INTERSPEECH 2009 BRIGHTON

Relative importance of formant and whole-spectral cues for vowel perception

Masashi Ito[1], *Keiji Ohara*[2], *Akinori Ito*[1], *Masafumi Yano*[2]

[1] Graduate School of Engineering, Tohoku University, Japan
[2] Research Institute of Electrical Communication, Tohoku University, Japan
itojin@makino.ecei.tohoku.ac.jp

Abstract

Three psycho-acoustical experiments were carried out to investigate relative importance of formant frequency and whole spectral shape as cues for vowel perception. Four types of vowel-like signals were presented to eight listeners. The mean responses for stimuli including both formant and amplitude-ratio feature were quite similar to those for the stimuli including only formant peak feature. Nonetheless reasonable vowel changes were observed in responses for stimuli including only amplitude-ratio feature. The perceived vowel changes were also observed even for stimuli including neither of these features. The results suggested that perceptual cues were involved in various parts of vowel spectrum.

Index Terms: vowel perception, formant frequency, whole spectrum model

1. Introduction

Formant frequency is one of the effective features to represent vowels. Peterson and Barney indicated that American-English vowels were well characterized by the first and second formant frequencies ($F1$ and $F2$) when they were uttered by a single speaker [1]. Based on the psycho-acoustical experiments, Klatt suggested the formant frequency to be the most important cue for vowel perception, in which only the formant frequency affected perceived vowel, while spectral tilt and filtering did not affect the perceived phonemic information although they were clearly audible [2].

However, this formant hypothesis has some problems with modeling vowel perception as claimed by Bladon [3]. For instance, it is not straightforward to estimate reliable formant frequencies for vowels spoken by female and children while human beings can perceive these vowels without any difficulties. Bladon proposed a whole spectrum model assuming that vowel might be perceived on the basis of its gross spectral shape without extracting formant frequencies.

The authors investigated perceptual effects of formant peak and whole-spectral shape in a series of experiments [4]. The result revealed that vowel could be consistently perceived for stimulus of which $F1$ or $F2$ peak was suppressed in the spectrum if the remained spectral shape was preserved as close as the original vowel. Further, it was indicated that an amplitude ratio, defined as the relative amplitude of the third to the first formant peaks, might be effective cue for articulatory place (front/back) of the perceived vowel. These results suggested that formant frequency was not exclusive cue for vowel perception. A similar result was obtained in another study at least for steady-state vowels [5].

The purpose of the present study is to evaluate relative importance of formant frequency and whole-spectral shape as a cue for vowel perception. Based on the finding of the previous studies, we will focus on the amplitude ratio of the whole-spectral shape, which was shown to correlate with articulatory place of perceived vowels. The second formant frequency is also known to affect the articulatory place of vowels. Thus, it is expected that relative importance of the amplitude ratio and $F2$ might be estimated by comparing responses for the stimuli including these features. In order to evaluate general characteristics of the perception, three psycho-acoustical experiments are carried out in which stimuli are designed to represent close, middle, and open vowels.

2. Experiment I

Japanese close vowels /u/ and /i/ are discriminated by the place of articulation. The perceptual cues for these vowels are investigated in the first experiment.

2.1. Method

2.1.1. Stimuli

Four types of vowel-like stimuli are used in the experiment. The first is a steady-state vowel of which spectral envelope is determined by a five-formant cascade-Klatt synthesizer [6]. The formant frequencies and bandwidths for the synthesizer are constant except for the second formant frequency ($F1$ = 250, $F3$ = 2500, $F4$ = 3500, $F5$ = 4500 Hz, $B1$ = $B2$ = $B3$ = 62.5 Hz, and $B4$ = $B5$ = 125 Hz), while $F2$ ranges from 1000 to 2250 Hz with a 125-Hz step. Fig. 1a shows spectral envelopes provided by the synthesizer. The spectral envelope was orderly raised with $F2$ increase at high frequency as shown in the figure. Hereafter a term $E_C(f)$ refers to logarithmic amplitude of the spectral envelope at frequency f determined by the cascade-Klatt synthesizer. For each of the eleven $F2$s, a harmonic signal is synthesized with $F0$ of 125 Hz, where logarithmic amplitude of the n-th harmonic component is set to $E_C(n \cdot F0)$, while phase of the component is determined by a cosine-phase manner. The duration of the signal is 400 ms, in which initial and final parts gradually increases and decreases with half-cosine windows of 40 ms. These signals are called as '*control*' stimuli corresponding to a simple model for naturally-uttered vowels.

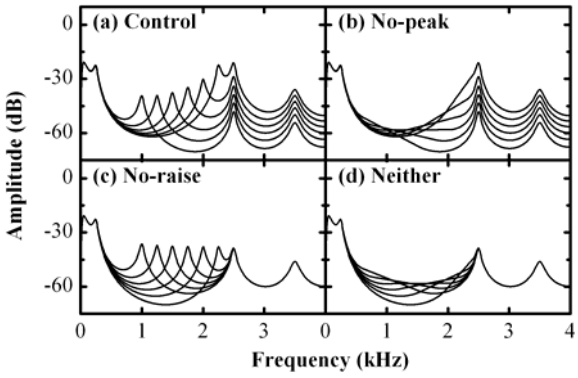

Figure 1: *Spectral envelopes of stimuli (Exp. I)*

The second type is the '*no-peak*' stimulus of which *F*2 peak is removed from the '*control*' one. The spectral envelope $E_P(f)$ is determined from $E_C(f)$ as follows.

$$E_P(f)=\begin{cases}E_C(f_L)+\alpha\cdot(f-f_L) & \ldots \quad f_L\le f\le f_H\\ E_C(f) & \ldots \quad others\end{cases} \tag{1}$$

$$\alpha=(E_C(f_H)-E_C(f_L))/(f_H-f_L) \tag{2}$$

where f_L and f_H represent lower and higher edge frequencies of the peak suppression ($F1 < f_L < F2 < f_H < F3$). In order to make both $E_P(f)$ and $dE_P(f)/df$ to be continuous at the edge frequencies, the frequencies are determined to satisfy Eq. (3) for each $E_P(f)$.

$$\alpha=dE_C(f)/df\big|_{f=f_L}=dE_C(f)/df\big|_{f=f_H} \tag{3}$$

As shown in Fig. 1b, it is quite difficult to estimate the original *F*2 in the suppressed envelope using spectral peak-picking. On the other hand, amplitude ratio of the original envelope is fairly preserved in the suppressed envelope, and orderly increases with the suppressed *F*2.

The third type is the '*no-raise*' stimulus of which amplitude ratio is kept constant regardless of *F*2, although it has prominent *F*2 peak. The spectral envelope $E_R(f)$ is determined from $E_C(f)$ as follows.

$$E_R(f)=\begin{cases}E_C(f)+\beta\cdot(f-F1)+\gamma & \ldots \quad F1\le f\le F3\\ R(f) & \ldots \quad others\end{cases} \tag{4}$$

$$\beta=(R(F3)-E_C(F3)-\gamma)/(F3-F1) \tag{5}$$

$$\gamma=R(F1)-E_C(F1) \tag{6}$$

where $R(f)$ represents $E_C(f)$ of which *F*2 is 1500 Hz that provides a reference amplitude ratio for every $E_R(f)$. As shown in Fig.1c, the envelope E_R(f) is constant at the frequency below *F*1 and above *F*3 regardless of the *F*2.

The last type is the '*neither*' stimulus including neither formant peak nor amplitude ratio features. The spectral envelope $E_N(f)$ is determined from the '*no-peak*' envelope $E_P(f)$ as follows.

$$E_N(f)=\begin{cases}E_P(f)+\beta'\cdot(f-F1)+\gamma' & \ldots \quad F1\le f\le F3\\ R(f) & \ldots \quad others\end{cases} \tag{7}$$

$$\beta'=(R(F3)-E_P(F3)-\gamma')/(F3-F1) \tag{8}$$

$$\gamma'=R(F1)-E_P(F1) \tag{9}$$

As shown in Fig.1d, $E_N(f)$ does not have prominent *F*2 peak in the spectrum. Further, its amplitude ratio is constant regardless of the suppressed *F*2. The spectral differences caused by the change of original *F*2 are limited in a range of 250 to 2500 Hz. For the envelopes $E_P(f)$, $E_R(f)$, and $E_N(f)$, harmonic signals are synthesized in a similar manner of the '*control*' stimuli. A total of 44 stimuli are used in the experiment. The amplitude of every stimulus is modified to have an identical root-mean-square power.

2.1.2. Procedure

The stimuli are presented to listener via a headphone (HDA-200, Sennheiser) in a soundproof chamber with an averaged most comfortable level. The order of the presentation is random, in which the 44 stimuli mentioned above are randomly presented regardless of their types (*control*, *no-peak*, *no-raise*, and *neither*). Each listener is required to answer one of five Japanese vowels (/a, i, u, e, o/) which is perceptually close to the presented stimulus. The re-presentation of the stimulus is not allowed. After responding the 44 stimuli in a set, listener can advance to the next set. Every listener responds to 30 sets of the stimuli. The listeners are eight adult males. They are native Japanese speakers and did not report any hearing impairments.

2.2. Results

From the responses of all listeners, mean recognition rate of each vowel was calculated for each stimulus. Fig. 2 shows the mean rates, in which most of the stimuli were perceived as either vowel /u/ or /i/ regardless of the stimulus types. There were no stimuli for which mean recognition rates of the other three vowels exceeded 12.9 % in this experiment.

As shown in Fig. 2a, the mean recognition rate of /u/ was gradually decreased with *F*2 while the rate of /i/ was contrarily increased with *F*2 for '*control*' stimuli. This result indicated that the stimuli involved crucial features to determine articulatory place of the perceived vowels. As shown in Fig.2c, recognition rates for '*no-raise*' stimuli were quite similar to those for the '*control*' stimuli. Since both of these stimuli included formant-peak features, the result supported that formant frequency might be effective cue for articulatory place of the perceived vowel. This was consistent with the previous study of Klatt [2].

For '*no-peak*' stimuli, mean recognition rates of /u/ and /i/ were monotonically decreasing and increasing with the suppressed *F*2, respectively (Fig.2b). This meant that articulatory place could be fairly perceived even for the stimulus of which *F*2 peak was removed from the spectrum. Thus, amplitude ratio seemed to be another effective cue for articulatory place of vowel as reported in the previous studies [4,5]. To compare the recognition rates for '*no-raise*' and '*no-peak*' stimuli, the former was more close to the rates for the '*control*' stimuli. Further, perceptual ambiguity was greater in the '*no-peak*' stimuli. These results implied that formant peak feature might affect to the articulatory place perception more greatly than amplitude ratio feature.

The most striking result was obtained from the responses for '*neither*' stimuli. The recognition rates of vowels /u/ and /i/ were orderly changed with the suppressed *F*2 for these stimuli when it was greater than 1625 Hz (Fig. 2d). Since the '*neither*' stimuli included neither formant peak nor amplitude ratio feature (Fig. 1d), another feature, involved in a spectrum from 250 to 2500 Hz, should be taken into account to explain this result.

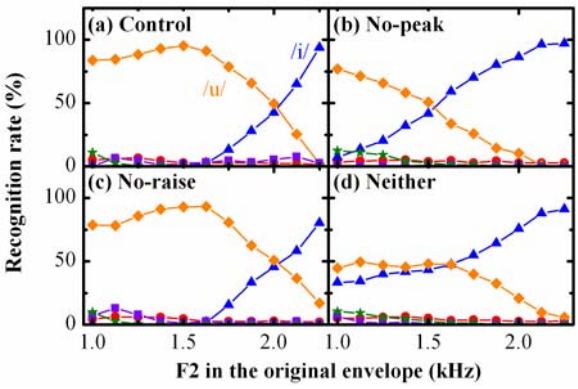

Figure 2: *Mean recognition rate (Exp. I)*

3. Experiment II

To investigate generality of the characteristics obtained in the experiment I, a similar experiment is carried out using stimuli representing Japanese close vowel /u/ and middle vowels /o/ and /e/ which are also discriminated by place of articulation.

3.1. Method

Four types of stimuli are synthesized in a similar way of the experiment I, where *F*1 for the cascade-Klatt synthesizer is set to 500 Hz while the other parameters are identical to those used in the previous experiment. Fig.3 shows spectral envelopes of the stimuli. A total of 44 stimuli are synthesized and presented to listeners in random order as same as the experiment I. For each stimulus, one of five Japanese vowels is required to be selected. Every listener responds to 30 sets of the stimuli. The listeners are eight adult males, who are identical to those participated in the experiment I.

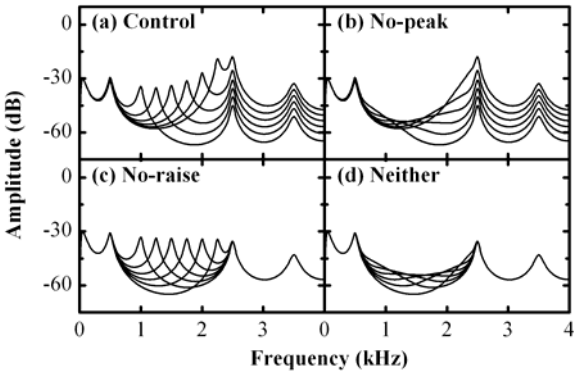

Figure 3: *Spectral envelopes of stimuli (Exp. II)*

3.2. Results

Fig. 4 shows mean recognition rates of the vowels for each stimulus, where most of the stimuli were perceived as vowel /o/, /u/, or /e/. There were no stimuli for which mean recognition rates of the other two vowels (/a/ or /i/) exceeded 15.4 % in this experiment.

As shown in Fig. 4a, perceived vowels for the '*control*' stimuli were systematically changed in an order of /o/, /u/, and /e/ according to *F*2. The maximum recognition rates of these vowels were 82.9, 84.6, and 80.8 %, respectively. These rates were a little smaller than maximum rates of vowels /u/ and /i/ (93.8 and 95.4 %) for the '*control*' stimuli in experiment I. The smaller maximum rates reflected greater listener variances in the responses. The recognition rates for the '*no-raise*' stimuli (Fig.4c) showed a quite similar pattern to those for the '*control*' stimuli (Fig.4a) as same as the experiment I. Thus, formant peak feature was still effective in this experiment.

The perceived vowels for the '*no-peak*' stimuli were consistently changed with the same order of the '*control*' stimuli according to the suppressed *F*2 (Fig. 4b). A remarkable difference from the '*control*' stimuli was that maximum recognition rate of vowel /u/ for the '*no-peak*' stimuli saturated at only 42.9 %. This increased recognition rates of vowels /o/ and /e/ with 12.1 % on the average. Nonetheless the result supported amplitude ratio to be another effective cue for perceiving articulatory place of the vowels as similar as the experiment I.

The mean recognition rates for the '*neither*' stimuli indicated weak but apparent dependency for the suppressed *F*2 as shown in Fig. 4d. The rates of vowels /e/ and /o/ were consistently increased and decreased with the *F*2, respectively, as similar as the '*control*' stimuli although perceptual ambiguity was much greater in the '*neither*' stimuli. This result supported the hypothesis that there might be another features correlated with perceptual articulatory place in the spectral shape ranged from 500 to 2500 Hz of these stimuli.

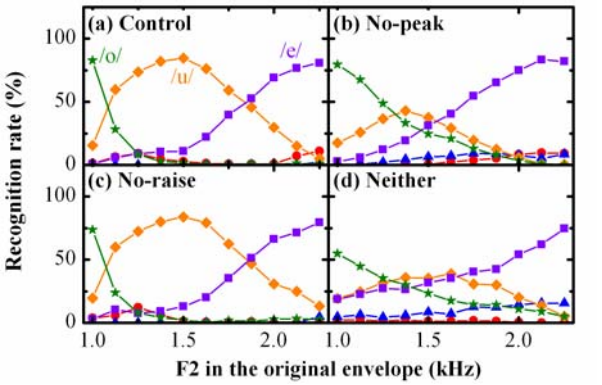

Figure 4: *Mean recognition rate (Exp. II)*

4. Experiment III

Finally, a similar experiment is again carried out using stimuli which represent Japanese close vowel /u/, middle vowels /o/ and /e/, and open vowel /a/.

4.1. Method

Four types of stimuli are synthesized in a similar manner of the experiments I and II, where *F*1 for the cascade-Klatt synthesizer is set to 750 Hz. Fig. 5 shows spectral envelopes for the stimuli. As shown in the figure, small differences are observed at frequency lower than *F*1 in the spectral envelopes for the '*control*' and '*no-peak*' stimuli (Fig. 5a and 5b). These differences simply reflect a boosting effect caused by adjacent formant peaks (*F*1 and *F*2) of the synthesizer, which is reported in [6]. The listeners and procedure are identical to those in the experiments I and II.

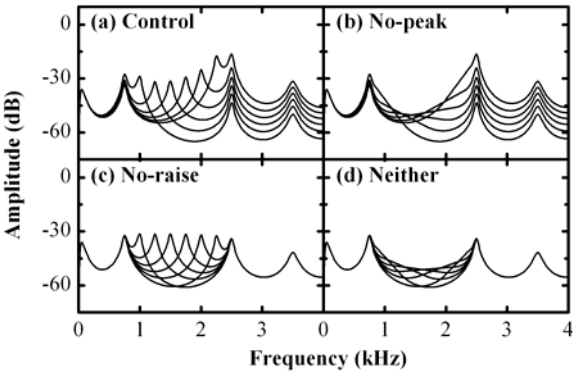

Figure 5: *Spectral envelopes of stimuli (Exp. III)*

4.2. Result

Fig. 6 shows mean recognition rates of the vowels for each stimulus, in which the rate of vowel /i/ did not exceed 7.1 % for every stimulus in this experiment. The maximum recognition rates for the '*control*' stimuli, were 50.4 % for vowel /o/ at *F*2 = 1000 Hz, 62.1 % for vowel /a/ at *F*2 = 1250 Hz, 36.7 % for vowel /u/ at *F*2 = 1750 or 1875 Hz, and 68.8 % for vowel /e/ at *F*2 = 2250 Hz. These are rather smaller than

those of the experiment I and II, which reflected the greatest listener variances of this experiment.

The mean recognition rates for the '*control*' stimuli (Fig. 6a) were quite close to those for the '*no-raise*' stimuli (Fig. 6c) as same as the previous experiments. Further, recognition rates for the '*no-peak*' stimuli did not greatly differ from those for the '*control*' and '*no-raise*' stimuli except for vowel /u/ that was hardly perceived for the '*no-peak*' stimuli (Fig. 6b). These results consistently supported that both formant peak and amplitude ratio feature might be effective perceptual cues for articulatory place of vowels. Moreover, mean recognition rates for the '*neither*' stimuli showed apparent dependencies for the suppressed *F*2. The result well agreed with those observed in the experiments I and II.

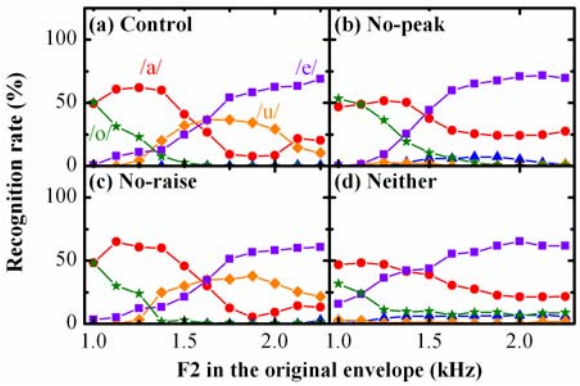

Figure 6: *Mean recognition rate (Exp. III)*

5. Discussion

To evaluate the consistency of the three experiments, the root-mean-square of the difference was calculated in the mean recognition rate between every pair of the stimulus types as shown in Fig. 7. The difference between the '*control*' and '*no-raise*' stimuli was the smallest in every experiment (3.4 to 4.0 %). Further, the difference between the '*no-peak*' and '*neither*' stimuli was rather smaller (9.0 to 10.1 %). This meant that the existence of the amplitude ratio feature did not greatly affect perceived articulatory place of vowels.

On the other hand, there were rather great differences between the '*control*' and '*no-peak*' (13.4 to 23.6 %), and between the '*no-raise*' and '*neither*' stimuli (15.4 to 22.7 %) in every experiment. This meant that the existence of the formant peak was a crucial factor for the articulatory place perception. These results led us to the conclusion that the formant peak feature had greater importance than the amplitude ratio feature for perceiving articulatory place of the vowels.

While these results almost agreed with the previous study of Klatt [2], it should be noted that the listeners could consistently perceive articulatory place of vowel even for the '*no-peak*' stimulus in which the formant peak was completely suppressed in the spectrum (Fig. 2b, 4b, and 6b). This result suggested that the formant peak feature was not exclusive cue for the perception as reported in the previous studies [4, 5].

The amplitude ratio of the third to the first formant peaks was a promising candidate for another cue for the perception. However, the definition of the feature should be carefully reconsidered because the articulatory place was fairly perceived for the '*neither*' stimulus which did not include this feature (Fig.2d, 4d, and 6d). Although the spectral envelope of the stimuli did not greatly change with the suppressed *F*2 (Fig. 1d, 3d, and 5d), the results obtained in all the three experiments indicated that these spectral shapes were informative in perceiving articulatory place of vowels. It is still open to discussion whether there might be a unified perceptual cue for reasonably explaining our results. Further research on this problem would clarify the appropriate definition of the perceptual cue for vowels.

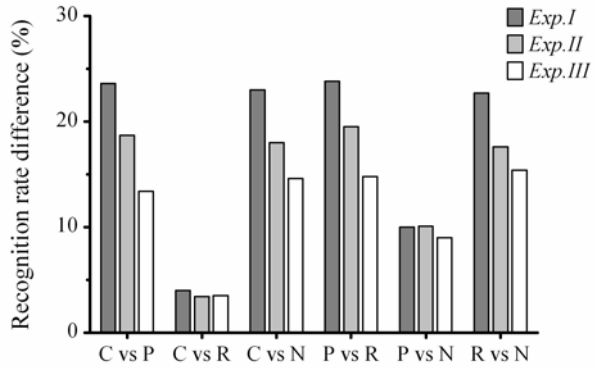

Figure 7: *Differences of the response pattern (C:Control, P:no-Peak, R: no-Raise, N: Neither)*

6. Acknowledgements

This work was supported by KAKENHI (21700282 and 17075003) of the Ministry of Education, Culture, Sports, Science and Technology.

7. References

[1] Peterson, G. and Barney, H. L., "Control methods used in a study of vowels," J. Acoust. Soc. Am. 20(2), 528-535, 1952.

[2] Klatt, D. H., "Speech processing strategies based on auditory models," in The Representation of Speech in the Peripheral Auditory System, Elsevier, Amsterdam, 181–196, 1982.

[3] Bladon, R. A. W., "Arguments against formants in the auditory representation of speech," in The Representation of Speech in the Peripheral Auditory System, Elsevier, Amsterdam, 95–102, 1982.

[4] Ito, M., Tsuchida, J. and Yano, M., "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Am. 110(2), 1141-1149, 2001.

[5] Kiefte, M. and Kluender, K. R., "The relative importance of spectral tilt in monophthongs and diphthongs," J. Acoust. Soc. Am. 117(3), 1395-1404, 2005.

[6] Klatt, D. H., "Software for a cascade-parallel formant synthesizer," J. Acoust. Soc. Am. 67(3), 971-995, 1980.