

ACOUSTICAL LETTER

Perception of multiple series of English /ra/–/la/ continuum having different end frequencies of formant transitions

Kanako Tomaru[1,*] and Takayuki Arai[2]

[1]*Graduate School of Science and Technology, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Department of Information and Communication Sciences, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*

(*Received 22 October 2013, Accepted for publication 4 December 2013*)

Keywords: Formant transition, End frequencies, English /r/–/l/ contrast, Identification, Discrimination
PACS number: 43.71.+m, 43.71.−k, 43.71.Es [doi:10.1250/ast.35.166]

1. Introduction

The first (F1) and third (F3) formant transitions are responsible for the English /r/–/l/ contrast [1]. The characteristics of F1 are temporal, whereas those of F3 are spectral. For /r/, the starting frequency of F1 remains stable for little while and the transition begins gradually; its F3 transition shows an upward movement [1–4]. For /l/, on the other hand, the starting frequency of F1 remains longer, resulting in a steeper transition than that for /r/; its transition of F3 is almost straight or may have a slight downward movement [1–4].

Such acoustic characteristics have an effect on perception [1–4]. When listening to continuous syllables, native speakers of English hear /r/ and /l/ according to the F1 and F3 transitional trajectories. Such a perceptual change can be visualized in identification and discrimination functions. The identification function illustrates how the number of /ra/ responses decreases at the "categorical boundary" [5], a point where listeners start hearing the stimuli as /la/. The discrimination function depicts the highest discrimination performance ("discrimination peak" [6]) for a stimulus pair, one member of which belongs to /r/ and the other of which belongs to /l/. In the present study, we investigated whether patterns of these functions are identical for multiple series of the /ra/–/la/ continuum with different end frequencies of transitions under equal transitional conditions.

2. Stimuli

The terminal frequencies of transitions *F1t*, *F2t*, and *F3t* in Fig. 1 correspond to the first to third formant frequencies (F1, F2, and F3) of the following vowel, i.e., /a/ in a /ra/–/la/ continuum. Thus, our three series of the /ra/–/la/ continuum (Series 1, Series 2, and Series 3) had different formant frequencies for the steady state of /a/. The series were generated by using XKL [7,8]. *F1t*, *F2t*, and *F3t* were decided on the basis of utterances of three male speakers, i.e., Speaker 1, Speaker 2, and Speaker 3, from the TIMIT corpus [9]. For each speaker, F1, F2, and F3 were obtained from the steady state of the vowel [ʌ] in "pro**nun**ciation" in the following sentence: "Clear pronunciation is appreciated." The series created on the basis of Speaker 1 was called Series 1, and so on. The values of F1, F2, and F3, which respectively equal the values of *F1t*, *F2t*, and *F3t*, are provided in Table 1.

Next, we added transitions to the formant frequencies of the steady state, or the terminal frequencies. The transitional conditions were set to be equal in the following manner. First, the values of the starting frequencies of the transitions (*F1s*, *F2s*, and *F3s* in Fig. 1) were found by multiplying the terminal frequencies by the ratio provided in Table 2. The ratio in Table 2 corresponds to that of the starting frequencies to the ending frequencies of the stimuli created by MacKain *et al.* [3]. Next, we linearly interpolated both ends of the frequencies in the transitions. For example, in the case of the F2 transition of Series 1, its *F2s* at each step was specified by multiplying the F2 of Speaker 1 in Table 1 by the ratio given in Table 2. After finding *F2s*, we interpolated *F2s* and *F2t* linearly. For F3, there was also a point of inflection (*F3i*) at 135 ms (see Fig. 1 and Table 2). *F3i* was found in the same way as *F2s*.

The same method was used for F1. However, note that *F1s* was identical among the steps because the F1 transitional variation was temporal not spectral. F1 varied in terms of the time point where the transition started ascending. Following Polka and Strange [2], the F1 temporal characteristic was varied in ten equal steps from Step 1 at 10 ms to Step 10 at 55 ms for every series. For Step 1, for example, the transition started ascending to the end frequency at 10 ms, and so on.

The stimuli also contained the fourth (F4) and fifth (F5) formants. Their values were identical among the series and were the default values of the synthesizer (see Fig. 1). Because these formants have not been focused on as important factors for the /r/–/l/ contrast [1–4], we assumed that using default values for these formants would not affect the present investigation. In addition, the values of F0 were obtained from the original utterances of the speakers.

The length of a syllable was 350 ms including 100 ms rising and falling periods. Digital outputs from the synthesizer (16-bit resolution and 10 kHz sampling rate) were converted to 16-bit resolution with a 16 kHz sampling rate.

*e-mail: himawari.kanako@gmail.com

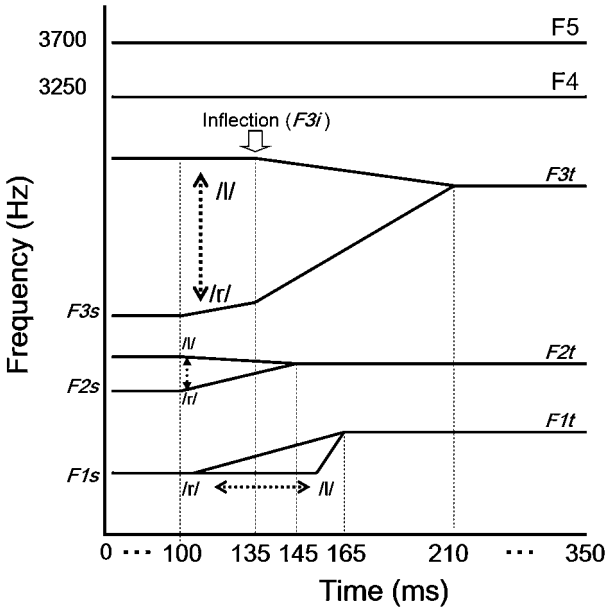

Fig. 1 Trajectories of the first five formants of the synthesized stimuli.

Table 1 F1, F2, and F3 (in Hz) obtained from steady state of the vowel [ʌ] for each speaker.

	F1 (= *F1t*)	F2 (= *F2t*)	F3 (= *F3t*)
Speaker 1	703	1,449	2,806
Speaker 2	670	1,357	2,788
Speaker 3	655	1,446	2,406

Table 2 Ratios (in percentage) of starting frequency (*F1s*, *F2s*, and *F3s*) to terminal frequency (*F1t*, *F2t*, and *F3t*). For F3, the ratio of the inflection (*F3i*) to *F3t* is also provided.

Step	*F1s* to *F1t*	*F2s* to *F2t*	*F3s* to *F3t*	*F3i* to *F3t*
1	56.1	89.0	57.7	61.6
2	56.1	90.4	63.0	66.2
3	56.1	91.7	67.6	70.7
4	56.1	93.0	72.2	74.8
5	56.1	94.4	77.1	79.3
6	56.1	95.7	82.2	83.4
7	56.1	96.4	87.1	88.4
8	56.1	97.8	91.7	92.3
9	56.1	99.2	96.4	97.1
10	56.1	100.7	101.4	101.4

3. Experiment

3.1. Participants

Five native speakers of English with normal hearing served as the participants (mean age = 23.2 years old).

3.2. Procedure

Participants completed an AXB discrimination test followed by a two-alternative-forced-choice (2AFC) identification test. Stimuli were presented diotically via Sennheiser HDA 200 headphones at a listening level comfortable to the participants in a sound proof studio. All sessions were carried out using Praat software [10]. The whole experiment took approximately 30 mins. No feedback was given during the experiment.

3.2.1. AXB discrimination

Participants judged whether the second sound matched the first sound or to the third sound. The inter stimulus duration was 0.3 s. For each series, listeners made 16 judgments for each of eight syllable pairs that were two steps apart along a continuum. The total number of judgments was 384 (16 judgments × 8 pairings × 3 series). Trials were blocked by three series of the continuum, and stimuli were presented within each block. A practice without feedback was given before the experiment.

3.2.2. 2AFC identification

Participants identified a syllable either as "ra" or "la." They heard four repetitions of one syllable. Thus, participants made a total of 120 judgments (4 repetitions × 10 stimuli × 3 series). A practice without feedback preceded the main task.

3.3. Results

One participant was excluded from the analysis because of her misunderstanding of the tasks.

3.3.1. 2AFC identification

For ease of discussion, the results of the identification task are provided first. The plotted scores in Fig. 2 represent the percentage response of /ra/ calculated for each listener. Figure 2 also shows the overall percentage response of /ra/ for each series in the form of a fitted function. The percentage response of the four listeners was fitted by using the following function:

$$y = \frac{100}{1 + e^{a(x-b)}},$$

where y represents the percentage response of /ra/ and x represents the step number. The parameter a corresponds to the slope of the curve. The parameter b corresponds to the 50% crossover point, or the categorical boundary. The fitted sigmoidal function had parameter values that minimize the residual sum of squares. The value of a was 2.0 for Series 1, 1.5 for Series 2, and 1.4 for Series 3. The values of b for Series 1, Series 2, and Series 3 were 6.1, 6.0, and 6.0, respectively. The values of the multiple correlation coefficient of the functions were above 0.90: 0.96 for Series 1, 0.93 for Series 2, and 0.96 for Series 3.

3.3.2. AXB discrimination

Figure 3 shows the results for Series 1, Series 2, and Series 3. From the identification functions, we expected a discrimination peak at pair 5–7, which crossed the categorical boundary (the cross-boundary pair), for all series. To assess the difference in performance between the cross-boundary pair and the other pairs, we combined the results for the /r/-side of pairs (within-/r/ pairs), and the /l/-side of pairs (within-/l/ pairs). The correct rates were compared among the three types of pairs, i.e., the cross-boundary pair, the within-/r/ pairs, and the within-/l/ pairs. ANOVA with repeated measures found the main effect of the pair type for Series 1 ($F(2, 6)$, $p = 0.002$) and Series 2 ($F(2, 6)$, $p = 0.024$). There was no main effect for Series 3 ($F(2, 6)$, $p = 0.869$). Further

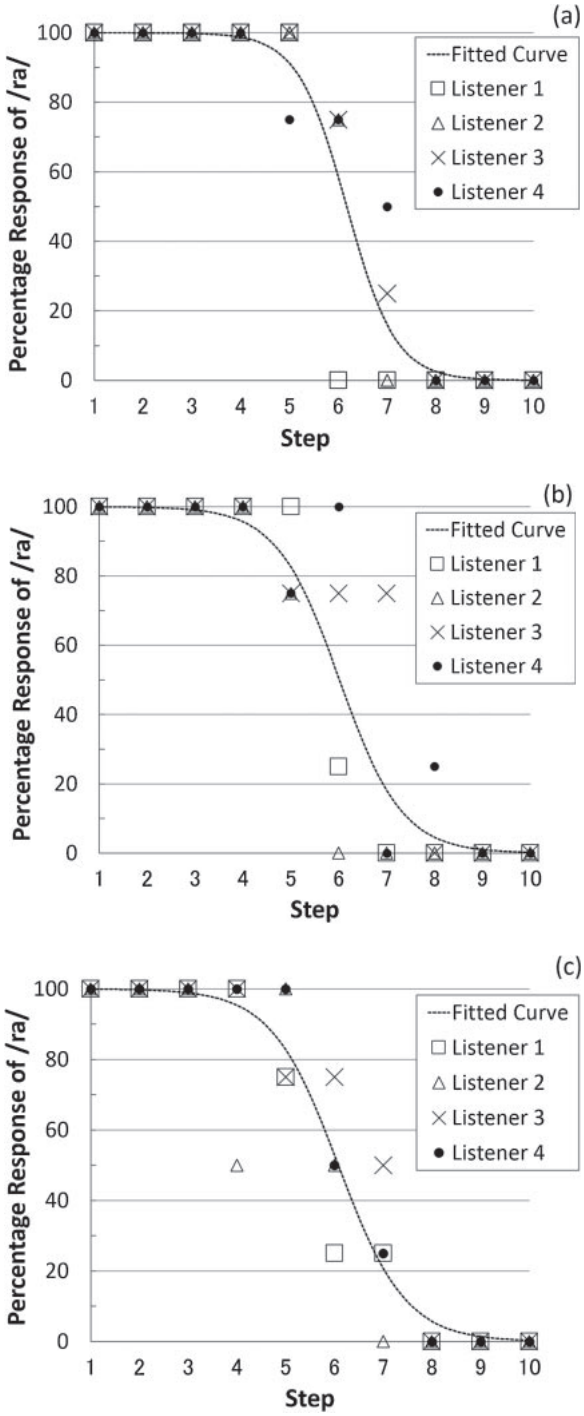

Fig. 2 Percentage response of /ra/ calculated for each listener according to the series, i.e., Series 1 (a), Series 2 (b), and Series 3 (c). Overall percentage response for each series is shown as a fitted function.

post-hoc testing was carried out for Series 1 and Series 2. For Series 1, the correct rate for the cross-boundary pair was significantly higher than that for the within-/r/ pairs ($p = 0.014$) and that for the within-/l/ pairs ($p = 0.009$). For Series 2, the significance was marginal: $p = 0.073$ for the cross-boundary pair versus the within-/r/ pairs and $p = 0.058$ for the cross-boundary pair versus the within-/l/ pairs.

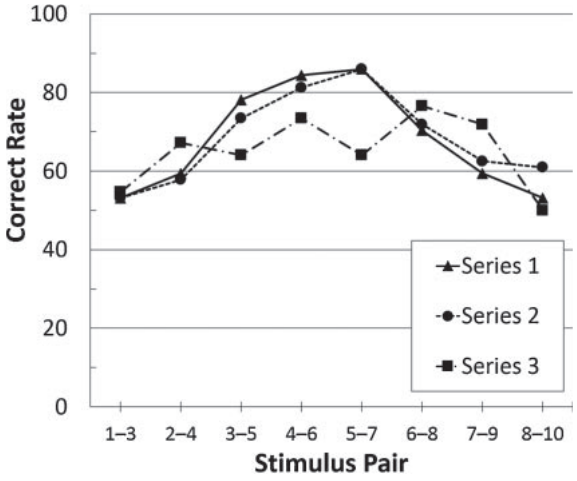

Fig. 3 Discrimination functions for Series 1, Series 2, and Series 3.

Table 3 F1, F2, and F3 (in Hz) of the vowel [ʌ] obtained in the present study and those provided by former studies [11,12]. The table also provides the distance (in Hz) between F2 and F3 and ratio (in percentage) of F2 to F3.

	Vowel	F1	F2	F3	Distance (F3 − F2)	Ratio of F2 to F3
Series 1		703	1,449	2,806	1,357	52
Series 2		670	1,357	2,788	1,431	49
Series 3	/ʌ/	655	1,446	2,406	960	60
Peterson & Barney [11]		640	1,190	2,390	1,200	50
Hillenbrand *et al.* [12]		623	1,200	2,550	1,350	47

4. Discussion and conclusions

Although the slopes of the identification functions seemed to vary slightly, listeners' perception similarly changed from /ra/ to /la/ according to the change in formant transitions in all series. In addition, the location of the categorical boundary was almost identical among the three series. Therefore, the identification functions of the three series with different end frequencies were shown to be nearly equal. In contrast, the discrimination functions appeared to differ. A discrimination peak was obtained only for Series 1 and Series 2. The function of Series 3 did not have an obvious peak.

The small distance between the terminal frequencies of the second and third formants, i.e., *F2t* and *F3t*, may account for the irregular discrimination function of Series 3. The distance between F2 and F3 of Series 3, i.e., 960 Hz (Table 3), was smaller than that of the other series. For the steady state of the vowel /ʌ/, the distance between F2 and F3 of less than 1,000 Hz is particularly small [11,12] (Table 3). Thus, it is possible that a peak in the discrimination function is observable only when F2 and F3 are separated sufficiently. In a further investigation, the effects of the distance between F2 and F3 on discrimination judgments should be clarified in detail using sufficient data.

Acknowledgements

The contents of this report were presented at the Fall Meeting of ASJ in 2012.

References

[1] R. D. Kent and C. Read, *Onsei no Onkyo Bunseki* (T. Arai and T. Sugawara, Trans.) (Kaibundo, Tokyo, 2004) (Original work published 1992).

[2] L. Polka and W. Strange, "Perceptual equivalence of acoustic cues that differentiate /r/ and /l/," *J. Acoust. Soc. Am.*, **78**, 1187–1197 (1985).

[3] K. S. MacKain, C. T. Best and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Pshycholinguist.*, **2**, 369–390 (1981).

[4] R. Dalston, "Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults," *J. Acoust. Soc. Am.*, **57**, 462–469 (1975).

[5] A. M. Liberman, K. A. Harris, H. S. Hoffman and B. C. Griffith, "The discrimination of speech sounds within and across phoneme boundaries," *J. Exp. Psychol.*, **54**, 358–368 (1957).

[6] S. Rosen and P. Howell, "Auditory, articulatory, and learning explanations of categorical perception in speech," in *Categorical Perception: The Groundwork of Cognition*, S. Harnad, Ed. (Cambridge University Press, New York, 1990), Chap. 4, pp. 113–160.

[7] D. H. Klatt, "The new MIT speech VAX computer facility," in *Speech Communication Group Working Papers IV, Research Laboratory of Electronics, MIT, Cambridge*, pp. 73–82 (1984).

[8] D. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, **87**, 820–857 (1990).

[9] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, **9**, 351–356 (1990).

[10] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]," Version 5.3.39, retrieved 6 January 2013 from http://www.praat.org/.

[11] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, **24**, 17–184 (1952).

[12] J. Hillenbrand, L. A. Getty, M. J. Clark and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, **97**, 3099–3111 (1995).