Acoust. Sci. & Tech. **35**, 5 (2014)

PAPER

Discrimination of /ra/–/la/ speech continuum by native speakers of English under nonisolated conditions

Kanako Tomaru* and Takayuki Arai

Faculty of Science and Technology, Sophia University,
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 6 August 2013, Accepted for publication 4 April 2014*)

Abstract: The discrimination of an /r/–/l/ speech continuum presented in an isolated monosyllabic context has been studied by a number of researchers. However, when one considers normal listening environments, it is unusual for people to hear syllables in isolation. In the present research, we investigated whether the characteristics of the discrimination performance for a /ra/–/la/ continuum presented under the isolated condition are also observed when the continuum is presented under nonisolated conditions, or more ordinary listening conditions. Two nonisolated conditions were employed: 1) the continuum was presented within a sentence, and 2) the continuum was preceded and followed by pure tones. Experiments revealed that the discrimination performance under condition 2) was similar to that under the isolated condition; however, the performance under condition 1) was different from those under the other conditions. The research suggests that the characteristics of the discrimination performance under nonisolated conditions are not necessarily identical to those under the isolated condition.

Keywords: Speech perception, Discrimination, /ra/–/la/ continuum, Nonisolated condition

PACS number: 43.71.–k, 43.71.Es, 43.71.Sy [doi:10.1250/ast.35.251]

1. INTRODUCTION

Phonetic categories can be distinguished by variations in acoustic parameters, e.g., formants, voice-onset time, and rise time. English /r/ and /l/ in onset positions, for example, are distinguished by the characteristics in the transition of the third formant (F3) (see [1] for a review). The F3 transition of /r/ shows an upward movement from low frequency to high frequency. On the other hand, /l/ has a slightly downward movement, or an almost straight trajectory.

According to previous studies, a listener's perception suddenly changes from /ra/ to /la/ (or /la/ to /ra/) at a certain point in a synthesized continuum, where only F3 transitions are varied from a /ra/ configuration to a /la/ configuration in equal steps [1–5]. Such a point is called the *categorical boundary* in the present study. When a listener attempts to discriminate between one syllable and another along a /ra/–/la/ continuum on the basis of only the variation in F3, it is considered that the listener's discrimination performance becomes most accurate for the syllable pairs that cross the boundary. For the other pairs, it is low, or almost at chance level. Many researchers have demonstrated such characteristics of the discrimination function for an /r/–/l/ continuum in line with *categorical perception* [2–5].

In previous research, listeners' discrimination performance was investigated using monosyllables in isolation. However, under ordinary listening conditions, it is rare that people hear a syllable under a completely isolated condition without any sounds around it. Thus, the question behind the current research is whether the characteristics of the discrimination performance obtained for continuous syllables *in isolation* are also observed when the syllables are presented in a more ordinary listening environment, or under *nonisolated* conditions. Under the nonisolated conditions here, the target /ra/–/la/ continuum is preceded and followed by sounds.

Two nonisolated conditions were prepared for the current research: 1) a sequence of speech sounds precedes and follows the target /ra/–/la/ syllables (Speech Condition), and 2) 1 kHz pure tones precede and follow the same target syllables (Nonspeech Condition). In Experiment 1, we first examined English native speakers' discrimination performance of a /ra/–/la/ continuum presented in isolation (Isolated Condition). In Experi-

*e-mail: ka-tomaru@sophia.ac.jp

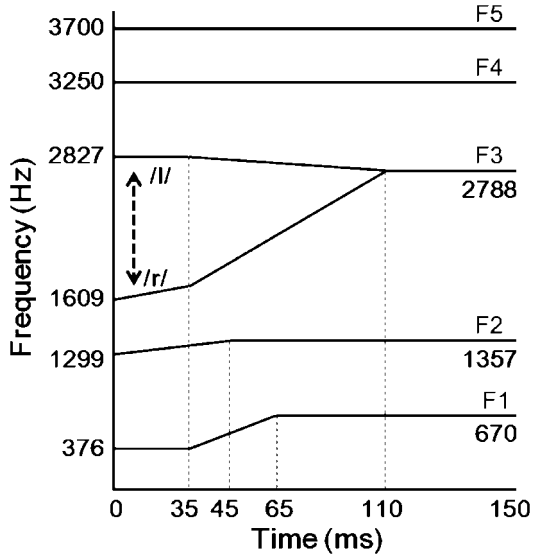

Fig. 1 Schematic representation of formant trajectories of the synthesized continuum.

ment 2, we investigated the discrimination of the same continuum presented under nonisolated conditions.

2. MATERIALS

2.1. Synthesizing /ra/–/la/ Continuum

For the present research, the English /ra/–/la/ continuum with a ten-step variation in the F3 transition was synthesized.

The ten-step continuum was synthesized using the XKL [6], originally designed by Klatt and Klatt [7]. The continuum was created on the basis of a male speaker's utterance from the TIMIT corpus (speaker ID: MKAM0) [8]. First, we decided parameter values for the first three formants, i.e., F1, F2, and F3, for the steady states of /a/ in the /ra/–/la/ continuum. For the steady-state values, the frequency values of F1, F2, and F3 were obtained from the vowel [ʌ] in "pro*nun*ciation" in the following sentence uttered by the speaker: "Clear pronunciation is appreciated" (sentence ID: sx236). The formant frequencies were obtained from this particular token because the same sentence spoken by the same speaker was used for the Speech Condition (see the following section for details). For F1, F2, and F3, frequencies were measured at several equidistant points in the selected part of the vowel, and the values at these points were averaged. The obtained values of F1, F2, and F3 were 670 Hz, 1,357 Hz, and 2,788 Hz, respectively (Fig. 1).

Figure 1 indicates a schematic representation of the formant trajectories in the synthesized stimuli. In the present research, only the F3 transition was varied. Thus, the transitions of F1 and F2 were fixed throughout the continuum. The F3 transition was decided using the data provided by MacKain *et al.* [5] (see also [9,10]). The starting frequency of F3 varied from 1,609 Hz (/ra/ configuration) to 2,827 Hz (/la/ configuration) in ten nearly equal steps (Step 1 through Step 10). In addition, the values at the inflection (Fig. 1) were also varied in ten steps from 1,717 Hz to 2,827 Hz. These values were again based on the data from MacKain *et al.* [5] (see also [9,10]). During the transition period, F3 of each step at 0 ms rose linearly to the value of the corresponding step at the inflection; then, the value at the inflection increased or decreased linearly to the steady state.

The transitions of F1 and F2 were fixed referring to a preliminary experiment [9,10] (see Appendix). Because characteristic changes in F1 and F2 also contribute to the perception of the /r/–/l/ contrast to some extent [4], fixed transitions were set to be as "neutral" as possible. That is, it was preferred that the F1 and F2 transitions do not cue /ra/ or /la/ in particular, so that the F3 transition is the only cue for the contrast. The transitional cue in F1 is temporal, whereas that in F2 is spectral [4]. On the basis of the findings of the preliminary experiment, the starting frequency of F1 at 0 ms was set to 376 Hz and its transition was decided to start at 35 ms. During the transition period, the value of F1 rose linearly from 376 Hz to 670 Hz. For the F2 transition, it was decided that F2 would start from 1,299 Hz at 0 ms, and the value increased linearly to the steady-state value.

The values of the fourth and fifth formants, i.e., F4 and F5, were set to the default values of the synthesizer (3,250 Hz and 3,700 Hz, respectively). These values were fixed throughout the continuum without any transitions. In addition, F0 was extracted from the word "pronunciation" in the original sentence and reflected in each of the synthesized syllables. First, the original F0 was sampled at 25 equidistant time points. Next, the number of F0 values was reduced to five by averaging every five values. The obtained five averaged values were used as F0 at the first five of six equidistant time points that were 70 ms away from each other in a synthesized syllable. The value at the sixth point was the same as that at the fifth point. The original F0 was approximated to minimize the unnaturalness when a syllable was inserted into a sentence in Experiment 2 (see Sect. 2.2 for details).

Each of the synthesized syllables had 100 ms rising and falling periods of amplitude. The rising and falling periods are omitted from Fig. 1. The amplitude in the rising period was changed linearly from 0 dB at 0 ms to 60 dB at 100 ms by adjusting the parameter, "amplitude of voicing (AV)" of the synthesizer. Using the same parameter, the amplitude in the falling period was also changed linearly from 60 dB to 0 dB in 100 ms. During the rising and falling periods, stimuli had formant values at 0 ms and at 150 ms, respectively (Fig. 1). The total duration of each of the continuous syllables was 350 ms.

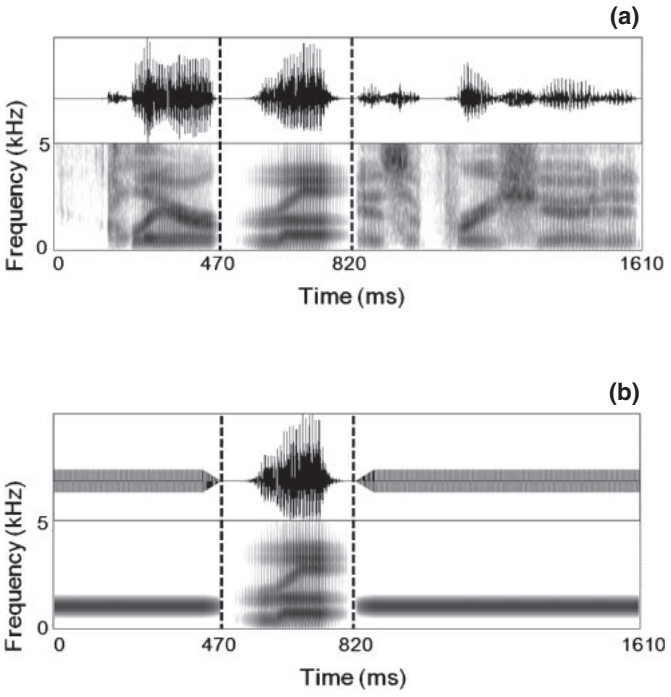

Fig. 2 Syllable Step 1 presented under (a) Speech Condition (*S*-Step 1) and (b) Nonspeech Condition (*NS*-Step 1).

Digital outputs from the synthesizer (16-bit resolution and 10 kHz sampling rate) were converted to 16-bit resolution and a 16 kHz sampling rate.

2.2. Syllables under Nonisolated Conditions

Two nonisolated conditions were adopted: 1) Speech Condition and 2) Nonspeech Condition.

A sequence of speech sounds for the Speech Condition was selected from the sentence produced by MKAM0: "Clear pronunciation is appreciated." For the experiment, the word "pronunciation" was deleted from the original sentence, i.e., "Clear _ is appreciated," and the /ra/–/la/ syllables were embedded into the blank portion of the sentence, e.g., "Clear /ra/ is appreciated" (Fig. 2(a)). To avoid co-articulatory effects, we tapered the offset of the word "clear" until a bilingual speaker of English and Japanese could not detect the following /p/ for the word "pronunciation" in the original sentence. Each syllable in the ten-step continuum was presented under the Speech Condition (see Experiment 2 for details). Syllable steps presented under the Speech Condition were called *S*-Step 1, *S*-Step 2, ..., *S*-Step 10.

For the Nonspeech Condition, a 1 kHz pure tone was employed to precede and follow the continuous syllables. As illustrated in Fig. 2(b), pure tones virtually replaced the sequence of speech sounds under the Speech Condition. Thus, the lengths of the pure tones preceding and following a syllable matched the length of the word "Clear" and that of the phrase "is appreciated" under the Speech Condition, respectively. The pure tones took 50 ms to fall before the onset of an embedded syllable and 50 ms to rise after the offset of the syllable. Syllable steps under the Nonspeech Condition were called *NS*-Step 1, *NS*-Step 2, ..., *NS*-Step 10.

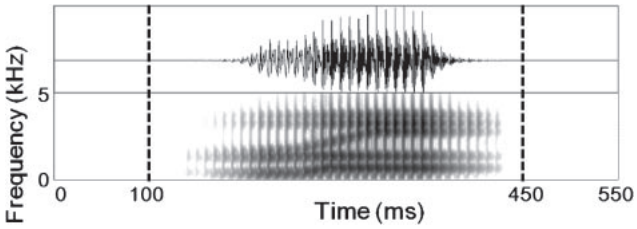

Fig. 3 Syllable Step 1 under Isolated Condition.

3. EXPERIMENT 1

Experiment 1 was conducted to examine listeners' discrimination performance for the continuum under the Isolated Condition.

3.1. Stimuli

The /ra/–/la/ continuum synthesized in Sect. 2.1 were presented in isolation. Under this condition, each syllable was preceded and followed by a 100 ms silent period (Fig. 3).

3.2. Listeners

Two groups of native speakers of English were recruited for the experimental tasks: one group was recruited for an identification task, and the other was recruited for a discrimination task.

One group (Group 1), which consisted of two listeners (1 male, 1 female) with normal hearing, participated in the identification task. The listeners were both from the United States. Their ages were 21 and 22 years (mean: 21.5 years). Both of them had been residing in Japan for 3 months at the time of the experiment.

The other group (Group 2) consisted of nine native speakers of English with normal hearing (6 males, 3 females). They participated in an AXB discrimination task. Eight of them were from the United States, and one of them was from the United Kingdom. Their ages ranged from 20 to 21 years (mean: 20.7 years). They had been residing in Japan for 3 to 11 months at the time of the experiment (mean: 4.1 months).

3.3. Procedure

Stimuli were presented using the Praat software [11]. All stimuli were presented diotically via Sennheiser HDA 200 headphones at a comfortable listening level for the participants.

3.3.1. Identification

The identification task took the form of a two-alternative forced choice (2AFC). Thus, in the identification task, listeners in Group 1 were asked to choose if the presented syllables were "ra" or "la." The listeners heard ten repetitions of the ten synthesized syllables (10 stimuli × 10 repetitions = 100 judgments). Stimuli were presented once, and listeners could not replay the stimuli. Stimuli were presented randomly to the listeners. A short practice task was assigned to the listeners prior to the experimental task. No feedback was given during the practice or the main experiment.

3.3.2. Discrimination

In the AXB discrimination task, synthesized syllables were paired such that each pair (AB) differed by two steps in the continuum, i.e., Step 1–Step 3, Step 2–Step 4, ..., Step 8–Step 10. Listeners in Group 2 were instructed to judge if the second syllable (X) matched the first (A), or the third (B). Listeners gave their responses by clicking on buttons labelled "first" and "third" on the screen. Paired stimuli were arranged into four permutations (AAB, ABB, BAA, and BBA). There were three repetitions of each presentation. Thus, listeners made 12 judgments for each pair. The total number of judgments was 96 (8 pairs × 4 presentations × 3 repetitions = 96 judgments). The AXB presentations were random. The inter-stimulus durations between A and X, and also between X and B, were 300 ms. The 300 ms interval was also used for AXB tasks under nonisolated conditions (Experiment 2). It has been shown that the discrimination performance is most accurate when the interval is set to around 300 ms [12]. Thus, the 300 ms interval was considered to be reasonable for discrimination tasks that require the use of working memory, such as the tasks in the current research. In Sect. 5 we discuss the effectiveness of the 300 ms interval in the current experiments.

For the purpose of familiarizing the listeners with the task, a short practice session was performed before the main experimental task.

3.4. Results

3.4.1. Identification

The averaged responses of /ra/ and /la/ over the two listeners are shown in Fig. 4. The figure shows an S-shaped curve. The number of /ra/ responses dropped from 75% at Step 4 to 55% at Step 5. From Step 6, the number of /ra/ responses decreased and the number of /la/ responses increased. Thus, the continuum was perceptually divided into two categories, and the boundary seemed to be at around Step 5.

3.4.2. Discrimination

The percentage of correct responses was averaged over the nine listeners. Figure 5 shows the averaged correct rate at each step. As indicated in the figure, there was a peak in the discrimination performance for the pair 4–6.

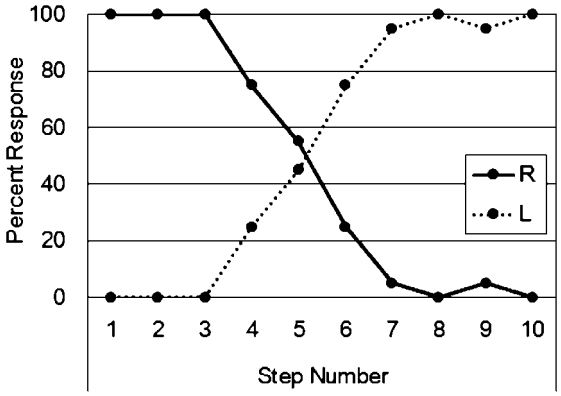

Fig. 4 Average percent responses (%) of /ra/ (solid line) and /la/ (dotted line) in the identification task.

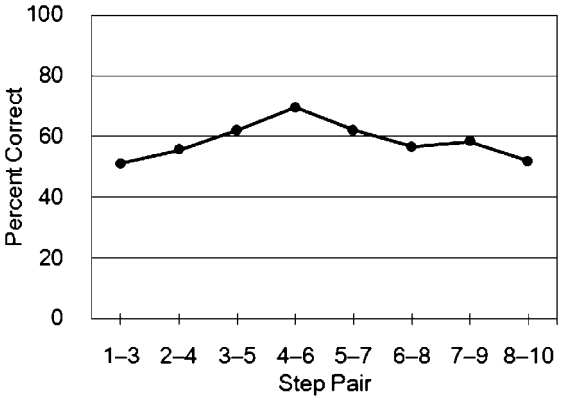

Fig. 5 Average percent correct (%) for discrimination of the target syllables presented in isolation.

In order to assess whether the discrimination performance for the pair 4–6 was significantly better than those for the other pairs, we compared the correct rate of the pair 4–6 and those of the other pairs. For statistical analysis, we grouped the stimulus pair(s) as 1) the central pair (the pair 4–6), 2) the /ra/-side of the polar pairs (the pairs 1–3, 2–4, and 3–5), and 3) the /la/-side of the polar pairs (the pairs 5–7, 6–8, 7–9, and 8–10). The ANOVA with repeated measures revealed the main effect of the group ($F(2, 16) = 7.57$, $p = 0.005$). A post hoc multiple comparison with the Bonferroni correction revealed that the difference between the central pair and the /ra/-side of the polar pairs was significant ($p = 0.009$). The difference between the central pair and the /la/-side of the polar pairs was also significant ($p = 0.028$). Thus, the discrimination performance was significantly better for the pair 4–6.

3.5. Discussion of Experiment 1

It is important to note that the pair 4–6 was assumed to be the pair crossing the categorical boundary, i.e., Step 5

(Fig. 4). That is, the discrimination performance was best when each syllable in the pair belonged to different phonetic categories. Such characteristics of the discrimination performance are consistent with the previous findings (for example, [3–5]), showing a typical pattern of the discrimination function as introduced in Sect. 1.

4. EXPERIMENT 2

In Experiment 2, we investigated the discrimination performance under two nonisolated conditions.

4.1. Stimuli

S-Step 1 through *S*-Step 10 and *NS*-Step 1 through *NS*-Step 10 were used as stimuli.

4.2. Listeners

The listeners were those in Group 1 and Group 2. However, one of the nine listeners in Group 2 was different from the one who participated in Experiment 1. The new participant was a 29-year-old female from the United States who had been residing in Japan for ten months at the time of the experiment.

For Group 1, Experiment 1 and Experiment 2 were held on the same day. For Group 2, Experiment 2 was carried out after Experiment 1 on a different day.

4.3. Procedure

The equipment used and basic procedure were identical to those of Experiment 1. Thus, only details that are specific to Experiment 2 are presented below.

4.3.1. Identification

Firstly, listeners identified the syllables under the Nonspeech Condition (*NS*-Step 1 through *NS*-Step 10). Under this condition, the listeners were asked to judge if a syllable appearing between pure tones was "ra" or "la." The listeners heard ten repetitions of the ten syllables under the Nonspeech Condition (10 syllables × 10 repetitions = 100 judgments).

Next, the listeners were presented with the stimuli of the Speech Condition (*S*-Step 1 through *S*-Step 10). In this task, the listeners were told that they would hear a syllable (***xx***) in the following form, once: *Clear **xx** is appreciated*. The listeners were then instructed to label the syllable in the sentence as "ra" or "la" by clicking the corresponding button on a screen. Ten repetitions of the ten syllables were presented to the listeners (10 syllables × 10 repetitions = 100 judgments). No feedback was given in either task.

4.3.2. Discrimination

Group 2 participated in a discrimination task under two nonisolated conditions. The first task was to discriminate the syllables under the Nonspeech Condition. In the instructions, listeners were told that they would hear a syllable between tone sounds (Fig. 2(b)), and that they would hear such a sequence three times. Next, they were instructed to judge whether the syllable in the second sequence (X) was more similar to that in the first sequence (A) or to that in the third sequence (B). There were three repetitions for each presentation, making a total of 96 judgments (8 pairs × 4 presentations × 3 repetitions = 96 judgments).

The second task was to discriminate syllables under the Speech Condition. Under this condition, an experimenter told the listeners that they would hear a sentence containing a syllable (***xx***) in the middle, i.e., *Clear **xx** is appreciated*, three times (Fig. 2(a)). Then, they were requested to judge whether the syllable in the second sentence (X) sounded more similar to that in the first sentence (A) or to that in the third sentence (B). As in Experiment 1, the interstimulus durations were 300 ms. The length of the intervals was set to be identical to that in the preceding experiment because 300 ms was considered to be reasonable for discrimination tasks. Other experimental conditions were the same as those in the discrimination task under the Nonspeech Condition, including the total number of judgments.

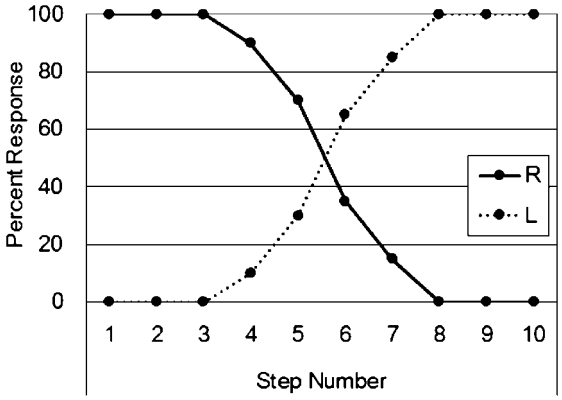

Fig. 6 Average percent response of /ra/ (solid line) and /la/ (dotted line) under Nonspeech Condition.

4.4. Results

4.4.1. Identification

Responses of /ra/ and those of /la/ were averaged over the two listeners for the Nonspeech Condition (Fig. 6) and Speech Condition (Fig. 7). For the Nonspeech Condition, the /ra/ responses dropped from 70% at Step 5 to 35% at Step 6. In other words, in Step 6 and subsequent steps /la/ was heard rather than /ra/ by listeners. Thus, it is suggested that the categorical boundary is located between Step 5 and Step 6 under the Nonspeech Condition. Also, under the Speech Condition, the /ra/ responses dropped from 65% at Step 5 to 30% at Step 6. Thus, in Step 6 through Step 10, /la/ was also heard under this condition. Therefore, a categorical boundary was assumed to exist between Step 5 and Step 6 also under this condition.

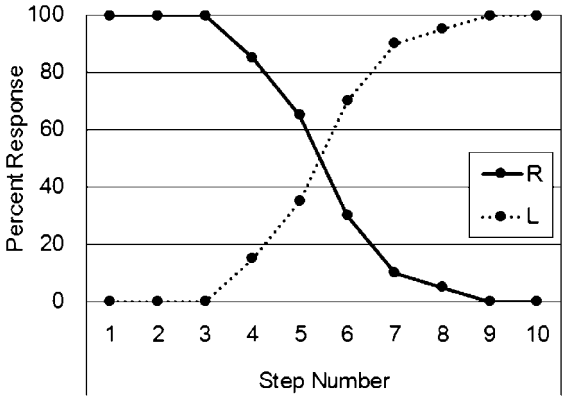

Fig. 7 Average percent responses of /ra/ (solid line) and /la/ (dotted line) under Speech Condition.

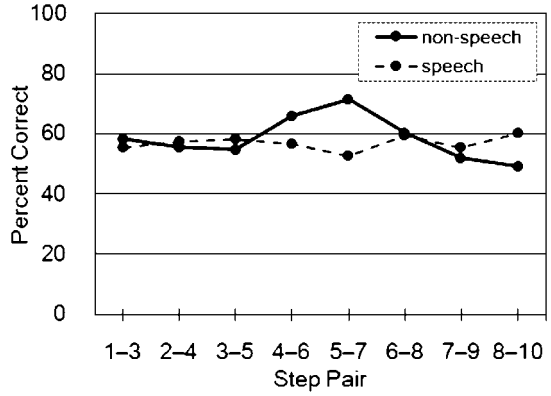

Fig. 8 Average percent correct (%) of discrimination under Nonspeech Condition (solid line) and Speech Condition (dashed line).

4.4.2. Discrimination

Figure 8 shows the averaged correct rates of syllable discrimination under the Nonspeech Condition (solid line) and the Speech Condition (dashed line). Firstly, when the continuum was presented under the Nonspeech Condition, listeners' discrimination of the central pairs, i.e., the pair 4–6 and the pair 5–7, was more accurate than that for the polar pairs. These central pairs were the pairs that crossed the categorical boundary, which was indicated to be located between Step 5 and Step 6 from the identification results (Fig. 6). Thus, in order to assess the discrimination performance of the pair crossing the boundary, the stimulus pairs were grouped as follows: 1) the central pairs (the pairs 4–6 and 5–7), 2) the /ra/-side of the polar pairs (the pairs 1–3, 2–4, and 3–5), and 3) the /la/-side of the polar pairs (the pairs 6–8, 7–9, and 8–10). The ANOVA with repeated measures revealed the main effect of the group ($F(2, 16) = 6.95$, $p = 0.007$). In addition, a post hoc multiple comparison with the Bonferroni correction showed that the percent correct for the central pairs was significantly higher than that for the /ra/-side of the polar pairs ($p = 0.038$). Additionally, it was also higher than that for the /la/-side of the polar pairs ($p = 0.049$).

When the same syllables were perceived under the Speech Condition, on the other hand, the discrimination performance was constantly low, indicating no peaks of performance. Similarly to under the Nonspeech Condition, the categorical boundary under the Speech Condition was shown to be located between Step 5 and Step 6 (Fig. 7). Thus, for the statistical analysis, the stimulus pairs were divided into three groups in the same manner as under the Nonspeech Condition. Results of the ANOVA with repeated measures indicated that the percent correct was not significantly different among the three groups ($F(2, 16) = 0.215$, $p = 0.809$).

4.5. Discussion of Experiment 2

The discrimination function under the Nonspeech Condition appears to be similar to the discrimination function under the Isolated Condition (Experiment 1). Under the Isolated Condition, listeners' judgments were most accurate for the central pair, i.e., the pair 4–6 (Fig. 5), which crossed the categorical boundary (Fig. 4). Thus, the discrimination function under the Nonspeech Condition and that under the Isolated Condition were similar to each other in the sense that the percent correct increased for the pairs crossing the categorical boundary.

On the other hand, under the Speech Condition, the discrimination performance was consistently low, indicating no peaks for any stimulus pairs. This suggests that a listener's performance of syllable discrimination under the Isolated Condition is not retained when the syllables are presented in a sentence.

5. DISCUSSION

In the present research, we investigated whether discrimination functions of a /ra/–/la/ continuum presented under the nonisolated conditions had the same characteristics as the function for the same continuum presented in isolation. Through two perceptual experiments, we found that the discrimination functions differed depending on the condition. Concretely, the discrimination function peaked, as under the Isolated Condition, when a target syllable was surrounded by nonspeech sounds (Nonspeech Condition). On the other hand, the function was rather flat and showed no specific peak of discrimination accuracy when the syllable was inserted into a sentence (Speech Condition).

In this section, we mainly discuss the reason why we observed no peak under the Speech Condition. As demonstrated by the results of the current experiments, the identification results in the case of the Speech Condition did not differ greatly from those in the case of the Isolated Condition or the Nonspeech Condition. The only difference

was the appearance of a discrimination peak. According to the previous studies, especially those concerned with categorical perception, a discrimination function is thought to be predicted by an identification function on the basis of the following assumptions: 1) stimuli are perceived by listeners in accordance with the category labeling, and 2) listeners use the same labeling in both identification and discrimination tasks [2–5]. These assumptions imply that listeners discriminate stimuli after labeling them. Furthermore, it is also implied that listeners are able to discriminate stimuli only when they assign different labels to the stimuli. The ability to discriminate stimuli that have different labels appears as a peak in a discrimination function. Thus, a discrimination function is predicted to have a performance peak for the stimulus pair that crosses the categorical boundary, as indicated by the identification results. In the case of the Nonspeech Condition, as well as the Isolated Condition, the discrimination function indicated a peak at the central pair(s) that crossed the boundary, as indicated by the results of the identification task. This suggests that, although the participant groups were different in each task, listeners used category labeling during discrimination tasks under these conditions. Under the Speech Condition, however, the discrimination function did not peak at the cross-boundary stimulus pairs. This suggests that under the Speech Condition, listeners were able to *label* the stimuli but they were not able to *use* the labels in the discrimination task.

According to the results of previous studies, it is not rare to obtain no peak for a discrimination function when listeners are not able to label the stimuli. For example, a discrimination function does not peak when listeners try to discriminate nonspeech stimuli that do not have known labels ([3,13], among others). In such cases, stimuli are not even labeled; thus, it is considered that these stimuli are perceived *noncategorically*. However, since the synthesized syllables in the current study were labeled in the identification task, the function without a peak under the Speech Condition should not be interpreted as evidence of noncategorical perception. Rather, it is more likely that one of the characteristics of categorical perception may be hidden under this condition.

The difference in the discrimination results can be explained by automatic speech processing (ATP). Although details of this processing are still unknown, many scholars believe that speech is processed automatically so that listeners cannot refrain from comprehending words and sentences spoken in their native language ([13,14], among others). ATP may include the observation of acoustic characteristics and the comprehension of meaning. When ATP is triggered, it becomes difficult for listeners to ignore incoming speech. Thus, under the Speech Condition, listeners may involuntarily start analyzing not only the target syllables that they must focus on, but also the carrier sentences that they do not have to attend to.

If ATP is triggered by the reception of speech sounds, the perceptual burden in an AXB discrimination task under the Speech Condition must become particularly high. Under the Speech Condition, on the basis of the assumption of ATP, the word “clear” in a stimulus sentence, i.e., “Clear /ra/–/la/ is appreciated” (see also Fig. 2(a)) must be the first sound to be processed. The analysis of the target /ra/–/la/ syllable subsequently follows. Finally, the analysis progresses to “is” and “appreciated,” which constitute the rest of the sentence. Note here that such a process of speech analysis is behind the actual presentation of speech stimuli. Therefore, listeners must analyze speech perceived a short time ago *and* receive new speech at the same time (analysis overlap). The problem of analysis overlap in a discrimination task under the Speech Condition is that it takes up a listener’s working memory during the whole of the AXB presentation. In the AXB discrimination task, three stimulus sentences were presented with relatively short intervals, i.e., 300 ms. The 300 ms stimulus interval may be appropriate under conditions where ATP is not triggered by the surrounding context, i.e., the Isolated Condition and Nonspeech Condition. However, it may not work efficiently when ATP is involved, i.e., the Speech Condition. That is, if the interval is too short under conditions that trigger ATP, listeners may encounter a new sentence before completing the analysis of a target syllable in the preceding sentence. As a result, listeners may become occupied with ongoing speech processing during the AXB task so that they do not have sufficient capacity for discrimination based on assigned labels. In this sense, the perceptual burden under the Speech Condition must have been higher than that under the the other conditions. Therefore, it is possible that the high perceptual burden caused by ATP influenced the use of labels in discrimination under the Speech Condition.

In contrast, ATP may have not affected the results of the identification task under the Speech Condition because there was no following sentence corresponding to X or B, unlike in the discrimination task. In other words, listeners were able to make judgments after completing the analysis of a presented sentence in the identification task.

If analysis overlap is the main cause of the results obtained under the Speech Condition, it is suggested that listeners may be able to compare the labels of target syllables if the analysis of speech in the preceding presentation does not overlap the reception of speech in the following presentation. In the present study, the stimulus interval was 300 ms, which may have been too short. In future experiments, it will be necessary to clarify whether or not a discrimination peak is obtained under the Speech Condition when the interval is sufficiently long.

In addition, in the present study, the target syllables were nonsense monosyllables. Thus, lexical knowledge could not help a listener remember the target syllables. Instead, the listener may have had to rely solely on working memory for discrimination, but this may not have been possible with their working memories already being fully used for the ongoing speech processing. It may be easier for listeners to retain the labels of syllables if the syllables have some meaning so that they can be stored in the long-term memory as words. Thus, the discrimination performance should be further tested using real words.

Furthermore, future research should be aimed at elucidating how the existence of speech-related features, e.g., frequency components, the structure of harmonics, and amplitude modulation, affects the discrimination performance under the nonisolated condition. It is assumed that the comparison of target syllables under the Speech Condition was disrupted by the carrier sentence because ATP was triggered. Thus, listeners are predicted to be able to use the labels if the sounds preceding the target do not trigger speech processing. In future experiments, the discrimination performance will be further tested under a nonisolated condition that contains less speech-related information than natural speech, but more information than a sinusoidal sound. The continuation of research toward this direction is expected to reveal the key features that initiate ATP and eventually the difference between speech and nonspeech.

Other questions to be answered in future research are as follows. Firstly, it is necessary to test the discrimination performance of stimuli that are more steps apart, e.g., 3 steps or even 4 steps, under the Speech Condition. The further apart they are from each other, the greater the acoustical difference between stimuli. Thus, it is necessary to reveal whether listeners are able to discriminate between paired stimuli that have a greater difference in terms of acoustics. The investigation of this point would clarify whether or not an acoustic difference is detectable by listeners during speech processing. In addition, we should discuss whether listeners are able to use labels when paired stimuli have a greater acoustic difference. Secondly, although a /ra/–/la/ continuum was employed as stimuli in the current experiments, using other parameters in other phonetic categories, e.g., rise time and the voice-onset time, is also an option. These are points that should be taken into account to further reveal the effects of *being nonisolated* on the discrimination of syllables.

ACKNOWLEDGMENTS

The contents of this paper were partially reported in the Auditory Research Meeting sponsored by the Technical Committee of Psychological and Physiological Acoustics [Tomaru and Arai, "Effects of surrounding contexts on English /ra/–/la/ perception: for educational purposes," *Proc. Audit. Res. Meet.*, pp. 591–596 (2012)] and in the International Congress of Acoustics [Tomaru and Arai, "Perception of /ra/–/la/ contrast in different contexts: mono-syllable vs. sentence," *Proc. Meetings on Acoust.*, Vol. 19, 060289 (2013)].

REFERENCES

[1] R. D. Kent and C. Read, *The Acoustic Analysis of Speech* (Singular Publishing Group, San Diego, 1992).
[2] A. M. Liberman, K. S. Harris, J. A. Kinney and H. Lane, "The discrimination of relative onset-time of the components of certain speech and nonspeech patterns," *J. Exp. Psychol.*, **61**, 379–388 (1961).
[3] K. Miyawaki, W. Strange, R. Verbrugge, A. M. Liberman, J. J. Jenkins and O. Fujimura, "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," *Percept. Psychophys.*, **18**, 331–340 (1975).
[4] L. Polka and W. Strange, "Perceptual equivalence of acoustic cues that differentiate /r/ and /l/," *J. Acoust. Soc. Am.*, **78**, 1187–1197 (1985).
[5] K. S. MacKain, C. T. Best and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Psycholinguist.*, **2**, 369–390 (1981).
[6] D. H. Klatt, "The new MIT speech VAX computer facility," in *Speech Communication Group Working Papers IV, Research Laboratory of Electronics* (MIT, Cambridge, 1984), pp. 73–82.
[7] D. H. 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).
[8] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, **9**, 351–356 (1990).
[9] K. Tomaru and T. Arai, "English speakers' perception of synthesized /ra/–/la/ continua with the same range of formant transition starting from different formant values," *Proc. Autumn. Meet. Acoust. Soc. Jpn.*, pp. 445–448 (2012).
[10] K. Tomaru and T. Arai, "Perception of multiple series of English /ra/–/la/ continuum having different end frequencies of formant transitions," *Acoust. Sci. & Tech.*, **35**, 166–169 (2014).
[11] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]," ver. 5.3.23, retrieved 7 August 2012 from http://www.praat.org/.
[12] D. B. Pisoni, "Auditory and phonetic memory codes in the discrimination of consonants and vowels," *Percept. Psychophys.*, **13**, 235–260 (1973).
[13] K. Johnson and J. V. Ralston, "Automaticity in speech perception: Some speech/nonspeech comparisons," *Phonetica*, **51**, 195–209 (1994).
[14] P. A. Tun, G. O'Kane and A. Wingfield, "Distraction by competing speech in young and older adult listeners," *Psychol. Aging*, **3**, 453–467 (2002).

Appendix

This appendix gives the details of the preliminary experiment, which are a subset of the report of Tomaru and Arai [9,10].

Although the F3 transition was assumed to be the primary cue for the /ra/–/la/ contrast, the transitions of F1 and F2 could also cue /ra/ or /la/ [4]. Thus, the F1 and F2

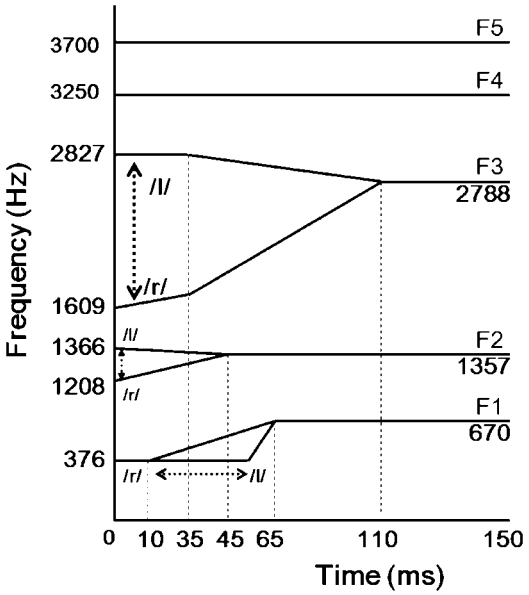

Fig. A.1 Schematic representation of formant trajectories of the synthesized continuum for the preliminary experiment.

transitions of the continuum for the main experiment were preferred to be as "neutral" as possible so that the F3 transition would be the only cue for the /ra/–/la/ contrast. In the preliminary experiment, "neutral transitions" were interpreted to be the transitional characteristics at a particular step in a continuum where listeners' identification judgments become close to the chance level. To find the neutral transitions of F1 and F2, a /ra/–/la/ continuum with simultaneous variations in F1, F2, and F3 was generated. The method of continuum synthesis was identical to that introduced in Sect. 2, except that the F1 and F2 transitions were also changed along the continuum (Fig. A.1). Former research [4] revealed that the cue in F1 is temporal whereas the cue in F2 is spectral (Fig. A.1). Thus, the continuum included variations in the temporal characteristics of F1 and the spectral characteristics of F2. For F1, the starting frequency value was set to 376 Hz on the basis of the data provided by MacKain *et al.* [5]. For the temporal variation of F1, the time point where the transition started was varied from 10 ms to 55 ms in ten equal steps following Polka and Strange [4]. For F2, the starting frequency was varied from 1,208 Hz to 1,366 Hz in ten steps based on MacKain *et al.* [5]. For both F1 and F2, the frequency at the onset of the transition and the frequency of the steady state were interpolated linearly during the transition period. The transitional characteristics of F3 were identical to those introduced in Sect. 2.

Using the synthesized stimuli introduced above, we conducted an identification task with five native speakers of English. The results of the experiment revealed that the listeners' judgments became almost the chance level at Step 6: Step 6 received 55% /ra/ responses and 45% /la/ responses. Thus, the transitional characteristics of F1 and F2 at Step 6 were judged to be neutral. The neutral starting point of the F1 transition was at 35 ms; the neutral starting frequency of F2 at 0 ms was 1,299 Hz.

Kanako Tomaru received her B.A. and M.A. degrees in linguistics from Doshisha Univ., Kyoto in 2009 and Sophia Univ., Tokyo in 2011, respectively. She is currently working on her PhD at the Faculty of Science and Technology, Sophia Univ. Her research interests include language perception in relation to the study of Phonetics and Phonology, Acoustics, and Psycholinguistics.

Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Professor of the Department of Information and Communication Sciences, Sophia Univ. In 2003–2004, he is a visiting scientist at Massachusetts Institute of Technology (Cambridge, MA, USA). His research interests include signal processing, acoustics, speech and hearing sciences, spoken language processing, and acoustic phonetics.