Japanese native speakers discriminate English vowel formant frequencies better than English native speakers

Sadao Hiroya, Takemi Mochida, and Makio Kashino

NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan

{hiroya, mochida}@idea.brl.ntt.co.jp, kashino@avg.brl.ntt.co.jp

Abstract. *We have shown that thresholds for vowel formant frequency discrimination are significantly correlated with the constraints of directly measured articulatory parameters that produce the vowel, not with the predictions based on the auditory frequency resolution. These findings suggest that native speakers of different languages should have different speech production models and this should result in differences in vowel formant frequency discrimination thresholds. In this study, we examine whether thresholds for vowel formant frequency discrimination between Japanese and English speakers differ for eight English monophthongal vowels. Psychophysics experimental results show different discrimination capabilities for various English vowels in English and Japanese speakers. This indicates that vowel formant frequency discrimination is conducted on the basis of the articulatory constraints of native language.*

1. Introduction

There are many models of speech perception, but no model is generally complete enough to account for all aspects of speech perception. One of the fundamental problems in speech perception is that there is no simple mapping between a phoneme and speech signals. As a solution, the motor theory of speech perception claims that the phonological units of speech are perceived by inferring the articulatory gestures of the speaker, and the relationship between articulatory gestures and speech signals is innately specified (Liberman, 1985). Many psychophysics experiments have shown that this theory can account for a large body of phenomena characteristic of speech perception, including categorical perception (Liberman, 1985), duplex perception (Rand, 1974), audio-visual integration (McGurk, 1976), and sine-wave speech perception (Remez, 1981). Moreover, recent neuroscience studies of speech perception have revealed a strong correlation between speech production and speech perception (Wilson, 2004). Perkell et al. have shown that the speakers' productions of vowel contrasts are related to their discrimination of the contrasts (Perkell, 2004). However, no psychophysics experiments have examined the relationship between the characteristics of speech perception and the directly measured articulatory parameters that produce the speech signals. On the other hand, anti-motor theorists claim that ordinary auditory processes are sufficient to explain speech perception (Stevens, 1980; Kewley-Port, 1998, 2005). It is still controversial whether speech perception involves a process incorporating the constraints of articulatory gestures.

We have examined the relationship between thresholds for vowel formant frequency discrimination and articulatory parameters that produce the vowel (Hiroya, 2006). The major finding was that the discrimination thresholds are significantly correlated with the ratio of formant change to articulatory movements, not with the predictions based on the auditory excitation-pattern model (Glasberg, 1990). This indicates that vowel formant frequency discrimination is not conducted only on the basis of the auditory frequency resolution, which is different from the previous findings (Kewley-Port, 1998).

In view of our findings, we have become interested in vowel formant frequency discrimination by native speakers of different languages because they should have different speech production models based on differences in the languages, and this should result in differences in vowel formant frequency discrimination. Previous studies have shown that there is no statistically significant difference in the thresholds among the Japanese, English, Swedish and Danish speakers for three English monophthongal vowels /æ, ɑ, ʌ/ (Kewley-Port, 2005). In this study, we extensively examined whether thresholds for vowel formant frequency discrimination between Japanese and English speakers differ among the thresholds for eight English monophthongal vowel formant frequency discrimination.

2. Methods

First, articulatory movements and speech signal data of English vowels were obtained from simultaneous measurements. The standard vowels and the test vowels to be discriminated from the standard ones were synthesized using this data. Thresholds for English vowel formant frequency discrimination estimated from psychophysics experiments were compared between Japanese and English speakers.

2.1. Stimuli

Eight monophthongal American English vowels /i, ɪ, e, æ, ɑ, ə, ʊ, u/ were examined. These steady-state vowels were synthesized from the first four formant frequencies, their bandwidths, and the fundamental frequency (F0) using the cascade branch of a formant synthesizer (Klatt, 1980). Simultaneous measurements of articulatory movements and speech signal data were obtained using the EMA system (Kaburagi, 1994) and an audio recording of American English CVC (beat, tit, peck, pack, tot, but, book, toot) produced by an American female English speaker. Formants and F0 were obtained from the middle of each vowel interval. The articulatory data and the palate positions were collected at a sampling rate of 250 Hz. The articulatory parameters were represented by the vertical and horizontal positions of six coils, which were placed on the lower incisor (LI), the upper and lower lips (UL, LL), and the tongue (T1, T2, T3; three positions). The speech signal was recorded at a sampling rate of 16 kHz. The speech signal was first pre-emphasized by first-order differencing. The first four formant frequencies and bandwidths were obtained based on the glottal closure interval. In order to remove the missing or incorrectly identified formants, each analysis was repeated using the linear prediction coding (LPC) order of 12 though 18, and an "optimal" order was chosen by visual inspection (Perkell, 2004). The fundamental frequency (F0) was extracted using the instantaneous frequency amplitude spectrum (Arifiant, 2004).

The synthesized vowel durations were set to 200 ms. The overall amplitude contour had a shallow rise-fall shape for naturalness. Bandwidths and F0 values were ob-

tained from recorded acoustic data, which is different from previous studies (Kewley-Port, 1998). This is because the articulatory parameters differ depending on bandwidths and F0 even when the formant frequencies have the nearly identical values (Hiroya, 2004). The eight vowels are referred to as the standard vowels.

The test vowels to be discriminated from the standard ones were synthesized using the change of the formant frequency of either the first (F1) or second (F2) formant frequency for each vowel. Therefore, eight vowels, with the two discrimination conditions, required 16 sets of test stimuli. For each stimuli set, the step size for a shift in formant frequency was calculated using a log ratio such that steps 13 or 14 would be easily discriminable from the standard. Each set included test vowels for an increment in either the F1 or F2, except for a decrement in F1 of /ə/, F1 of /ɑ/, and F1 of /u/. This was because formant frequencies for these three sets decreased in the recorded data.

2.2. Subjects

Five adult American English Native speakers, three males and two females 20 to 21 years old, and Five adult Japanese Native speakers, three males and two females 25 to 35 years old, with normal hearing participated in the experiments. One of the female English speakers was the speaker who provided articulatory-acoustic data. Subjects were highly trained so that reliable and valid thresholds could be obtained. Each subject performed at least 24 practice sessions. Sessions were run in two hours a day. Subjects took a break for five minutes every two sessions.

2.3. Procedures

Stimuli were presented diotically through Sennheiser HD 280 Pro headphones in a soundproof chamber. In the experiments, the stimuli were retrieved from hard disk with a personal computer and converted to a sampling rate of 22 kHz. Thresholds of formant frequency discrimination were estimated on the basis of the average of the mean reversals from the last five blocks using Levitt's adaptive-tracking method (Levitt, 1971) with the three-down, one-up rule. In each trial, stimuli were presented in a modified, three-interval two-alternative force-choice task with feedback. The standard was always presented in the first of the three intervals, followed by two test stimuli, one identical to the standard and the other selected from the appropriate set of test vowels. Subjects were instructed to select which interval, two or three, contained the vowel that was "different." Each vowel condition was individually tested once or twice. Mean values across individual subject's thresholds were calculated as the group estimates of the formant-frequency threshold for each vowel condition. To preclude the use of the overall level as a cue, the levels of every stimulus were randomized within a 10-dB range from 60 dB SPL.

2.4. Articulatory constraints

A detailed analysis of vocal-tract acoustics has shown that there appear to be ranges of the articulatory parameters for which there is very little change in the formant frequencies and other ranges where the formant frequencies are more sensitive to changes in articulation (Fant, 1960). We defined the ratio of formant change to directly measure mid-sagittal articulatory movement as the articulatory-formant sensitivity (AFS). In this study, this is called articulatory constraints. Formant frequencies were converted to the

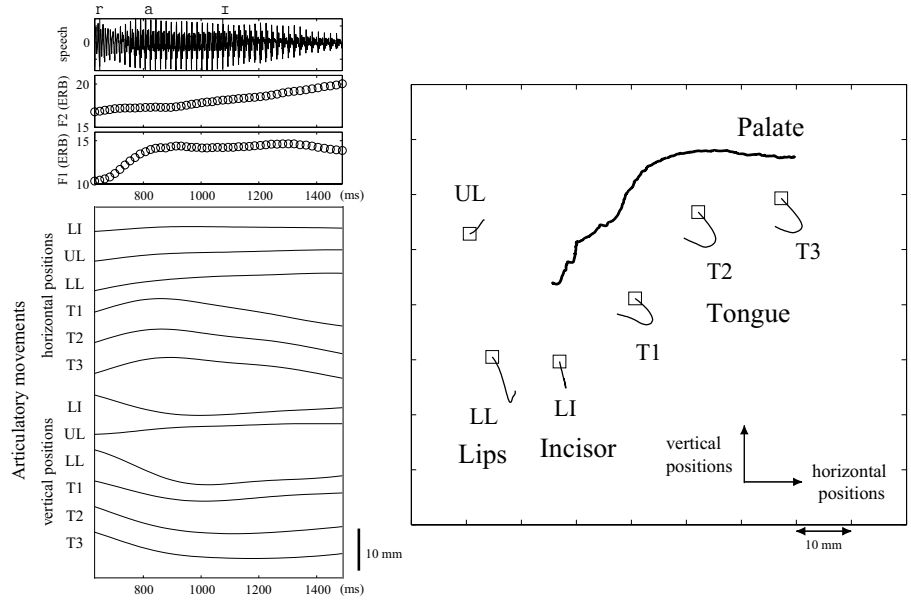

Figure 1. Left: Formant trajectory and articulatory movements for voiced interval of the English word /rise/. Right: Articulatory movements of the mid-sagittal section for the same word. Squares indicate the positions at the beginning of /r/.

equivalent rectangular bandwidth (ERB) scale, which represents an auditory frequency resolution (Glasberg, 1990). This ratio quantifies the non-linear relationship between articulatory parameters and formant frequencies in vowels on the basis of the directly measured articulatory-acoustic data.

Here, we explain the AFS by using an English diphthongal vowel. Figure 1 shows an example of formant trajectory and articulatory movements of the English word /rise/ on the time axis and articulatory movements for a mid-sagittal section. In Fig. 2, the first and second formant frequencies are plotted against for the average shift of every articulatory position from /r/. The ratio, AFS, was calculated as linear regression coefficients of formant frequency to articulatory movements for the F1 and F2 of every standard vowel. The ratio of formant change to articulatory movements is not uniform when formant frequencies are changed as a result of small perturbation of the articulatory parameters. For example, the second formant frequency is hardly changed at all by small perturbations of the articulatory parameters in the transition between vowel /r/ and /a/, but largely changed in vowel /ɪ/. The former condition is called small AFS and the latter large AFS. If discrimination of vowel formant frequency is conducted on the basis of the articulatory constraints, it is expected that discrimination thresholds will be small for small AFS and large for large AFS.

3. Results

3.1. Comparison of discrimination thresholds with AFS

First, we compared thresholds for vowel formant frequency discrimination of English speakers with the AFS. Discrimination thresholds were significantly correlated with the

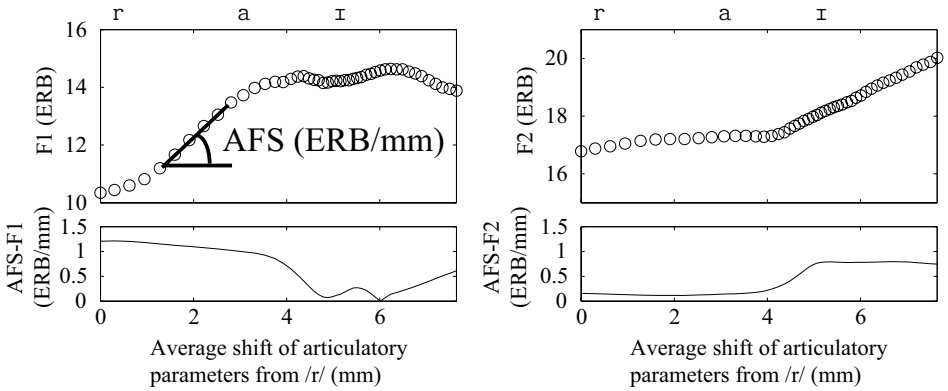

Figure 2. Top: The first and second formant frequencies for the average shift of every articulatory position from /r/. Bottom: The AFS values of the first and second formant frequencies.

AFS ($r = 0.89$). This suggests that, for English speakers, articulatory constraints of their native language are used in vowel formant frequency discrimination.

Next, we compared Japanese speaker's formant frequency discrimination thresholds in English vowels with the AFS. The correlation ($r = 0.71$) was less significant than for English speakers. This would support the idea that speech production models of Japanese speakers differ from those of English speakers.

3.2. Cross-language analysis

Figure 3 shows the thresholds for vowel formant frequency discrimination of English and Japanese speakers. An analysis of variance (ANOVA) showed that there is a statistically significant difference in the thresholds between Japanese and English speakers for F2 of /æ/ and F2 of /e/ ($p < 0.05$). Interestingly, although the thresholds for English speakers were larger than those for Japanese ones, the thresholds were highly correlated with the AFS. One possible reason for the smaller thresholds for Japanese speakers is that the relationship between articulatory parameters and formant frequencies in English vowels is not represented well in the central nervous system for Japanese speakers. Another possible reason may be the categorical perception for vowels. However, determining the exact reason for the thresholds requires further investigation.

This result would support the idea that the articulatory constraints, which are used for vowel formant frequency discrimination, are acquired by the learning of one's native language, not innately specified (Liberman, 1985). However, it is still controversial whether the articulatory gestures are directly inferred in the central nervous system during vowel perception (Liberman, 1985). We may say that auditory processes are adapted on the basis of the relationship between formant frequencies and articulatory gestures (Stevens, 1989). Thus, a more detailed analysis is required.

3.3. Sine-wave speech stimuli

As a control experiment, we examined thresholds for sine-wave speech (Remez, 1981). We used four sinusoids signal, each of which had the same frequency as the corresponding first four formants of the standard vowel sound. The amplitudes of the sinusoids were

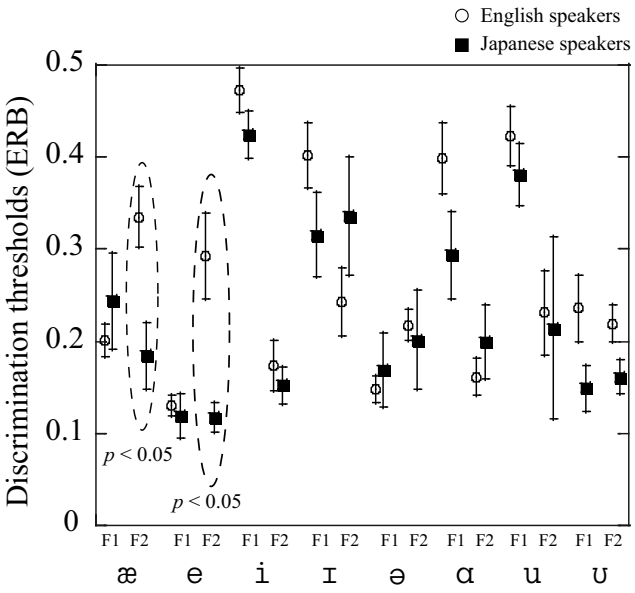

Figure 3. Average values of thresholds of English vowel stimuli (either F1 or F2 for each vowel) calculated for English and Japanese speakers. Vertical bars indicate the standard error of mean.

set equal to those of the first four formants of the test vowel and changed over time in the same way. Such signals are quite different from natural speech: They lacked the harmonic structure of speech and do not have the pulsing structure associated with voicing. Six standard stimuli (F2 of /e/, F1 and F2 of /ə/, F1 and F2 of /u/, and F2 of /ʊ/) were used. Figure 4 shows the thresholds of sinusoid signals between English and Japanese speakers. Results of ANOVA showed that there is no statistically significant differences in the thresholds between English and Japanese speakers, especially for F2 of /e/. This indicates that the peripheral level of processing is a similar between English and Japanese speakers and that different thresholds for English vowels between English and Japanese speakers can not be explained by the auditory frequency resolution.

4. Discussion

We calculated the distances of articulatory parameters among the standard Japanese and English vowels. Average articulatory positions on the basis of 359 Japanese sentences produced by five Japanese male subjects were used for the standard Japanese vowels /a i u e o/ (Hiroya, in press). For English vowels, the standard vowels from the experiments were used. Figure 5 shows a result of a multi-dimensional scaling analysis of the distances. A comparison between Figs. 3 and 5 indicates that English vowels /e, æ, ɑ, ʊ/ are relatively far from Japanese vowels and their thresholds differ between Japanese and English speakers.

To examine this in detail, we performed an English vowel identification test for Japanese and English subjects. In this task, Japanese subjects were asked to categorize the English standard vowels according to the Japanese vowels /a i u e o/ while English subjects categorized them into the English vowels. Results showed that English subjects could almost perfectly identify the English vowels, except for vowels /ɪ/ and /ə/. For

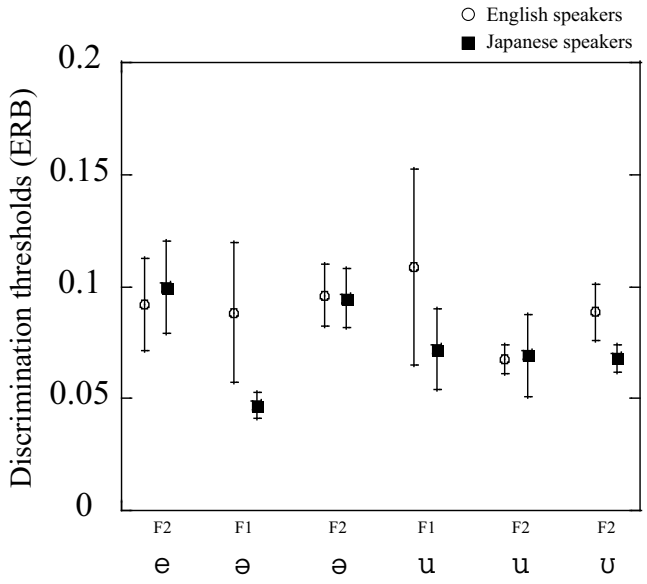

Figure 4. Average values of thresholds of sine-wave speech stimuli calculated for English and Japanese speakers. Vertical bars indicate the standard error of mean.

Japanese subjects, English vowels /i/, /ɑ/, /ə/ and /u/ seemed to correspond to Japanese vowels /i/, /a/, /a/ and /u/, respectively. However, English vowels /æ/, /e/, /ɪ/ and /ʊ/ seemed to be lie between Japanese vowels /a/ and /e/, /a/ and /e/, /i/ and /e/, and /u/ and /o/, respectively. This would support the idea that native English speakers raise the thresholds of the second formant frequency changes for vowels /æ/ and /e/ of their own language in order to robustly identify the phonological units of those vowels.

5. Conclusions

We examined whether thresholds for vowel formant frequency discrimination between Japanese and English speakers differ for eight English monophthongal vowels and found different discrimination capabilities for English vowels /æ/ and /e/ in English and Japanese speakers. This indicates that vowel formant frequency discrimination is conducted on the basis of the articulatory constraints of native language.

Acknowledgments

The authors thank Ewen Chao of the California Institute of Technology for conducting the experiments and many useful discussions.

References

Arifiant, D., Tanaka, T., Masuko, T., and Kobayashi, T. Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency. *IEICE Trans. Inf. & Syst.*, E87-D(12):2812–2820, 2004.

Fant, G. *Acoustic theory of speech production.* Monton & Co.'s-Gravenhage, 1960.

Glasberg, B.R. and Moore, B.C.J. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47:103–138, 1990.

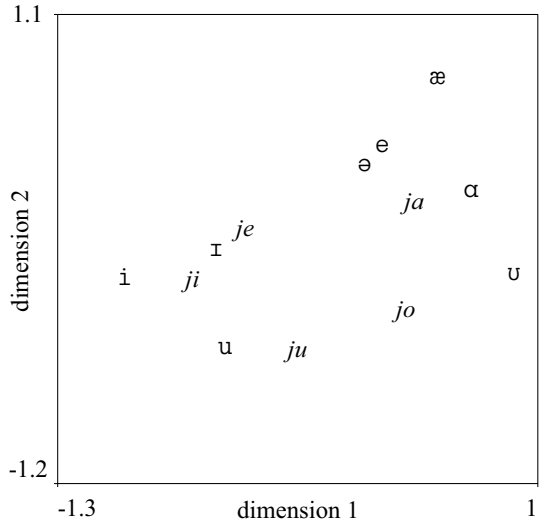

Figure 5. Multi-dimensional scaling analysis of articulatory parameters among the Japanese and English vowels. The vowels 'ja', 'ji', 'ju', 'je', and 'jo' indicate the Japanese vowels /a/, /i/, /u/, /e/, and /o/, respectively.

Hiroya, S., Mochida, T., and Kashino, M. Reducing redundancy in acoustic-to-articulatory inversion by fundamental frequency. In *Proc. From sound to sense: 50+ years of discoveries in speech communication*, page 19, 2004.

Hiroya, S., Mochida, T., and Kashino, M. Articulatory gestures, not auditory frequency resolution, determine formant frequency discrimination thresholds in vowels. In *Proc. The 29th ARO MidWinter Meeting*, page 249, 2006.

Hiroya, S. and Mochida, T. Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs. *Speech Communication*, in press.

Kaburagi, T. and Honda, M. Determination of sagittal tongue shape from the positions of points on the tongue surface. *J. Acoust. Soc. Am.*, 96(3):1356–1366, 1994.

Kewley-Port, D. and Zheng, Y. Auditory models of formant frequency discrimination for isolated vowels. *J. Acoust. Soc. Am.*, 103(3):1654–1666, 1998.

Kewley-Port, D., Bohn, O-S., and Nishi, K. The influence of different native language systems on vowel discrimination and identification. *J. Acoust. Soc. Am.*, 117(4):2339, 2005.

Klatt, D.H. Software for cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, 67(3):971–995, 1980.

Levitt, H. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49(2B):467–477, 1971.

Liberman, A.M. and Mattingly, I.G. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

McGurk, H. and Macdonald, J. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M., and Zandipour, M. The distinctness of speaker's production of vowel contrasts is related to their discrimination of the contrasts. *J. Acoust. Soc. Am.*, 116(4):2338–2344, 2004.

Rand, T.C. Dichotic release from masking for speech. *J. Acoust. Soc. Am.*, 55(3):678–680, 1974.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. Speech perception without traditional speech cues. *Science*, 212(22):947–950, 1981.

Stevens, K.N. Acoustic correlates of some phonetic categories. *J. Acoust. Soc. Am.*, 68(3):836–842, 1980.

Stevens, K.N. On the quantal nature of speech. *J. Phonetics*, 17(1-2):3–45, 1989.

Wilson, S.M., Saygin, A.P., Sereno, M.I., and Iacoboni, M. Listening to speech activates motor areas involved in speech production. *Nat. Neuroscience*, 7(7):701–702, 2004.