Role of labeling mediation in speech perception: Evidence from a voiced stop continuum perceived in different surrounding sound contexts

Kanako Tomaru* and Takayuki Arai

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

(Received 25 November 2015, Accepted for publication 23 June 2016)

Abstract: The theory of categorical perception of speech sounds traditionally suggests that speech sound discrimination is conducted based on phonemic labeling, which is an abstract speech representation that listeners are hypothesized to have. However, recent research has found that the impact of labeling on perception of an English /t/–/l/ contrast may depend on surrounding sound contexts: the effects of phonemic labeling may disappear when the speech sounds to be discriminated are presented in a sentence. The purpose of the present research is to investigate (1) the effects of the sound contexts on categorical perception of speech sounds, and (2) cross linguistic extensibility of such an effect. The experiments employed a Japanese voiced stop consonant continuum, i.e., /ba/~da/, and tested discrimination of sounds on the continuum by native speakers of Japanese. Experiment 2 in particular investigated whether sounds on such a continuum are discriminated in accordance with the labeling when the sound in question is inserted into a sentence. Through experiments, the cross linguistic effects of surrounding sound contexts are found although there may be some exceptional cases. The research proposes reconsideration of the role of labeling mediation in speech perception.

Keywords: Speech, Categorical perception, Labeling mediation, Context effects

PACS number: 43.71.−k, 43.71.Es, 43.71.Hw, 43.71.Sy, 43.72.Ja [doi:10.1250/ast.37.303]

1. INTRODUCTION

The flipside of the importance of phonemic categories in speech perception is the negligibility of acoustic details. Although listeners may be aware of acoustic variation of allophones under some experimental conditions [1,2], traditional studies have often discussed human’s selective sensitivity to phonemic contrast during speech perception in terms of categorical perception [3–5].

1.1. Categorical Perception and Its Characteristics

The term categorical perception refers to a particular type of perception that indicates sensitivity to phonemic categories. The sensitivity is assessed using both identification and discrimination tasks. The former task shows how listeners label speech stimuli. For example, when presented with a sound continuum that gradually changes from English /t/ to /l/, a native speaker of English labels stimuli that are on the /t/-side from a phonetic perspective as “r” and those on the /l/-side as “l” with great frequency [5–7]. In addition, such a continuum contains stimuli that could be heard as either “r” or “l,” where the cumulative percentage of phonemic labeling responses becomes 50%. The point is called a categorical boundary [3–7]. The discrimination task, on the other hand, shows listeners’ sensitivity to acoustic difference of syllables along a continuum; the difference between adjacent syllables along the continuum is called a step. In the discrimination task, syllables are paired such that each pair differs by steps of arbitrary size of acoustic difference along the continuum. For example, in the case of a one-step comparison along a continuum that moves from /t/ to /l/ in ten steps, i.e., from step 1 (/t/) to step 10 (/l/), step 1 may be compared with step 2, step 2 with step 3, and so on. When it is a two-step comparison, step 1 will be paired with step 3, and so on. If stimuli are perceived categorically, listeners’ discrimination performance is most accurate when comparing stimuli that stride a categorical boundary, which is indicated by the identification results. For instance, if the categorical boundary of this hypothetical /t/–/l/ continuum is at step 5, the percentage of correct discrimination judgments is expected to be highest when listeners...
compare stimuli from step 1 through 4 to stimuli from step 6 through 9. High discrimination sensitivity at a categorical boundary is called a discrimination peak [8].

According to the original work by Liberman et al. [3], categorical perception expects that the location of a categorical boundary in a continuum agrees with a discrimination peak on a basis of a hypothesis that listeners are not able to distinguish two speech segments unless each segment has a different phonemic label. Concretely, categorical perception in an original sense predicts (1) discrimination peak at a pair that strides a categorical boundary and (2) close to the chance level of discrimination within the same phonemic category.

Traditionally, categorical perception had been considered to be observable under any circumstances of speech perception. However, later studies have revealed that the degree of categorical perception depends on methodology of experimental tasks [1], and surrounding sound contexts [9]. For example, Gerrits and Schouten [1] showed that perception tends to be biased by discrimination tasks. Their finding was that perception was fairly categorical when tested using a two-interval two-alternative forced-choice discrimination task (2I2AFC), but it was not when tested with a four-interval two-alternative forced-choice discrimination task (4I2AFC) (see Sect. 2 for details). The 2I2AFC demands discrimination judgments to be made with top-down labeling criteria leading to categorical perception, while 4I2AFC does not require listeners to judge the degree of difference based on such higher level language related information leading perception to be non-categorical, or auditory basis. Their suggestion is: that “[i]n normal everyday speech perception, we perceive [speech sounds] categorically,” but experimental tasks operate on perceptual process and perception can be either categorical or non-categorical depending on criteria the task demands.

However, more recent report has found that the effects of categorical perception diminish, if they do not disappear, even in everyday listening settings. Tomaru and Arai [9] investigated identification and discrimination performance of native speakers of English using synthesized English syllables that continuously change from /za/ to /la/ under three conditions. First was an Isolated Condition, in which each syllable in the continuum was presented in isolation. In the second condition, particularly referred to in the present study as a Sentence Condition, the same syllables were presented as part of a sentence to mimic ordinary listening condition: “Clear /za/ is appreciated.” Finally, in the third condition (a Nonspeech Condition), the syllables were sandwiched by pure tones. They found fairly sharp identification with a categorical boundary under all conditions; however, the discrimination peak was not observable under the Sentence Condition unlike other conditions. Since the degree of categorical perception must be assessed by a combination of identification and discrimination performances, as discussed below, perception under the Sentence Condition was not fully categorical against traditional expectations. This implies that ordinary speech perception is non-categorical, or at least less categorical than formerly believed.

1.2. Purpose of the Present Study

The main purpose of the present research is to replicate the former results of Tomaru and Arai [9] using a voiced stop continuum. The reason why we employed voiced stop continuum is because of their perceptual tendency toward categorical perception [10]. Healy and Repp [10] found this tendency through perceptual experiments using vowel, fricative and voiced stop, i.e., /ba/--/da/, continua as stimuli. When listeners were asked to label stimuli first then discriminate between them (overt labeling task), a discrimination peak was observable for each kind of continuum. However, when listeners were asked to discriminate the stimuli without labeling the stimuli beforehand (covert labeling task), only the voiced stop continuum indicated a discrimination peak. Therefore if the effect of being-in-a-sentence found by the former research is caused by deprivation of overt labeling mediation process, we should expect a discrimination peak for a /ba/--/da/ continuum presented within a sentence. If, on the other hand, we do not observe a peak even for the voiced stop continuum, it provides supportive evidence for the given hypothesis that surrounding speech sounds interfere categorical perception of any consonants. Additionally, the present research aims at exploring cross linguistic extensibility of the effects of surrounding sound context beyond the case of specific English continuum discussed in the previous study [9]. For this purpose, this study employed Japanese /ba/--/da/ continuum to test in the Japanese language context.

2. EXPERIMENTAL TASKS

According to earlier studies [1,3,8], the three characteristics of categorical perception are: (1) sharp identification results indicating a categorical boundary, (2) a discrimination peak at the boundary, and (3) chance discrimination performance for within-category comparison. Sharpness in identification results alone cannot be used as evidence of categorical perception because there are no legitimate criteria for measuring sharpness to be compared among stimuli that vary in acoustic dimensions [8]. In addition, the difference in methodology causes variation in sharpness. In fact, regular identification tasks encourage perception to be categorical because they directly ask listeners to answer the phonemic label of a presented stimulus. Any identification task of this sort forces listeners to use top-down criteria, which eventually provides them bias toward categorical
perception [1]. Thus, if we are to assess the degree of categorical perception with as less instructional bias as possible, we must observe listeners’ discrimination performance in addition to the identification results. This study employed a two-alternative forced-choice (2AFC) paradigm for identification and an AXB for discrimination.

For discrimination, we must observe a peak of discrimination performance for cross-category comparison, and a chance-level performance for the other within-category comparisons. However, to achieve the latter is rarely the case because within-category discrimination is usually better than chance. It is in part because listeners tend to be more sensitive to acoustic differences between stimuli than predicted by the original hypothesis of categorical perception, that is, people cannot discriminate between speech sounds unless the sounds have different phonemic labels [1,3]. Furthermore, the degree of the “peak” performance varies depending on experimental tasks. Therefore, by taking these facts into account, we translate the two characteristics into a single assumption: a cross-category comparison outperforms within-category comparisons. In this study, results of discrimination tasks are discussed based on this assumption.

Discrimination paradigm employed in the current experiments is an AXB comparison. This paradigm is preferred over the others because it is assumed that it would cause less response bias and would not unnecessarily rise difficulty of the task. Neither the AX nor the ABX, both of which seem simpler than the AXB, was used because of the response bias toward “same” response in the AX, and “X = B” response in the ABX [1,11]. Gerrits and Schouten [1] offered two additional options: 2I2AFC and 4I2AFC. In the former, listeners are provided with a pair of stimuli in the order of which is always different, e.g., AB, and are asked to tell the correct order of the stimuli. In the latter, one “same” stimulus pair and one “different” stimulus pair are presented, e.g., AA and BA, and listeners are to find one odd. The 2I2AFC was discarded for the present purpose because it induces labeling criteria when listeners explain what is first and what is second. The 4I2AFC was not preferred because it would make experiments under the Speech and the Nonspeech Conditions considerably difficult. Other details including the procedure of experiments of the current study are further explained in the corresponding sections.

3. MATERIALS

3.1. Acoustic Characteristics of /ba/ and /da/

The acoustic differences between /ba/ and /da/ appear in the transitions of the first and the second formants ($F_1$ and $F_2$), and to some extent the third formant ($F_3$) [12]. In natural speech, /b/ and /d/ each has a brief burst before formant transitions begin. The amplitude of the burst helps identify the place of articulation for both voiced and voiceless stops, although the effect is greater for the latter than for the former [13]; in the former, transitional formant information is sufficient to indicate difference in place of articulation [3,14]. Thus, in this paper, we synthesized a /ba/–/da/ continuum solely by changing formant trajectories following former research [3].

The overall trajectory of $F_1$ in /ba/ and /da/ is almost the same because $F_1$ is mainly affected by the opening of closure. For both stop closures, an articulator is attached at a particular position: at the lips for /b/ and at the tongue tip and alveolar ridge for /d/. Since $F_1$ frequency at the closure is close to zero, $F_1$ draws a very similar trajectory for both /b/ and /d/. Although the trajectories look similar, the difference between /b/ and /d/ is reported to manifest in transitional duration of $F_1$ [15], which is shorter for /b/ and longer for /d/. Thus, in the current research, transitional duration of $F_1$ was varied to help convey the distinction between /b/ and /d/. In addition, for /b/, $F_2$ transition rises relatively rapidly from the lower region to a steady-state frequency. For /d/, on the other hand, the $F_2$ transition falls relatively slowly to the steady-state from a starting frequency around its locus [15]. Although different acoustic properties are also observed for $F_3$, transition, the current experiment did not vary the characteristics of $F_3$ because varying $F_1$ and $F_2$ was sufficient to synthesize a reasonable continuum for the present purpose. Schematic representation of formant trajectories is shown in Fig. 1(b).

3.2. Synthesizing the /ba/–/da/ Continuum

For the current experiments, we created a nine-step /ba/–/da/ continuum using a cascade-formant synthesizer designed by Klatt and Klatt [16] (see also Klatt [17]).

First, the parameter values of synthetic syllables were decided on the basis of a recorded utterance of a syllable /ba/ presented in the following Japanese sentence produced by a male native speaker of Japanese: sorekara /ba/ ga aruto omoimasu (/sorekara ba ga aruto omoimasu/), “And I think that there is /ba/.” One token was selected from several tokens for synthesis. This carrier sentence was employed mainly for two reasons. First, the sentence did not contain voiced stops of our interest. During perceptual experiments, naturally produced speech sounds may be reliable reference source of perceptual judgments for listeners; thus, no potential reference should be included in a carrier sentence. Second, the sentence begins with a sibilant consonant. The carrier sentence of the former research started with a stop consonant which caused a short pause at the beginning. During an AXB discrimination task, the pause at the beginning ends up adding a brief extra time to stimulus interval; longer interval may affect discrimination performance.
During a recording, the speaker read the sentence several times as naturally as possible. The speaker was not instructed to read with an intended pause before or after the syllable, /ba/. However, probably because of an unusual experimental setting, the speaker often read the sentence with a brief pause before the target syllable.

After the recording, the mean values of the first four formants, i.e., $F_1$, $F_2$, $F_3$ and $F_4$, were obtained from the steady-state portion of one selected target token of /ba/, using the “Formant Listing” function of the Praat software [18]. The mean value of the fifth formant ($F_5$) was obtained using “Get Formants” function of the same software. These formant values were used to synthesize the steady state portion of the /ba/-/da/ continuum. In addition to the formant frequencies, the bandwidth of each formant, extra tilt of voicing spectrum and the quotient of vocal-fold opening were manipulated using the “tilt” (TL) function and the “quotient of vocal-fold opening” (OQ) function respectively (Appendix A) in order to sound naturally like the voice in the recorded utterance enough when the syllables were inserted back into the original sentence for the experiment under the Sentence Condition in Experiment 2 (see Sect. 3.3 for details). In Experiment 1, the continuum was presented in isolation, i.e., the Isolated Condition, but in Experiment 2, each syllable on the continuum was presented within a sentence, i.e., the Sentence Condition and within pure tones, i.e., the Non-speech Condition.

Next, we added formant trajectories to the $F_1$ and $F_2$ steady states (Figs. 1 and 2). The variation in $F_1$ and $F_2$ trajectories in the synthesized syllables progressed in nine steps from canonical /ba/ (Step 1) to /da/ (Step 9). Figure 1(b) provides overall schematic representation of formant trajectories of the continuum with a $F_0$ contour. $F_1$ and $F_2$ trajectories from 0 ms to 138 ms are dotted to show mobility of transition according to steps. This particular part of trajectories is enlarged in Fig. 2 to show precise transitional movements of $F_1$ and $F_2$ of each step. As illustrated in the figures, the $F_1$ trajectory differed from step to step in respect of the duration of the transition. The duration of the transition is the time length between the starting point (385 Hz at 100 ms) and the time point where the frequency reaches its steady-state level (790 Hz). The duration varied from 22 ms for Step 1 to 38 ms for Step 9 at intervals of 2 ms (Fig. 2). In contrast, the trajectory of $F_2$ differed in starting frequency (from 885 Hz to 1,690 Hz) as well as duration of transition, which changed from 22 to 38 ms as it did for $F_1$.

An $F_0$ contour was added (Fig. 1(b)) so that the created syllables would fit the original sentence as naturally as possible (see Sect. 3.3). $F_0$ was extracted from a recorded syllable /ba/ in the recorded sentence. First, the $F_0$ values of the recorded syllable were sampled at 20 equidistant time points; the average value, i.e., 135 kHz, was used as the peak value of the contour. Then, six other values at different time points were added to create a contour (Fig. 1(b)): 125 Hz at 0 ms, 125 Hz at 100 ms, 130 Hz at 150 ms, 135 Hz at 200 ms, 130 Hz at 300 ms and 125 Hz at 330 ms. These values are connected linearly. The value and the shape of the contour were adjusted as above on
the basis of the two authors’, who are native speakers of Japanese, perceptual impression regarding naturalness of a stimulus sentence with a synthesized syllable in the middle: for instance, for Step 1, sorekara Step 1 ga aruto omoimasu.

Each of the synthesized syllables had 100 ms silence at the beginning (Fig. 1). The amplitude was adjusted using the “amplitude of voicing” (AV) of the synthesizer. In the rising period, the amplitude rose abruptly from 0 dB to 60 dB between 95 ms and 100 ms (Fig. 1(a)). The amplitude in the falling period was changed linearly from 60 dB at 230 ms to 0 dB at 330 ms.

The digital output from the synthesizer, which was 16-bit resolution and a 10 kHz sampling rate, was converted to a 16 kHz sampling rate for experiments. The created syllables in Experiment 1 were called I-Step 1, I-Step 2, ..., I-Step 9, where “I” stands for the Isolated Condition. Figure 3 shows an example of I-Step stimuli (I-Step 1).

### 3.3. /ba/–/da/ Continuum in Different Contexts

For Experiment 2, two non-isolated conditions were prepared. In the Sentence Condition, each synthesized syllable was presented in a carrier sentence, i.e., sorekara ga aruto omoimasu, and in the Nonspeech Condition, in-between 1 kHz pure tones.

To create stimuli for the Sentence Condition, each of the synthesized syllables was inserted into the underlined portion of the sentence, for instance, sorekara Step 1 ga aruto omoimasu (Fig. 4). Stimuli under the Sentence Condition were called S-Steps 1 through 9. Recall that the synthesized syllables had a 95-ms complete silence and a 5-ms amplitude rising period at the beginning; thus, a target syllable reaches its maximum amplitude of voicing at 100 ms. Therefore, an inserted target syllable had approximately 100 ms silence after the preceding phrase “sorekara.” The preceding phrase ends at 769 ms (Fig. 4). Then, the target syllable was presented, and the phrase “ga aruto omoimasu” followed the target from 1,099 ms to the end, i.e. 2,295 ms.

For the Nonspeech Condition, 1 kHz pure tones replaced the naturally produced speech sounds that had been used under the Sentence Condition, as illustrated in Fig. 5. The stimuli under the Nonspeech Condition were called NS-Steps 1 through 9. The rising and falling periods of the pure tones differed depending on whether the tone preceded (Tone 1) or followed (Tone 2) the target syllable. Tone 1 took 50 ms to rise from 0 dB at 0 ms to 60 dB at 50 ms, and 100 ms to fall from 60 dB to 0 dB immediately before the target syllable. On the other hand, Tone 2 rose to 60 dB from 0 dB during the first 100 ms, and took 50 ms to fall to the end. The amplitude slopes that immediately preceded and followed the target, i.e., the falling period of Tone 1 and the rising period of Tone 2, were decided on the bases of amplitude change in the carrier sentence. The carrier that preceded a target in the S-Step stimuli, i.e., “sorekara,” has amplitude that decreased to the minimum from its maximum in approximately 100 ms, i.e., 102 ms. Similarly, the carrier after the target, i.e., “ga aruto…,” rises its amplitude from low level for “g” to the maximum level for “a” in about 100 ms, i.e., 106 ms. The amplitude slopes of the tones are to simulate these amplitude changes in the carrier sentence. On the other hand, 50-ms rising and
falling periods of Tone 1 and 2 were not intended to reflect amplitude contours of the carrier sentence; rather, we attempted to follow the amplitude pattern employed by Tomaru and Arai [9]. In their experiment, the inner ends of tones which directly preceded and followed a target stimulus inclined toward a target stimulus from higher to lower level of amplitude while the outer ends kept its maximum level with no amplitude decreasing period. Our stimuli were created to retain the similar pattern; however, to avoid impulsive onset and offset of NS stimuli which was assumed to be the case in the previous research, our stimuli included a slight degree of amplitude slope at both outer ends of the tones.

The amplitude of tones was adjusted so that auditory impressions of loudness of NS-Step stimuli did not greatly differ from those of S-Step stimuli. The intensity of Tone 1 and Tone 2 were 80.3 dB and 80.5 dB respectively according to the function of the Praat [18], “Get intensity (dB).” The intensity of a target was 77.9 dB, according to the same function. The difference of the intensity between Tone 1 and the target, and that between Tone 2 and the target were 2.4 dB and 2.6 dB, respectively. Since these differences were little, and the loudness of NS-Steps was almost the same as S-steps, our stimuli were considered to be adequate for the purpose of the investigation.

4. EXPERIMENT 1

4.1. Stimuli

Experiment 1 investigates discrimination of the /ba/-/da/ continuum presented in isolation. I-Steps 1 through 9, were used as stimuli.

4.2. Participants

Ten native speakers of Japanese (6 males and 4 females) with normal hearing participated in the experiment (mean age = 28.3 years old).

4.3. Procedure

Each participant had a familiarization session followed by an experimental session consisting of identification and discrimination tasks. Both the familiarization session and the experimental sessions were conducted using Praat software [18] and headphones (Audio-Technica ATH-M50) connected to a computer via a USB audio interface.

4.3.1. Familiarization

The familiarization session was conducted in order to let participants become accustomed to hearing synthesized sounds as speech syllables since a pilot research showed that some listeners may not be able to realize synthetic sounds as speech sounds. First, the participants heard five repetitions of the two edge stimuli of the continuum, i.e., I-Steps 1 and 9. Half of the participants heard I-Step 1 first, and then I-Step 9; the others did opposite. After hearing the edge stimuli, the whole continuum was presented. To those who heard I-Step 1 first, the continuum was presented from I-Steps 1 to 9. Those who heard I-Step 9 first did the reverse. During the familiarization session, participants were told to listen to the stimuli carefully and make internal self-judgments about which stimuli sounded like /ba/ and which sounded like /da/. No feedback was provided. Participants were able to adjust the sound to a comfortable listening level during the familiarization session, and they were told not to change the volume once the experiments had started.

4.3.2. Identification task

The 2AFC identification task was conducted. Participants reported whether the syllable they heard was “ba” or “da” by clicking on a button that appeared in Japanese kana letters on a computer screen. Each syllable along the continuum was repeated 10 times in a random order; thus, the participants made 90 identification judgments (9 stimuli × 10 repetitions). Other than the familiarization session that has been mentioned above, no additional practice session was conducted before the task. Participants did not receive any feedback during the task.

4.3.3. Discrimination task

For the discrimination task, the AXB discrimination paradigm was employed. In this paradigm, a participant is presented with three stimuli in AXB order, where X matches A or B. After hearing the stimuli, s/he judges whether the second sound (X) matched the first sound (A) or the third (B). The stimuli were paired such that each pair (AB) differed by two steps in the continuum: I-Step 1 with I-Step 3, and so on. Additionally, the edge stimuli, I-Steps 1 and 9, were paired as an “edge pair” and presented together with the other stimulus pairs as fillers.

All stimulus pairs, including the edge pair, were arranged into four permutations: AAB, ABB, BAA and BBA. Each permutation was repeated three times. The stimuli were presented to participants in a random order. The stimulus interval was 300 ms following Tomaru and Arai [9]. The total number of discrimination judgments was 96: 8 pairs, including the edge pair × 4 presentations × 3 repetitions. In order to get accustomed to the AXB paradigm, participants had a short practice session before the main experiment.

4.4. Results of Experiment 1

Because one participant reported that he could not recognize the stimuli as either /ba/ or /da/, data provided by this participant were excluded from the analysis.

4.4.1. Identification results under Isolated Condition

Results under the Isolated Condition are summarized in Fig. 6. Each thin line represents the average response of each listener: if a listener reported /ba/ five times out of 10 repetitions of the same stimulus, the average response
The nine listeners’ averaged responses are fitted using a logistic model (1).

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

(1)

In the model (1), \( y \) represents the percentage responses of /ba/, and \( x \) represents the step number on the continuum. The parameter \( a \) represents the slope of the curve; the parameter \( b \) corresponds to the 50% crossover point, which is interpreted as a categorical boundary. The bold line in Fig. 6 indicates the result of the fitting. The slope of the fitted curve is 2.3 and the 50% crossover, the categorical boundary, is 5.2. Therefore, if categorical perception is to be observed under this condition, we expect a discrimination peak at the stimulus pair 4–6, which crosses the boundary.

4.4.2. Discrimination results under Isolated Condition

For each listener, the percentage of correct judgments was calculated for each pair by dividing the number of correct response to the pair by the total number of presentation of the pair. The results are shown in Fig. 7. Discrimination accuracy reached a peak for the pair 4–6, which crosses the boundary at 5.2, as indicated by the identification function. Recall that in our definition, the peak performance of discrimination is assessed by comparing performance for a cross-boundary pair with that for within-category pairs. Assuming that the pair 4–6 is the only pair that crosses the categorical boundary based on the identification results, it is reasonable to divide the stimulus pairs into three groups: (1) pairs which were predominantly categorized as /ba/, i.e., pairs 1–3, and 2–4 (within-/ba/ pairs), (2) pairs which were predominantly categorized as /da/, i.e., pairs 6–8, and 7–9 (within-/da/ pairs), and (3) the pair with different labels, i.e., pair 4–6 (cross-boundary pair). Note that the pairs that included I-Step 5 are eliminated from the groups because the step is almost a boundary between the two categories, and is assumed to be heard as a good example of neither /ba/ nor /da/; no rigid prediction can be made for such pairs that contain a boundary step in terms of categorical perception. A one-way ANOVA with repeated measures revealed a main effect (\( F(4, 32) = 13.29, p < 0.001 \)). A post hoc multiple comparison with Bonferroni correction indicated that the percent correct for the cross-boundary pair and that of the within-/ba/ pairs were significantly different (\( p < 0.01 \)). Similarly, the difference between the cross-boundary pair and pairs in the within-/da/ group was also significant (\( p < 0.01 \)). No significant difference was found between the other combinations of the pairs. The results ascertained that the /ba/-/da/ continuum in isolation was categorically perceived as reported by previous studies.

5. EXPERIMENT 2

5.1. Stimuli

Identification and discrimination performances were investigated under two non-isolated conditions. S-Steps 1 through 9, were used for the Sentence Condition, and for the Nonspeech Condition, NS-Steps 1 through 9.

5.2. Participants

Twenty native speakers of Japanese, who had not participated in Experiment 1, were recruited as participants (9 males and 11 females, mean age = 21.1 years old) for experiments under both conditions.

5.3. Procedure

For both conditions, each participant completed an identification task followed by a discrimination task. Half of the participants completed the tasks under the Sentence Condition first, and the others completed the tasks under the Nonspeech Condition first. As in Experiment 1, each
participant had a familiarization session for each condition prior to the relevant main experimental session. All other experimental conditions were identical to those in Experiment 1, except as noted below.

5.3.1. Experiment under the Sentence Condition

Each participant completed a short familiarization session before the main task, as in Experiment 1. First, half of the participants were presented with five repetitions of S-Step 1 and then five repetitions of S-Step 9; the other half heard five repetitions of S-Step 9 first and then five of S-Step 1. Next, all syllables in a continuum were presented in order from S-Step 1 to S-Step 9 to the first half of the participants, and in the reverse order to the other half.

Then, participants completed an identification task. In this task, participants were told that they would hear sentences they heard during the familiarization session, and were instructed to report whether they had heard /ba/ or /da/ in the blank portion of the carrier sentence: sorekara ga aruto omoimasu, which was presented to the listeners in Japanese transcription. Each participant heard 10 repetitions of each stimulus (9 syllables × 10 repetitions). No practice session was prepared for the identification task, and no feedback was given during the task.

After the identification task, a discrimination task was conducted. In the instructions, each participant was told that s/he would be presented successively with three sentences from the familiarization session, and was requested to judge whether the syllable in the second sentence (X) matched that in the first sentence (A) or that in the third sentence (B). A short practice session was given in order to accustom the participants to the experimental procedure. As in Experiment 1, inter-stimulus durations were 300 ms. Moreover, an edge pair, S-Steps 1–9, was also prepared under the Sentence Condition. Thus, the total number of judgments was 96: 8 pairs (the edge pair included) × 4 presentations × 3 repetitions.

5.3.2. Experiment under the Nonspeech Condition

As the experiments under the Sentence Condition, a familiarization session was conducted before an identification task. The procedure of the familiarization session was identical to that held under the Sentence Condition.

In the identification task, a participant heard a NS stimulus; then, s/he then judged whether the syllable appearing between the pure tones was /ba/ or /da/. Each stimulus was repeated 10 times (9 syllables × 10 repetitions). No feedback was given.

Next, a discrimination task was conducted in which each pair of NS stimuli was presented successively in AXB order. Each participant was instructed to judge whether the syllable in X matched that in A or that in B. The participants had a practice session before the main task. All other experimental conditions were the same as those in the discrimination task under the Sentence Condition, including the number of stimulus repetitions. The edge pair was included as a filler under this condition, as well.

5.4. Results under the Sentence Condition

5.4.1. Identification results

Identification data provided by all participants are summarized in Fig. 8. As in Experiment 1, each thin line in the figure represents each listener’s average response of /ba/ for each stimulus. The average responses of the 20 participants were fitted using the model introduced in Experiment 1 (Eq. (1)). The slope of the fitted curve was 1.9, and the point of 50% crossover was 5.9.

Notice that the crossover point under this condition shifted toward the /da/ side of the continuum compared to the boundary observed under the Isolated Condition. Under the Sentence Condition, the categorical boundary is located almost right at S-Step 6, expecting a discrimination peak for a stimulus pair that strides S-Step 6. In addition, because of the boundary shift, Step 1 and Step 2 may have sounded extreme examples of /ba/ under the Sentence Condition. Therefore, it was reasonable to expect irregular discrimination judgments for the pairs that included S-Steps 1 and 2.

5.4.2. Discrimination results

The percentage of correct judgments was calculated by using the same method as in Experiment 1. As foretold, many listeners’ percentage of correct judgments for the pairs 1–3 and 2–4 was below a chance level ranged from 8 to 42%, suggesting that listeners responded the opposite of what they were expected to respond for these extreme pairs: for the “same” pair, they responded “different” and for the “difference” pair, they responded “same.” Such a tendency may be caused by the boundary shift. Therefore, we discarded the results for the pairs 1–3 and 2–4.

Fig. 8 Identification results under theSentence Condition. All listeners’ responses (thin lines) are combined using a logistic model (bold line).
In addition, analysis for each listener found fairly clear tendency of categorical perception for four listeners while the others showed no such tendency. Since it was possible that these four listeners used criteria that were different from the others, results of the four listeners were excluded from the overall analysis of discrimination data; the results are discussed in the following section and Sect. 6.

The percentages of correct discrimination judgments were averaged over the remaining 16 participants (Fig. 9). Based on the assumption of categorical perception, a discrimination peak was expected for the stimulus pair 5–7 that crosses the categorical boundary. As in Experiment 1, the stimulus pairs were categorized into three groups according to the identification results. However, again, since we could make no prediction for such pairs that included the boundary step, those that included S-Step 6 were excluded from the statistical analysis, although the pairs are included in the figure. Thus, pairs that were subject to statistical analysis were: (1) the pair mostly heard as /ba/, i.e., the pair 3–5 (within-/ba/ pair), (2) that mostly heard as /da/, i.e., the pair 7–9 (within-/da/ pair), and (3) the pair which strides the categorical boundary, i.e., pair 5–7 (cross-boundary pair). A one-way ANOVA with repeated measures found no main effect among pairs in these groups ($F(1,3;19.44) = 1.68, p > 0.05$).

5.4.3. Interim summary and discussion

The identification results showed that the 50% crossover point under the Sentence Condition moved closer to the /da/ end compared to the point under the Isolated Condition, which was 5.2. This means that listeners tend to hear more /da/ than /ba/ when stimuli were presented within a sentence. This boundary shift is consistent with the former research, where a categorical boundary shifted toward /l/ in a /la/-/la/ continuum when stimuli were presented in non-isolated contexts. The shift was observed for both speech and pure-tone conditions [9]. Thus, we expected the shift under the Nonspeech Condition in the present research as well.

The discrimination performance did not reflect labeling results under the Sentence Condition. That is, a discrimination peak was hidden, if not completely disappeared, under this condition. However, analysis for each individual listener found that some people showed tendency of categorical perception. As seen in the figures in Appendix B, the four listeners showed a remarkable improvement of discrimination performance at the boundary pair. It is possible that these listeners used criteria different from the others. Or, it is also possible that the degree of effects of surrounding sound contexts may differ among individuals. Brief discussion on this topic is given again in Sect. 6. Any detail report on this topic should be one of the future issues.

5.5. Results under the Nonspeech Condition

5.5.1. Identification results

Each listener’s average response was calculated for each stimulus (thin line in Fig. 10). All listeners’ average responses were fitted using the same model (Eq. (1)) as used for the other analysis (bold line in Fig. 10). The 50% crossover point was 6.0, indicating that the categorical boundary was right at NS-Step 6. The slope of the function was 2.4. If the continuum under this condition is perceived categorically, a discrimination peak is expected to be observed for the pair 5–7, which crosses the boundary.

5.5.2. Discrimination results

The percentage of correct judgements was calculated using the same method as in the other analysis. As expected, the boundary shift to the /da/ end was observed under this condition. However, unlike under the Sentence Condition, there were only two people whose percentages of correct judgments for the extreme pairs, i.e. the pairs 1–3 and 2–4, were remarkably low, that is, 17%. Such difference of performance between the two non-isolated conditions may also be the effects of the type of
surrounding sounds (discussed further below). Therefore, for the analysis of this condition, we included the pairs 1–3 and 2–4 and excluded the two listeners. In addition, another listener who marked the percentage correct of 17% for the pair 3–5 was excluded from the analysis because it seemed possible that this pair was still an extreme case for this listener.

Figure 11 indicates the percentage of correct discrimination responses averaged over the remaining 17 participants. Again, a peak in discrimination accuracy was predicted for pair 5–7, which crossed the categorical boundary indicated by the identification results, that is, NS-Step 6. The stimulus pairs were divided into three groups, and analyzed: (1) pairs frequently heard as /ba/, that is, pairs 1–3, 2–4, and 3–5 (within-/ba/ pairs), (2) those frequently heard as /da/, that is, the pair 7–9 (within-/da/ pair), and (3) the pair that crosses the categorical boundary, that is, the pair 5–7 (cross-boundary pair). As in the Sentence Condition, the pairs that have the boundary step, i.e., NS-Step 6, were excluded from the analysis, although they are shown in the figure. A one-way ANOVA with repeated measures found the main effect ($F(4, 64) = 12.2$, $p < 0.001$); a post hoc test with Bonferroni Correction revealed the significant difference between each pair in the within-/ba/ group and the cross-boundary pair ($p < 0.01$). The test also found that difference between the within-/da/ pair and the cross-boundary pair was significant ($p < 0.01$). The comparison between the any other pairs was found not to be significant ($p > 0.05$). The results under the Nonspeech Condition showed a discrimination peak.

5.5.3. Interim summary and discussion

The results under the Nonspeech Condition showed characteristics what we have obtained under the Isolated and the Sentence Conditions: a discrimination peak like under the Isolated Condition, and the categorical boundary shift toward the /da/ end like under the Sentence Condition. As noted earlier, the former research [9] mentioned about boundary shift under non-isolated conditions. Therefore, it was reasonable to observe a boundary shift under the Nonspeech Condition. Since the shift is observable under both types of non-isolated conditions, it is probable that it has something to do with the overall length of comparing stimuli rather than the type of surrounding sounds. Nevertheless, the effect of boundary shift may act differently on listeners depending on the surrounding sound contexts. Recall that while some listeners’ responses showed the opposite of what was intended for the /ba/-side of pairs that included extreme steps, i.e., the pairs 1–3 and 2–4, under the Sentence Condition, such response tendency was not observed except for few cases, under the Non-speech Condition.

6. GENERAL DISCUSSION

6.1. Role of Labeling Mediation in Speech Perception

At this point, several explanations for the results seem reasonable. For example, as Tomaru and Arai [9] points out, surrounding speech sounds may cause high perceptual burden in terms of working memory so that a listener has to sacrifice the process of labeling to complete discrimination task. However, instead of listing the possible explanations, discussion here focuses on the fact that a discrimination peak was not observable even for the voiced stop continuum when heard within a sentence. Recall that a voiced stop continuum needs no overt labeling mediation to show categorical perception, or more precisely, a discrimination peak. Therefore, a peak should have been observed for the voiced stop continuum under the Sentence Condition if it is only the overt labeling process that is interfered by surrounding speech context; however, it was not the case. Therefore, not only does the present research support the original argument of Tomaru and Arai [9], it also implies that the process of naturally spoken everyday speech does not depend on labeling mediation of any sort. The results of the current experiments further provide the idea that speech perception has both a categorical and a non-categorical nature. This idea is worth being mentioned...
given that it has been a challenge for researchers over the years to reliably define a line between continuous and categorical speech perception. In fact, the two types of perception have often been discussed as two separate perspectives. However, if speech perception may show both continuous and categorical aspects depending on contexts as suggested, scholars must be aware which kind of perception it is that they should be interested in in a given case.

6.2. Issues for Further Consideration

6.2.1. Rich vs. sparse perceptual data

It is worth noting that perceptual tendency is often influenced by stochastic data [19–22], e.g., the relative appearance probability of particular phonemes, words and so on. In the present research, as well as in Tomaru and Arai [9], the sentence stimuli were grammatically acceptable, but might be meaningless for many people because the sentences would not be said or heard in everyday conversational settings. The topic is not thoroughly discussed here; however, effects of stochastic distribution of not just phonemes but also words and sentences on discrimination performance should be investigated in the future research.

6.2.2. Individual differences

Finally, issues related to response variability attributes to individual differences need to be addressed. As noted above, four listeners showed tendency of categorical perception under the Sentence Condition. Therefore, even though surrounding speech sounds interfere with categorical perception to the great extent, there might be some ways to escape from the effects. One possibility is that some listeners may use specific discrimination criteria that is not affected by surrounding speech sounds. It is also possible that the ability to ignore carrier sounds, or to focus on target sounds, may vary among each individuals: some may be able to ignore surrounding sounds that are unrelated to a task easily whereas others may not. Speech-likeness, which involves linguistic and acoustic naturalness, of carrier sounds may also be one of the factors that cause individual differences. Future study should clarify this point.

ACKNOWLEDGMENTS

The contents of this paper were partially reported at the 2014 Autumn Meeting of the Acoustical Society of Japan [Tomaru and Arai, “Discrimination and identification of Japanese /ba/-/da/ continuum within a sentence by Japanese-speaking listeners,” Proc. Autumn Meet. Acoust. Soc. Jpn., pp. 489–492 (2014)]. They were also partially reported in the doctoral dissertation by the first author, Kanako Tomaru, submitted in partial fulfillment of the doctoral requirements at Sophia University. This research is supported by Japan Society for the Promotion of Science Kakenhi Grant Number 26884051.

REFERENCES

[1] E. Gerrits and M. E. H. Schouten, “Categorical perception depends on the discrimination task,” Percept. Psychophys., 66, 363–376 (2004).
[2] B. McMurray, R. N. Aslin, M. K. Tanenhaus, M. J. Spivey and D. Subik, “Gradient sensitivity to within-category variation in words and syllables,” J. Exp. Psychol.: Hum. Percept. Perform., 34, 1609–1631 (2008).
[3] A. M. Liberman, K. S. 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).
[4] 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).
[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] M. Mochizuki, “The identification of /r/ and /l/ in natural and synthesized speech,” J. Phon., 9, 283–303 (1981).
[7] L. Polka and W. Strange, “Perceptual equivalence of acoustic cues that differentiate /r/ and /l/,” J. Acoust. Soc. Am., 78, 1187–1197 (1985).
[8] P. Howell and S. Rosen, “Natural auditory sensitivities as universal determiners of phonemic contrasts,” Linguistics, 21, 205–236 (1983).
[9] K. Tomaru and T. Arai, “Discrimination of /ra/-/la/ speech continuum by native speakers of English in Non-isolated conditions,” Acoust. Sci. & Tech., 35, 251–259 (2014).
[10] A. F. Healy and B. H. Repp, “Context independence and phonetic mediation in categorical perception,” J. Exp. Psychol.: Hum. Percept. Perform., 8, 68–80 (1982).
[11] E. Gerrits, The Categorisation of Speech Sounds by Adults and Children, Ph. D. Netherlands National Graduate School of Linguistics (2001).
[12] K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre and F. S. Cooper, “Effect of third formant transitions on the perception of the voiced stop consonants,” J. Acoust. Soc. Am., 30, 122–126 (1958).
[13] R. N. Ohde and K. N. Stevens, “Effects of burst amplitude on the perception of stop consonant place of articulation,” J. Acoust. Soc. Am., 74, 706–714 (1983).
[14] A. M. Liberman, P. C. Delattre, F. S. Cooper and L. J. Gerstman, “The role of consonant-voice transitions in the perception of the stop and nasal consonants,” Psychol. Monogr.: Gen. Appl., 68(8), pp. 1–13 (1954).
[15] R. D. Kent and C. Read, The Acoustic Analysis of Speech (Singular Publishing Group, San Diego, 1992).
[16] 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).
[17] D. H. Klatt, “The new MIT speech VAX computer facility,” in Speech Communication Group Working Papers IV, Research Laboratory of Electronics (MIT, Cambridge, MA, 1984), pp. 73–82.
[18] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer” [computer program], v. 5.3.23, retrieved 7 August 2012 from http://www.praat.org/.
[19] J. B. Pierruhumbert “Exemplar dynamics: Word frequency, lenition and contrast,” in Frequency and the Emergence of
Table 1  Parameter values of the synthesized steps.

| Parameter name            | Value                        |
|---------------------------|------------------------------|
| $F_0$                     | 125–135 Hz                   |
| $F_1/F_1$ bandwidth       | 790 Hz/60 Hz                 |
| $F_2/F_2$ bandwidth       | 1,190 Hz/105 Hz              |
| $F_3/F_3$ bandwidth       | 2,640 Hz/150 Hz              |
| $F_4/F_4$ bandwidth       | 3,700 Hz/200 Hz              |
| $F_5/F_5$ bandwidth       | 4,800 Hz/1,000 Hz            |
| Open quotient             | 80%                          |
| Extra tilt of voicing spectrum | 8 dB                      |

APPENDIX A

Table 1 shows parameters manipulated to synthesize the /ba/-/da/ continuum. Open Quotient is calculated as voicing open time divided by period.

APPENDIX B

Results of the four listeners excluded from the analysis under the Sentence Condition are summarized in Figs. 12 and 13. In Fig. 12, the percentage of /ba/ responses are averaged over the four listeners, and fitted using the Eq. (1). The slope was 2.0, and the point of 50% cross over, i.e., categorical boundary, was 5.9. Thus the boundary for these listeners is almost at S-Step 6. Figure 13 shows discrimination performance; the four listeners did remarkably well for the pair that crosses the categorical boundary, i.e. the pair 5–7.

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 finished her PhD at the Faculty of Science and Technology, Sophia Univ., in 2015. 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.