1-3-12

English speakers' perception of synthesized /ra/-/la/ continua with the same range of formant transition starting from different formant values[*]

☆Kanako Tomaru, Takayuki Arai (Sophia Univ.)

Introduction

1.1 Background

Acoustic difference between English /r/ and /l/ can be found in their formant frequencies. Firstly, the third lowest formant frequency (F3) plays an important role when English-speaking listeners perceptually differentiate /r/ from /l/ [1]. The main difference is the direction of F3 transition. For /r/, you see upward F3 transition from low to high between /r/ and the following vowel. For /l/, on the other hand, there is no such upward F3 transition; the transition looks almost straight. You can see such difference in Fig. 1, which illustrates formant patterns of liquids.

Secondly, we find /r/-/l/ difference also in the first lowest formant frequency (F1). As illustrated in Fig.1, onset steady state of F1 is short and transition is gradual for /r/. On the other hand, for /l/, onset steady state of F1 is long and transition duration is brief [1, 2]. Brief transition for /l/ comes from the fact that tongue movement is more rapid for /l/ production than /r/ production [2]. For /l/, tongue tip contacts with alveolar for a shorter moment than for /r/, and it rapidly moves to produce the following vowel [2]. This durational difference in terms of F1 transition is called temporal variation in this paper.

1.2 Purpose of the present research

Researches on /r/ and /l/ perception have long been benefitted from speech synthesis techniques [3-8]. A number of researchers examined perception of /r/-/l/ using synthesized continua to find what exact cue is important for /r/-/l/ difference for native speakers of English [6], or to show if such cues are effective for non-native speakers of English as well, or not [5,7,8]. In addition, some researches imply that listeners rely on other information about speaker's individual characteristics in speech perception [9-11]. For example, it is said that second language phonetic training with talker variability [9-11] is more effective than training with one speaker. But, even after effective training with various talkers, learners find some speakers' utterance is easier to discriminate, or to identify, than other speakers' utterance [10]. And, it is worth examining what exact characteristics of speaker makes that happens.

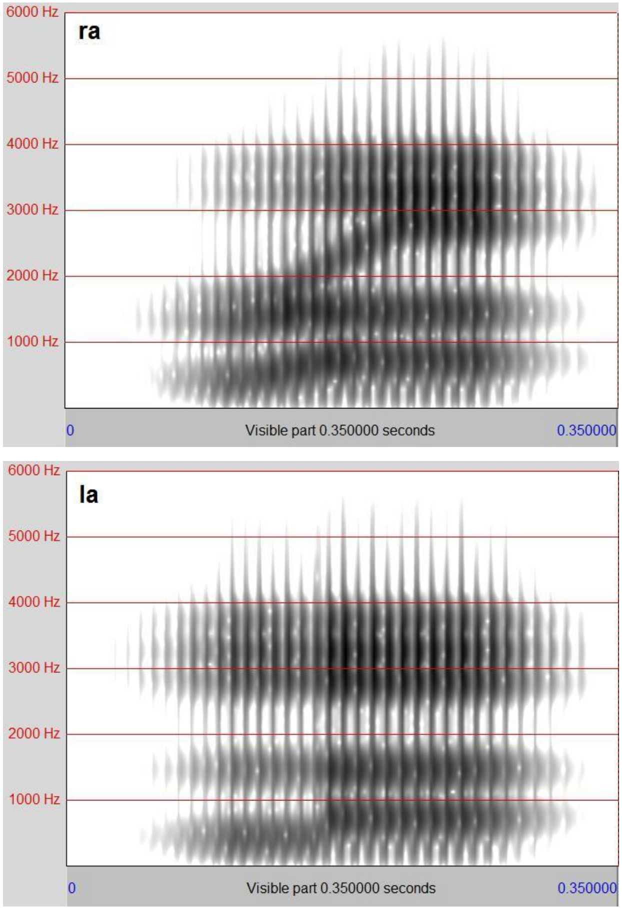

Fig. 1 Spectrograms of /ra/ and /la/, which are synthesized for the present research.

The current research focuses on vocal tract size variation and tests if it has effects on perception of /r/-/l/ continuum. Variation of vocal tract size appears as variation of formant frequencies. For example, a tall adult man tends to have a long

[*] 英語話者による/ra/-/la/連続体の知覚—遷移幅を固定し開始周波数を変化させた場合，渡丸嘉菜子，荒井隆行(上智大・理工).

vocal tract, and it should be shown as having low formant frequencies. So, in the present study, we created three series of /ra/-/la/ continuum which had variation in terms of the first three formant frequencies, which partially reflect the size of vocal tract.

In the present study, we had three types of /ra/-/la/ continuum which had the same formant transition range, but had different actual values for the first three formants. Through the perceptual experiments with English-speaking listeners, we found similar perceptual patterns across the continua, although we did obtain some differences.

2 Method

2.1 Stimuli

We synthesized three series of /ra/-/la/ continua using cascade-formant software synthesizer designed by Klatt and Klatt [4]. We synthesized /ra/-/la/ syllables based on three male speakers' utterance from TIMIT corpus [12]. The speakers were MDPB0, MKAM0, and MTJM0, who will be called as Speaker1, Speaker2, and Speaker3, respectively, hereafter. The speakers recorded the same sentence: "Clear pronunciation is appreciated." Synthesized syllables were 350ms long including 100ms rising and falling periods.

To synthesize continua, we first obtained formant frequency values of each speaker from a vowel [ʌ] in "pronunciation" in the sentence. First three formants for each speaker are shown in Table 1. These values were used as steady state formant values that came after transition in /ra/-/la/ series (*d1*, *d2*, and *d3* in Fig. 2). The steady state durations are also shown in parenthesis in Table 1.

Transitions started at 100ms because we had 100ms rising period before onset of a liquid (Fig.2). Transition ranges for F1, F2, and F3 were adopted from MacKain *et al.* [5]. Following MacKain *et al.*, we calculated transition ratio of starting frequencies (*Fs1*, *Fs2*, and *Fs3* in Fig. 2) to the following vowel's steady state format values. *Fs* varied in ten nearly equal steps from /ra/ configuration (step1) to /la/ configuration (step10). For F2 transition, for example, Fs2 of each step can be found by multiplying a speaker's F2 in Table 1 by given formant transition ratio in Table 2. For F3, there is a point of inflection at 135ms (Fig. 2). To add temporal variation to F1, duration of *F1 steady state* and that of *F1 transition* (Fig. 2) was varied following Polka & Strange [6]. The temporal duration varied in ten equal steps from step1 (10-ms F1 steady state and 55-ms F1 transition) to step10 (55-ms F1 steady state and 10-ms F1 transition). The three continua also had different F0. F0 values were tracked from originally recorded sentences.

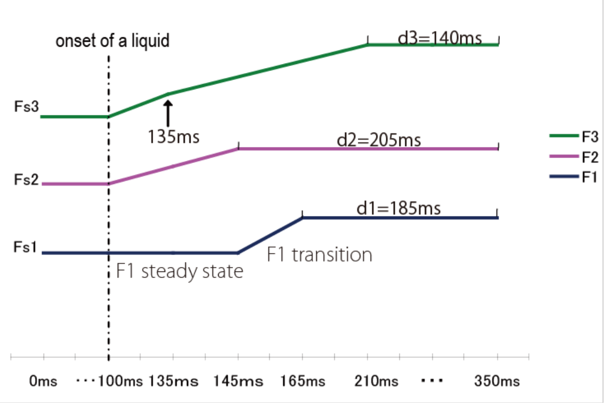

Fig. 2 Parameters for stimuli synthesis. F1, F2, and F3 are shown in blue, pink and green, respectively. The horizontal axis shows duration.

Table 1. F1, F2 and F3 of [ʌ] spoken by three speakers in Hz.

	F1 **(165-350ms)**	**F2** **(145-350ms)**	**F3** **(210-350ms)**
Speaker 1	703	1449	2806
Speaker 2	670	1357	2788
Speaker 3	655	1446	2406

Table 2. Formant transition ratio of an onset frequency to the frequency of the following vowel in percent.

step	**F1**	**F2**	**F3**	**F3(at 135ms)**
1	56.1	89.0	57.7	61.6
2	-	90.4	63.0	66.2
3	-	91.7	67.6	70.7
4	-	93.0	72.2	74.8
5	-	94.4	77.1	79.3
6	-	95.7	82.2	83.4
7	-	96.4	87.1	88.4
8	-	97.8	91.7	92.3
9	-	99.2	96.4	97.1
10	-	100.7	101.4	101.4

We named these series of continuum as Continuum1, Continuum2, and Continuum3, after speakers' identification number. Digital outputs from the synthesizer (16-bit resolution and 10-kHz sampling rate) were converted to 16-bit resolution and 16-kHz sampling rate.

2.2 Participants

Five native speakers of English participated in an AXB discrimination test, as well as in a two-forced-choice (2FC) identification test. Participants ranged in age from 20 to 29 years old. None reported any known hearing problems.

2.3 Procedure

Participants were tested in a sound-proof studio in Arai Laboratory, Sophia University. Participants completed the AXB discrimination test followed by the identification test. Stimuli were presented diotically via Sennheiser HDA 200 headphones at participants' comfortable listening level. All sessions were carried out using Praat software [13].

2.3.1 AXB test

First, participants completed the AXB test. They made 16 judgments for each of the 8 possible pairs of stimuli that were 2 steps apart along the continuum, i.e., 1-3, 2-4, 3-5, 4-6, 5-7, 6-8, 7-9, and 8-10. Therefore, they made total of 128 judgments for each speaker (16 judgments × 8 pairings = 128 judgments). Trials were blocked by three continua, and stimuli were randomly presented to participants within each block. Participants took a short practice session to be familiarized with stimuli and procedure before the experimental session.

2.3.2 Identification test

After the discrimination test, participants completed the 2FC identification test, where they were to choose what they've heard is either "ra" or "la". Participants heard four repetitions of each of the ten stimuli, all of which were presented randomly to participants. Thus, participants made total of 40 judgments for each continuum (4 repetitions × 10 stimuli). Again, participants took a practice session to be familiarized with stimuli and procedure before the experimental session.

Results

2.4 AXB discrimination

Results of the discrimination test are shown in Fig. 3. As a quick look at the figures tells you, discrimination function does not exactly match across continuum types. The function for Continuum2 clearly has a discrimination peak around pair 5-7. Similarly, Continuum1 also has a kind of discrimination peak around that area, although the peak doesn't stand out as it does in Continuum2. Discrimination function for Continuum3 is rather flat showing no clear discrimination peak. In fact, we can see a little dent at pair 5-7 where we received discrimination peaks for the other continua.

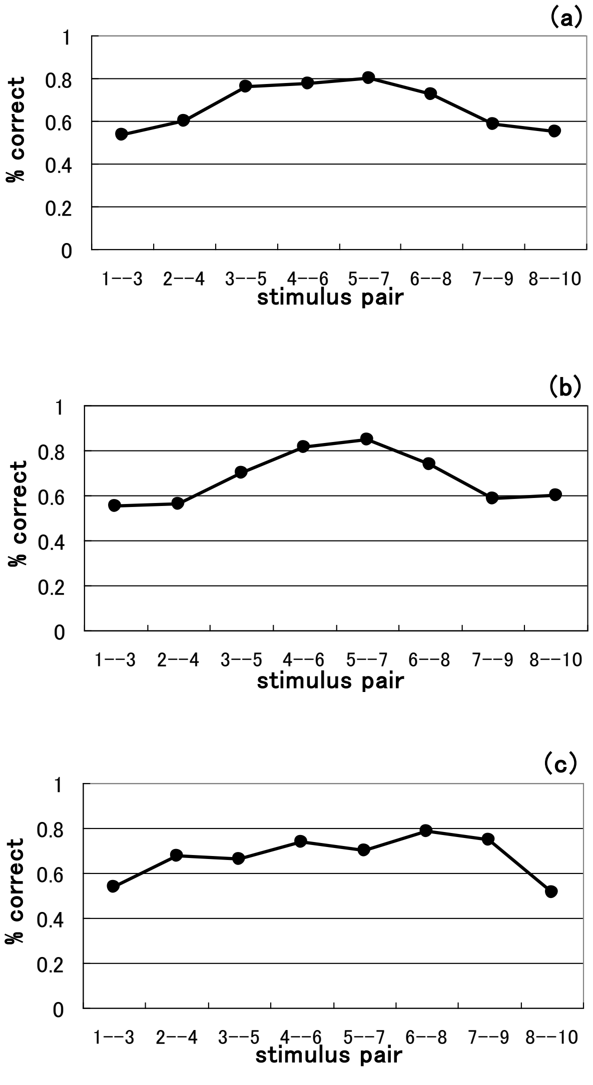

Fig. 3 Obtained discrimination function for Continuum1 (a), Continuum2 (b), and Continuum3 (c).

2.5 2FC identification

Results of the identification test are shown in Fig. 4. Although identification functions look similar among the continua, we can also notice a little difference: while categorical boundaries for Continuum1 and Continuum2 locate almost right at Stimulus6, boundary for Continuum3 locates in between Stimulus6 and Stimulus7.

3 Summary and Discussion

The current experiments showed how perceptual pattern varied among three types of /ra/-/la/ continuum which had the same range of formant transition, but had different formant frequency values. The difference in formant values meant to reflect vocal tract size variation of speakers. Results of the identification test as well as those of the discrimination test showed similar functions across continua; however, those for Continuum3 looked a bit dissimilar from the rest. Because stimuli series of the current experiment had different F0 values in addition to spectral variation, there is still a possibility that the slight difference obtained in Continuum3 comes from F0 variation. Therefore, further experiments should reveal whether the difference comes from formant variation, which reflect vocal tract size variation, or not in order to find what factor would make speech easier to hear.

References

[1] Kent and Read, Onsei no onkyo bunseki. (Arai and Sugahara, Trans.), 2004. (Original work published 1992).

[2] Dalston, JASA, 57 (2), 462-469, 1975.

[3] Klatt, JASA, 67 (3), 971-995, 1980.

[4] Klatt and Klatt, JASA, 87 (2), 820-857, 1990.

[5] MacKain *et al.*, Appl. Pshycholing., 2 (4), 369-390, 1981.

[6] Polka and Strange, JASA, 78 (4), 1187-1197, 1985.

[7] Miyawaki *et al.*, Percept. Psychopys., 18 (5), 331-340, 1975.

[8] Mochizuki, J. Phon., 9, 283-303, 1981.

[9] Strange and Dittman, Percept. Psychophys., 36 (2), 131-145.

[10] Logan *et al.*, JASA, 89 (2), 874-886, 1991.

[11] Lively *et al.*, JASA, 96 (4), 2076-2087.

[12] Zue *et al.*, Speech Comm., 9 (4), 351-356, 1990.

[13] Boersma & Weenink, Glot International, 5 (9), 341-345.

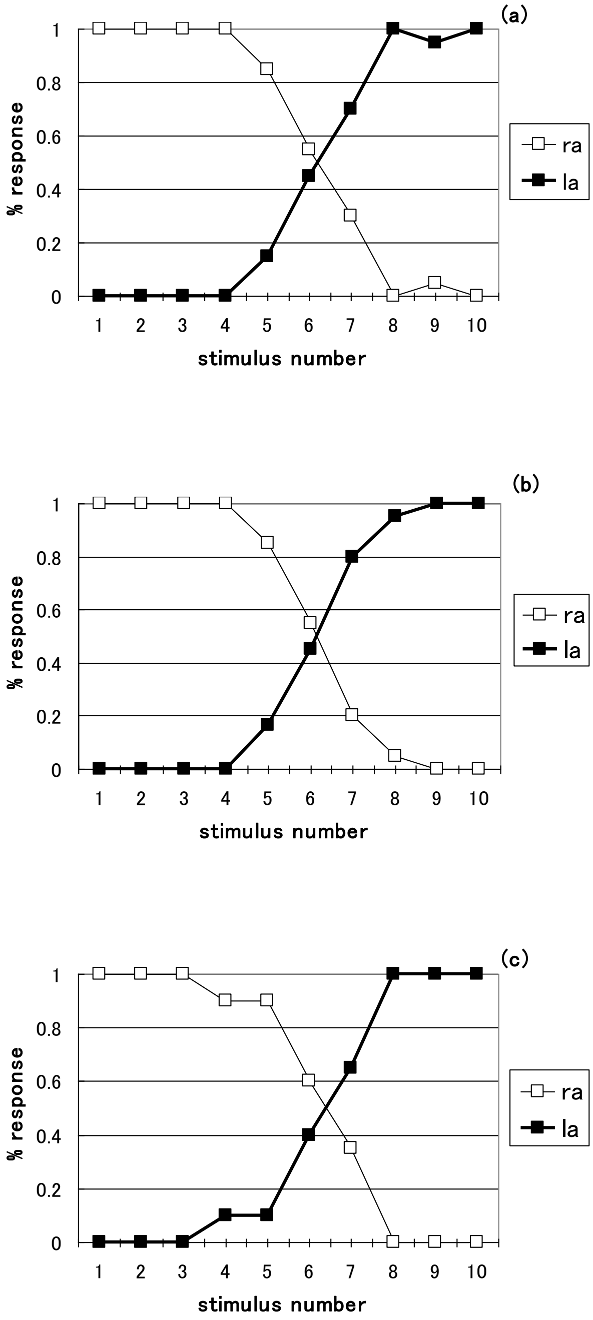

Fig. 4 Obtained identification function for Continuum1 (a), Continuum2 (b), and Continuum3 (c).