding instructions. The experimenter asked the participants to think of two musical notes—*So* for a high tone and *Do* for a low tone—in the hopes that it would elicit more hyperarticulated speech with relatively higher f0 values than with the relatively lower f0 values obtained in the carrier phrase context (Smiljanić & Bradlow 2009). Smiljanić and Bradlow found that entire sentences produced in clear speech tend to be produced with higher f0 than in conversational speech. The carrier sentence was 'I say *hVd (So)*, *hVd (Do)*, *hVd (So)*, *hVd (Do)* again'. The high tone words in the sample sentence were printed in bold face and in larger size to make them stand out and to prompt the speaker to increase the amplitude, f0 and temporal length. The participants practised producing the given words in a clear speaking style. The data collected from the clear speech will be referred to as the 'production data' hereafter. The participants used a headset with the microphone fixed at a distance of approximately 7 cm from the mouth. Their speech was recorded on a digital recorder at a sampling rate of 22,050 Hz in the 16-bit LPCM mode. Participants were asked to imagine a situation in which they were talking to a child or a person who was hard-of-hearing on a noisy street to encourage them to speak more slowly and open their mouths wider. Following Johnson *et al.* (1993), participants were instructed to repeat and make more exaggerated productions of the two words *had* and *odd*, which appeared four times at roughly similar intervals in the list of words. Generally, the experimenter observed that the participants maintained exaggerated jaw openings for open vowels and lip protrusions for round vowels throughout the recording.

3.4. Data Collection and Analysis

The recorded data were transferred to the notebook computer. Goldwave (v5.58) was used to segment and save four productions of each target word as a separately numbered file. Praat (v. 5.2.21 (Boersma & Weenink 2013)) was used to read each participant's file. From the four productions in a tonal sequence of high, low, high and low in a waveform display, the first author listened to all the words and made it a general rule to choose the first production in the display for the production data. However, he sometimes chose the third production in the display if it were considered to represent the style better. The first author then collected the f0 and first three formant values while watching a wideband spectrogram with speaker-specific settings. The measurement was made at one-third the vowel duration (see Yang 1996) using a Praat script. Average acoustic values within a 25 ms window around the measurement point were obtained. Acoustic values of the f0 and formants were collected after checking their plausibility. Halving errors in f0 sometimes occurred in the female data, which were corrected by zooming into the waveform and computing f0 from the average duration of two adjacent glottal cycles. The first three formants were collected carefully within the window size. The number of expected formants within 5 kHz was set on a speaker-specific basis: generally up to 5 for male participants and up to 4.5 for female participants. The number after the decimal point led to a slight shift in the formant measurement in Praat, which yielded more valid formant values. For vowels with closely neighbouring formants such as [ʊ, u], a number of formants were hand-corrected by visually checking the estimated formant values at the centre of the dark band of the spectrographic energy display. In most cases, the Praat measurements seemed valid. The final production data of the f0 and formant values were collected after calculating the average of the two measurements of the same target vowel produced by each participant. The total number of f0 and formant values was 648 (9 vowels × 18 participants × 4 acoustic measurements). The statistical analyses of the production and perception data were performed using SPSS (v.20) and *R* (v.2.14.0).

3.5. Results and Discussion

Tables 3 and 4 provide the average f0, F1, F2 and F3 values in Hz for the vowels produced by the nine male (Table 3) and nine female (Table 4) speakers in a clear speaking style with their average values.

Cross-sex comparisons of the production data will be discussed here briefly because they are not a primary concern of this paper and any such comparison requires appropriate normalization procedures (Flynn 2011; Yang 1992, 1996). Hillenbrand *et al.* (1995) reported 130 Hz as an average f0 value for males and 220 Hz for females. The means of the current male and female production data (146 Hz for males and 270 Hz for females) deviate slightly upward from their average f0 values because the first author asked the participants to produce the vowels in a somewhat hyperarticulated speech style. The average f0 values of the higher vowels were generally higher than those for the lower vowels in both the male and female

Table 3 Average f0, F1, F2 and F3 values in Hz of vowels produced by the nine American males in the clear speaking style

Vowel	f0 m	F1 m	F2 m	F3 m
i	160 (18)	313 (17)	2320 (124)	3081 (125)
ɪ	153 (25)	443 (24)	1996 (116)	2667 (197)
ɛ	139 (21)	626 (43)	1824 (138)	2629 (219)
æ	135 (20)	777 (86)	1751 (92)	2548 (171)
u	167 (26)	327 (28)	949 (151)	2298 (194)
ʊ	149 (21)	477 (45)	1166 (109)	2479 (155)
ʌ	144 (21)	656 (34)	1258 (71)	2620 (136)
ɔ	132 (16)	692 (68)	1001 (122)	2638 (119)
ɑ	132 (19)	801 (53)	1192 (116)	2612 (134)
Average	146 (21)	568 (44)	1495 (115)	2619 (161)

Note: SD is given in parentheses. Average values of each acoustic parameter across all the vowels are given in the bottommost row.

production data, which reflects the vowel intrinsic f0 effect (Lehiste 1967; Whalen & Levitt 1995; Yang 1996).

Figure 5 illustrates the produced vowel spaces of the American male and female participants in bark. In the figure the two vowel spaces appear to have nearly the same pattern but they show the sex difference clearly, with a larger female space in leftward and downward directions from a smaller male space.

A Mann-Whitney *U* test was conducted between the male and female production data in bark using SPSS. All the statistical comparisons of the F1 and F2 values of the nine vowels except for vowel [i] were significantly different between the two sex groups. The *p* value for F1 for the vowel [i] just passed over the alpha level $p<0.05$ ($n=18$, Mann-Whitney $U=83$, $p=0.050$). Thus, one can conclude that the produced vowel spaces of the two sex groups are different.

Table 4 Average f0, F1, F2, F3 values in Hz of vowels produced by the nine American females in the clear speaking style

Vowel	f0 f	F1 f	F2 f	F3 f
i	278 (39)	349 (47)	2994 (92)	3557 (147)
ɪ	281 (38)	589 (74)	2503 (101)	3291 (109)
ɛ	266 (41)	840 (102)	2230 (135)	3236 (124)
æ	261 (43)	1081 (60)	1926 (107)	3066 (126)
u	288 (36)	419 (84)	1283 (246)	2992 (157)
ʊ	285 (41)	657 (89)	1583 (171)	3133 (108)
ʌ	268 (40)	856 (58)	1633 (129)	3162 (135)
ɔ	253 (42)	867 (109)	1223 (175)	3005 (268)
ɑ	254 (46)	975 (89)	1366 (142)	2986 (305)
Average	270 (41)	737 (79)	1860 (144)	3159 (164)

Note: SD is given in parentheses. Average values of each acoustic parameter across all the vowels are given in the bottommost row.

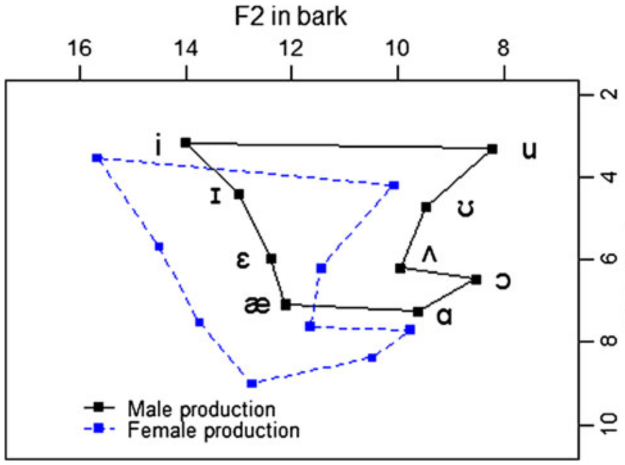


Figure 5 Vowel chart (F2×F1) of the average male and female participants' production in a clear speaking style. The units are in bark. Each space was connected peripherally by a line. The vowel sounds are noted near the data points

4. Comparison of Perceived and Produced Vowel Spaces

In this section, the perceived and produced vowel spaces of the male and female participants will be compared to explore whether perception can be predicted from production. Although their perceptions were similar, the statistical analyses were conducted on the two sex groups separately because male and female participants had different vocal tract lengths.

Figure 6 illustrates the vowel chart of the nine male speakers' perception of the base synthetic stimulus set and their production data. Generally, the produced vowel space is smaller than the perceived vowel spaces, as was observed in Johnson *et al.* (1993). However, the perceived space expansion appears mostly leftward and downward for the front vowels, whereas little expansion is seen for the back vowels, as can be found in the similar pattern of vowel space expansion in Whalen *et al.* (2004a). The produced vowel space of the male speakers appears the smallest compared with their perceived vowel spaces. As was observed in the perception experiment, the perceived vowel spaces gradually expanded from the down-scaled set to the up-scaled set. Two back low vowels [ɔ, ɑ] on the right bottom corner are relatively crowded compared with the other vowel points, which reflects the fact that certain male speakers did not distinguish the vowels clearly.

Table 5 lists the statistical probabilities for the comparisons of produced and perceived vowel formants in bark by the male speakers. Fewer than half of all cases (21 out of 54) yielded statistically significant differences. The produced vowel space of the male speakers falls roughly within the down-scaled set, except for the F2 values of the three front vowels [ɪ, ɛ, æ].

All these male data tend to show that the perception data were larger than the production data. Based on just the data for male speakers, it would be possible to

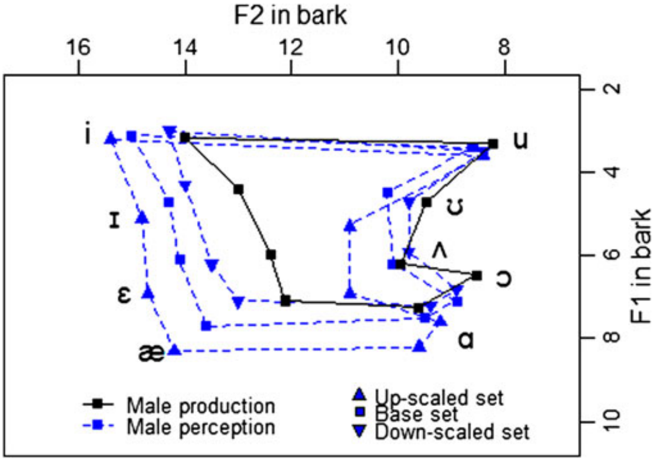


Figure 6 Vowel chart (F2×F1) showing the method of adjustment results for the average male perception and production data. The units are in bark. Each space is connected peripherally by a line. 'Base set' indicates the perception data on the base synthetic stimulus set. 'Down-scaled set' and 'Up-scaled set' refer to the perceptions of the synthetic stimuli that were scaled down or up by 15% from all the formant values of the base synthetic stimulus set, respectively. The vowel sounds are noted near the data points

conclude that the ideal model of the perceived vowel space may have more extreme boundaries. However, such a claim that is based on a simple comparison of perceptions and productions may not work for female data because female productions generally yield a more peripheral vowel space than male data because of the proportionately higher female formants.

Figure 7 shows the perceived and produced vowel charts of the nine female speakers. The produced vowel space of the female speakers appears to be largest compared with the three perceived vowel spaces except for the F2 values of the three front vowels [ɪ, ɛ, æ]. As was observed in Figure 6, the perceived vowel spaces gradually expanded from

Table 5 Statistical results of the comparisons of produced and perceived vowel formants in bark by the male speakers

Scale	Down		Base		Up	
Vowel	F1	F2	F1	F2	F1	F2
i	0.094	0.258	0.546	0.000*	0.546	0.000*
ɪ	0.796	0.001*	0.040*	0.006*	0.000*	0.000*
ɛ	0.258	0.001*	0.489	0.000*	0.008*	0.000*
æ	0.931	0.008*	0.063	0.004*	0.000*	0.000*
u	1.000	0.546	0.546	0.387	0.340	0.666
ʊ	1.000	0.489	0.387	0.161	0.077	0.014*
ʌ	0.161	0.340	0.931	0.605	0.000*	0.008*
ɔ	0.161	0.796	0.019*	0.931	0.004*	0.666
ɑ	0.387	0.340	0.297	0.190	0.000*	0.113

Note: The cells with asterisks indicate significant difference at $p<0.05$.

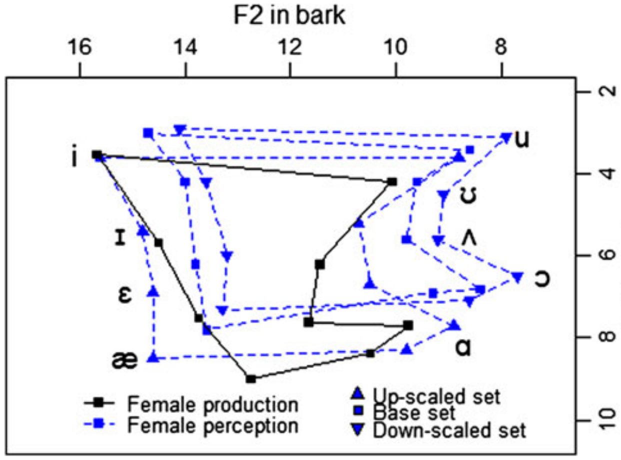


Figure 7 Vowel chart (F2×F1) showing method of adjustment results of the average female perception and production data. The units are in bark. Each space was connected peripherally by a line. 'Base set' indicates the perception data on the base synthetic stimulus set. 'Down-scaled set' and 'Up-scaled set' refer to the perceptions of the synthetic stimuli that were scaled down or up by 15% from all the formant values of the base synthetic stimulus set, respectively. The vowel sounds are noted near the data points

the down-scaled set to the up-scaled set. The expansion from the base set to the up-scaled set appears larger than that in the male spaces. Statistically, the majority of the comparisons between the female group's perceived and produced vowel formant values in bark were found to be significant (39 out of 54 cases in Table 6). Although there were significant differences in seven of the 18 comparisons, the produced vowel space of the female speakers roughly falls within the up-scaled set.

All these data indicate that the female production data generally form a greater vowel space in both the F1 and F2 dimensions than the perception data. Here, a simple comparison between the perceived and produced vowel spaces of the female

Table 6 Statistical results of the comparisons of produced and perceived vowel formants in bark by the female speakers

Scale	Down		Base		Up	
Vowel	F1	F2	F1	F2	F1	F2
i	0.008*	0.000*	0.031*	0.000*	1.000	0.605
ɪ	0.000*	0.004*	0.000*	0.161	0.297	0.387
ɛ	0.000*	0.297	0.000*	0.863	0.136	0.024*
æ	0.000*	0.094	0.000*	0.011*	0.136	0.000*
u	0.006*	0.002*	0.031*	0.014*	0.094	0.040*
ʊ	0.000*	0.000*	0.000*	0.000*	0.004*	0.190
ʌ	0.000*	0.000*	0.000*	0.000*	0.000*	0.006*
ɔ	0.004*	0.000*	0.011*	0.004*	0.931	0.031*
ɑ	0.000*	0.000*	0.000*	0.011*	0.931	0.050

Note: The cells with asterisks indicate significant difference at $p<0.05$.

group may lead to the opposite conclusion from that of the male group data: an ideal model of the perceived vowel space may be far below that of hyperarticulated vowels. Moreover, the vocal tract characteristics of a different group of participants may lead to different conclusions. For example, if more participants with shorter vocal tracts (who would have yielded higher formant values) had been included, such as small females or young children, the difference in their production and the base perception vowel spaces would have been much greater. Further perceptual experiments with synthetic stimuli changing modification rates or recruitments of a new group of participants with much shorter vocal tract lengths may support this argument.

These results have implications for the controversial issue mentioned in the introduction. The proposed hyperspace effect (Johnson *et al.* 1993) relies on a mismatch between a perception space and the production space of the listener's vowels. Previous studies (Whalen *et al.* 2004a, 2004b) have asserted that this does not account for perception by female listeners, given that they have vocal tracts that are shorter than that of the presumed speaker of the synthetic vowels used in the perceptual test. The present experiments show that this is, in fact, the case. The perception space for females is approximately the same size as their production space, and it is in a reduced direction instead of the expanded direction observed in the male data. By contrast, the perception space for males appears to support the hyperspace effect. Whalen *et al.* (2004b: 378) suggest that, although a range of synthetic voices is required to fully test the hyperspace effect, 'such a demonstration could be interpreted as showing only that vowel spaces are different, both in extent and location, and that listeners are sensitive to the information in the vowel signal that tells them what that speaker is like'. The current findings support this assertion. The participants responded to the synthetic stimuli independently of their own produced vowel spaces. Moreover, the male and female participants perceived the three sets of synthetic stimuli according to the modification rate, which indicates that they are sensitive to the information in the synthetic stimuli. As an alternative to the hyperspace effect, a more sophisticated comparison of perceived and produced vowel spaces could be made on an individual basis. In such a study, the synthetic stimulus set must be made to be equivalent to the listener's produced vowel space. If all the comparisons of the perceived and produced vowel spaces of the participants yield expanded directions for an ideal space, a conclusive statement on the existence of the hyperspace effect might be made. However, the current experimental results clearly show that the perception and production of vowels work independently, which is not the result of the hyperspace effect.

5. Conclusion

This study examined the perceived vowel spaces of 18 American male and female participants and compared such perceived vowel spaces with their produced vowel spaces to explore the relationship between perception and production in general. Three synthetic stimulus sets simulating three people with different vocal tract lengths were presented to the participants. Their productions were also controlled to collect hyperarticulated speech tokens to match the best exemplar in the perception task. The results are as follows.

First, the perceptual response patterns were similar across males and females, despite their different-sized vocal tracts, which indicates that both sexes were attributing the synthetic vowels to a similar vocal tract. The vast majority of statistical comparisons between the perceived vowels of the two sex groups were not statistically significant.

Second, the present experiments show that the listeners are willing to adjust their perception to match the apparent vowel space of the speaker they are presented with. Here, although the 'speaker' was a speech synthesizer, listeners applied a reasonable compensation for the presumed vocal tract that produced the speech sounds. There was a steady increase from the down-scaled synthetic stimulus set through the base set to the up-scaled set, although the incremental rate did not exactly match the modification rate (15%) of the down-scaled and up-scaled sets.

Third, there was a significant difference between the production data of the male group and those of the female group, which reflects their vocal tract differences. The female vowel space was much larger than that of the male group.

Finally, a comparison between the perceived and the produced vowel spaces reveals an independent relationship between production and perception. The male vowel space appears to support the hyperspace effect, but the female vowel space definitely does not.

Overall, these results show that the perception of synthetic vowels is made without reference to the listener's own production. Simple matching to the listener's own production to derive a theory can be misleading, as was attempted in previous studies on the hyperspace effect. It is clear that data from a range of speakers (i.e. not just males, at least) need to be incorporated in theories of vowel perception.

ORCID

Byunggon Yang http://orcid.org/0000-0002-6608-5962
Doug Whalen http://orcid.org/0000-0003-3974-0084

References

Barreda S & TM Nearey 2012 'The direct and indirect roles of fundamental frequency in vowel perception' *The Journal of the Acoustical Society of America* 131(1): 466–477. doi:10.1121/1.3662068

Boersma P & D Weenink 2013 *Praat: Doing Phonetics by Computer*. Available at: http://www.fon.hum.uva.nl/praat/ accessed 15 May 2013.

Clopper CG, DB Pisoni & K De Jong 2005 'Acoustic characteristics of the vowel systems of six regional varieties of American English' *Journal of the Acoustical Society of America* 118(3): 1661–1657. doi:10.1121/1.2000774

Fant G 1970 *Acoustic Theory of Speech Production* The Hague: Mouton.

Flynn N 2011 'Comparing vowel formant normalisation procedures' *York Papers in Linguistics Series* 2(11): 1–28.

Hillenbrand J, LA Getty, MJ Clark & K Wheeler 1995 'Acoustic characteristics of American English vowels' *Journal of the Acoustical Society of America* 97(5): 3099–3111. doi:10.1121/1.411872

Jacewicz E, RA Fox, & J Salmons 2007 'Vowel space areas across dialects and gender' ICPhS XVI 1465–1468. Available at: http://www.icphs2007.de/conference/Papers/1252/1252.pdf

Jacewicz E, RA Fox & L Wei 2010 'Between-speaker and within-speaker variation in speech tempo of American English' *Journal of the Acoustical Society of America* 128(2): 839–850. doi:10.1121/1.3459842

Johnson K, E Flemming & R Wright 1993 'The hyperspace effect: phonetic targets are hyperarticulated' *Language* 69(3): 505–528. doi:10.2307/416697

Johnson K, E Flemming & R Wright 2004 'Response to Whalen et al.' *Language* 80(4): 646–648.

Johnson K, EA Strand & M D'Imperio 1999. 'Auditory–visual integration of talker gender in vowel perception' *Journal of Phonetics* 27(4): 359–384. doi:10.1006/jpho.1999.0100

Klatt, DH & LC Klatt 1990 'Analysis, synthesis, and perception of voice quality variations among female and male talkers' *Journal of the Acoustical Society of America* 87(2): 820–857. doi:10.1121/1.398894

Ladefoged P & DE Broadbent 1957 'Information conveyed by vowels' *Journal of the Acoustical Society of America* 29(1): 98–104. doi:10.1121/1.1908694

Lehiste I 1967 *Readings in Acoustic Phonetics* Cambridge, MA: MIT Press.

Nearey TM 1989 'Static, dynamic, and relational properties in vowel perception' *Journal of the Acoustical Society of America* 85(5): 2088–2113. doi:10.1121/1.397861

Peterson GE, & HL Barney 1952. 'Control methods used in a study of the vowels' *Journal of the Acoustical Society of America* 24(2): 175–184. doi:10.1121/1.1906875

Pickett JM 1987. *The Sounds of Speech Communication* Austin, Texas: Pro-ed.

Repp BH, & AM Liberman 1987. 'Phonetic category boundaries are flexible' in SN Harnad (ed.) *Categorical Perception* New York: Cambridge University Press, pp. 89–112.

Samuel AG 1982 'Phonetic prototypes' *Perception and Psychophysics* 31(4): 307–314. doi:10.3758/BF03202653

Sharf B 1970. 'Critical bands' in JV Tobias (ed.) *Foundations of Modern Auditory Theory I* New York: Academic Press, pp. 157–202.

Smiljanić R & AR Bradlow 2009 'Speaking and hearing clearly: talker and listener factors in speaking style changes' *Language and Linguistics Compass* 3(1): 236–264. doi:10.1111/j.1749-818X.2008.00112.x

Thomas ER 2001 *An Acoustic Analysis of Vowel Variation in New World English*. Durham, NC: Duke University Press.

Traunmüller H 1988 'Paralinguistic variation and invariance in the characteristic frequencies of vowels' *Phonetica* 45: 1–29.

Whalen DH & AG Levitt 1995 'The universality of intrinsic F0 of vowels' *Journal of Phonetics* 23(3): 349–366. doi:10.1016/S0095-4470(95)80165-0

Whalen DH, HS Magen, M Pouplier, AM Kang & K Iskarous 2004a 'Vowel production and perception: hyperarticulation without a hyperspace effect' *Language and Speech* 47(2): 155–174. doi:10.1177/00238309040470020301

Whalen DH, HS Magen, M Pouplier, AM Kang, & K Iskarous 2004b 'Vowel target without a hyperspace effect' *Language* 80(3): 377–380. doi:10.1353/lan.2004.0156

Yang B 1990 *Development of Vowel Normalization Procedures: English and Korean*. PhD thesis, The University of Texas at Austin, Austin.

Yang B 1992 'An acoustical study of Korean monophthongs produced by male and female speakers' *Journal of the Acoustical Society of America* 91(4): 2280–2283. doi:10.1121/1.403664

Yang B 1996 'A comparative study of American English and Korean vowels produced by male and female speakers' *Journal of Phonetics* 24(2): 245–261. doi:10.1006/jpho.1996.0013

Zwicker E & E Terhardt 1980 'Analytical expressions for critical-band rate and critical bandwidth as a function of frequency' *The Journal of the Acoustical Society of America* 68(5): 1523–1525. doi:10.1121/1.385079