Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 5aSCb: Production and Perception II: The Speech Segment (Poster Session)

5aSCb45. Perception of /ra/-/la/ contrast in different contexts: mono-syllable vs. sentence

Kanako Tomaru* and Takayuki Arai

***Corresponding author's address: Sophia University, Chiyoda-ku, 102-0094, Tokyo, Japan, himawari.kanako@gmail.com**

A strict assumption that underlies categorical perception hypotheses is that two speech sounds are discriminable only when they cross a categorical boundary emerging from an identification function. A number of researchers have attempted to show the categorical perception of the sounds in question (particularly consonants) through discrimination and identification tests. However, these tests are usually used for mono-syllables or mono-syllabic words. In this study, we first investigated whether the perception of /ra/-/la/ contrast indicate any categorical perception in a mono-syllabic CV context, using synthesized syllables. Next, we tested whether categorical perception can be also observed in a sentence through perceptual discrimination and identification experiments. Results showed that 1) a discrimination peak predicted by the identification function was obtained only for the mono-syllabic context, and 2) discrimination accuracy in the sentence condition was consistently low. These results suggest that categorical perception in a strict sense may not be evident in the perception of a sentence.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4798755]
Received 8 Jan 2013; published 2 Jun 2013

INTRODUCTION

It is widely known that native speech sounds, especially consonants, are perceived categorically. Such perception is called "categorical perception" [1]. According to previous researches that are concerned with native speech perception, evidence of the categorical perception appears in listeners' identification and discrimination functions. In the identification function, we see an s-shaped curve in percent response. For example, when listeners perceive syllables in a continuum that changes from /ra/ to /la/, /la/-responses rapidly increases at a certain point where listeners suddenly hear the stimuli as /la/; such point is called a categorical boundary. In the discrimination function, on the other hand, listeners' discrimination accuracy reaches the peak at the stimulus pair which crosses the categorical boundary. For instance, in the case of the same /ra/-/la/ continuum perception, listeners discriminate syllables from the /ra/-side of the edge (e.g. Step 1) to the /la/-side of the edge (e.g. Step 10). Usually, the syllables are paired such that each pair differs by arbitrary steps from one edge to the other along the continuum, e.g. Step 1 ― Step 3, Step 2 ― Step 4, and so on. Listeners' discrimination performance becomes most accurate when discriminating the paired stimuli one of which belongs to /ra/ category, and another of which belongs to /la/ category. In a strict sense of the categorical perception, an identification function is assumed to predict a discrimination function. That is, only the stimuli identified as different categories can be perceived to be different.

In previous studies, the categorical perception was demonstrated using mono-syllables, or mono-syllabic words presented in isolation (for example, [1, 2]). The present study attempts to reveal whether or not the categorical perception is observed in conditions other than mono-syllables (or mono-syllabic words) in isolation: in a sentence. In the present report, we first demonstrated that perception of a /ra/-/la/ continuum indicated a categorical perception in isolation condition (Experiment 1). Next, we tested whether categorical perception could be observed in perception of the same syllables in the continuum presented in a sentence (Experiment 2). Results showed that 1) a discrimination peak predicted by the identification function was obtained only for the mono-syllabic context, and 2) discrimination accuracy in the sentence condition was consistently low. These results suggest that categorical perception may not be evident in the perception of syllables presented in a sentence.

MATERIALS

For the experiments, we synthesized a series of /ra/-/la/ continuum using cascade-formant software synthesizer designed by Klatt and Klatt [3]. The created syllables were used as stimuli in both Experiment 1 and Experiment 2. In Experiment 1, the synthetic /ra/-/la/ syllables were presented in isolation. In Experiment 2, each of the syllables were inserted into a sentence, and presented to listeners as part of the sentence.

Synthesizing a /ra/-/la/ Continuum

The /ra/-/la/ syllables were created based on a male speaker's utterance from the TIMIT corpus [4]. The speaker ID of the speaker was MKAM0. To synthesize the continuum, we obtained formant frequency values of the speaker from a vowel /ʌ/ in "pronunciation" of a sentence "Clear pronunciation is appreciated" (sentence ID: sx236) produced by the speaker. The speaker's first three formant frequencies of the selected part of the vowel /ʌ/ were averaged over time. The averaged F1, F2, and F3 were 670 Hz, 1357 Hz, and 2788 Hz, respectively. These values served as steady state values of the vowel /a/ in the /ra/-/la/ syllables (Fig. 1). The same sentence produced by the same speaker was used as a sentence into which the synthetic syllables were inserted (See the following section for details).

Next, we calculated starting frequencies of F1, F2, and F3 (*F1s*, *F2s*, and *F3s* in Fig. 1) following MacKain *et al.* [2]. For F3, we also calculated the values at the inflection (at 135 ms). Figure 1 illustrates schematic representation of trajectories of the first five formants. Because F3 transition is one of the primary cues of the perception of the /r/-/l/ contrast, only *F3s* and the value at the inflection varied in nearly equal ten steps from /ra/ configuration (Syllable-Step1) to /la/ configuration (Syllable-Step10). *F3s* varied from 1609 Hz to 2827 Hz. The inflection values varied from 1717 Hz to 2827 Hz. *F1s* and *F2s* were fixed throughout the continuum. F4 and F5 were 3250 Hz and 3700 Hz, respectively. The values of F4 and F5 were default values of the synthesizer, and were fixed throughout the continuum.

All synthesized syllables lasted for 350 ms including 100-ms rising and falling periods of amplitude. Figure 1 shows the 250-ms long period without the rising or the falling period. Amplitude in the rising and the falling periods

were changed linearly in the decibel scale from 0 dB at 0 ms to 60 dB at 100 ms, and from 60 dB at 250 ms to 0 dB at 350 ms by using the parameter, "amplitude of voicing (AV)," of the synthesizer. During the rising period, all formants had the onset values which are illustrated in Fig. 1. For the experiment, additional 100 ms silent periods were added before and after each of the synthesized syllables.

In addition to the first three formants, the speakers' F0 contour of the "pronunciation" part of the original utterance was approximated, and reflected in the synthesized syllables in order to reduce "un-naturalness" when the syllables were inserted back into the original sentence (See Fig. 3). In order to approximate the F0 contour of the original utterance, we first sampled the F0 of the "pronunciation" part of the sentence at 25 equidistant time points. Then, we approximated the original F0 with six time points by averaging the sampled values at every five time points along the 25 equidistant time points. All synthetic syllables had the same F0 contour.

Digital outputs from the synthesizer (16-bit resolution and 10-kHz sampling rate) were converted to 16-bit resolution and 16-kHz sampling rate. An example of the synthesized syllables is presented as Fig. 2.

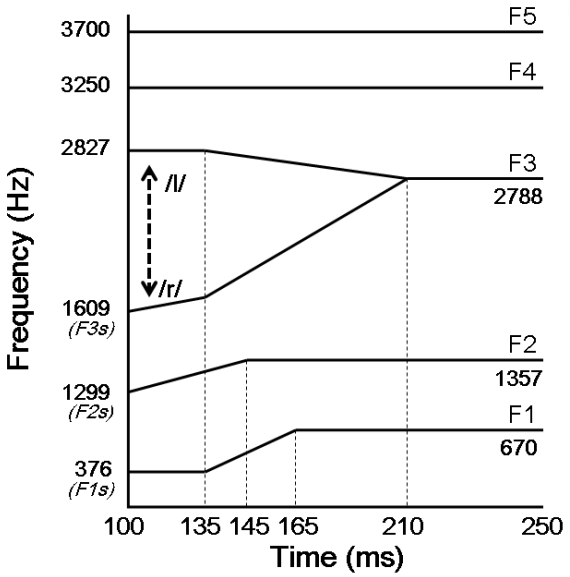

FIGURE 1. Schematic representation of trajectories of format frequencies from F1 to F5.

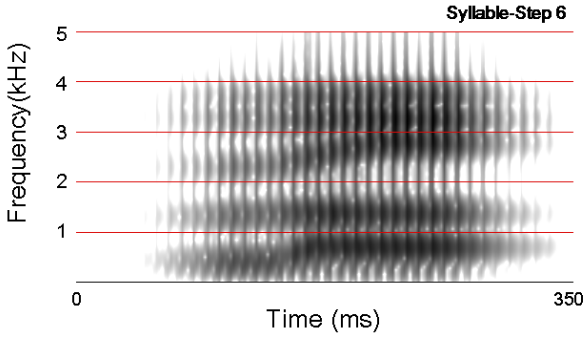

FIGURE 2. Spectrograms of a synthesized syllable (Syllable-Step 6).

Inserting the Syllables into Sentences

The created synthetic syllables (Syllable-Step1 through Syllable-Step 10) were replaced with the word "pronunciation" of a sentence "Clear pronunciation is appreciated" produced by MKAM0. Thus, new ten sentences with the synthetic syllables were created. Such sentences looked like this: "Clear /ra/-/la/ is appreciated." The new sentence with Syllable-Step 1 was called as Sentence-Step 1, and so on. These stimulus sentences were used for Experiment 2. Because F0 contour of the original utterance was approximated, experimenters found the ten

sentences natural enough to be perceived as completed sentences as if the syllables were originally there. Fig.3 shows an example of the newly created sentences.

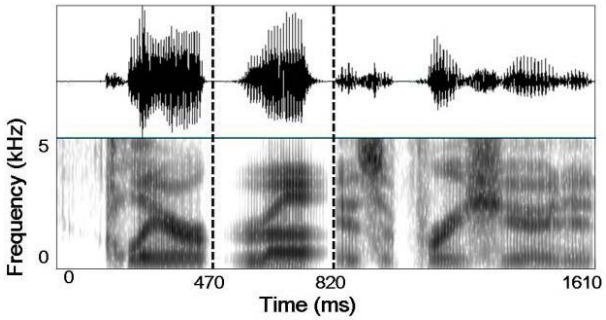

FIGURE 3. An example of the spectrograms of the sentences with synthetic syllables (Sentence-Step 1).

EXPERIMENT 1

The purpose of Experiment 1 was to investigate perception of a /ra/-/la/ continuum in isolation. The experiment attempted to demonstrate categorical perception of the continuum by native speakers of English in such isolation condition. In Experiment 1, two groups of native speakers of English were recruited: one was for an identification task and the other was for a discrimination task.

Stimuli

Syllables created in MATERIALS (Syllable-Step 1 through Syllable-Step 10) were used as stimuli.

Listeners

Two groups of native speakers of English participated in Experiment 1. One group that consisted of nine native speakers of English (6 males, 3 females) participated in the AXB discrimination task (AXB-exp1). Eight of them were from the United States, and one of them was from United Kingdom. Ages ranged from 20 to 21 years (mean 20.7 years). All of the participants were exchange students at Sophia University, Japan. They have resided in Japan for 3 months to 11 months at the time of the experiment (mean 4.1 months).

The second group consisted of two listeners (1 male, 1 female) participated in the identification task (ID-exp1&2). The listeners were both from the United States. Ages were 21 and 22 years (mean 21.5 years). Both of them were exchange students at Sophia University, and have resided in Japan for 3 months at the time of the experiment.

None of the listeners reported any hearing problems at the time of the experiment.

Procedure

The experimental and the practice sessions were given to listeners by using the Praat software [5]. All stimuli were presented diotically via Sennheiser HDA 200 headphones at participants' comfortable listening level.

Identification Task

In the identification task, listeners of ID-exp1were instructed to judge if the presented syllables were "ra" or "la". Thus, the identification task took the form of two-alternative forced-choice (2AFC). The listeners heard ten repetitions of the ten synthesized syllables (10 stimuli × 10 repetitions = 100 judgments). Stimuli were presented once, and listeners were not able to replay. Stimuli were presented randomly to the listeners. The listeners had a practice session for task familiarization.

AXB discrimination Task

Listeners of AXB-exp1 discriminated the synthetic syllables with an AXB paradigm. In this task, the synthesized syllables were paired such that each pair (AB) differed by two steps in the continuum, i.e. Syllable-Step 1－3, Syllable-Step 2－4, …, Syllable-Step 8－10. Two-step comparison was preferred because a pilot AXB experiment with both two- and three-step comparison revealed that the three-step comparison was too easy for English listeners that their discrimination accuracy hit more than 60% at every pair, and we did not obtain discrimination peak for the three-step comparison.

The listeners were instructed to judge if the second syllable (X) matched to the first (A), or to the third (B), and were also instructed to guess if necessary. Listeners made responses by clicking buttons, which said “first” and “third” on a computer screen. Paired stimuli were arranged into four permutations (AAB, ABB, BAA, and BBA). The experimental session had three repetitions for each presentation, so listeners made 12 judgments for each pair. This makes total of 96 judgments (8 pairs × 4 presentations × 3 repetitions = 96 judgments). The AXB presentations were made randomly. For purpose of familiarizing the listeners with the tasks, a short practice session was held prior to the main experimental session.

Results

Identification Task

The responses were averaged over two listeners (Fig. 4). As Figure 4 indicates, a categorical boundary seemed to locate at Step 5, where the /ra/ responses dropped from 75% at Step 4 to 55%. Thus, the peak of a discrimination function was expected to appear at such pair that strides across Step 5, i.e. the pair 4－6. In fact, the /ra/ responses for Step 4 indicated 75% and the /la/ responses for Step 6 were also 75%. Therefore, in other words, the pair 4－6 was the only pair, elements of which belonged to different categories, i.e. /ra/ or /la/. This also suggested that a discrimination peak, if there would be any, would appear at the pair 4－6.

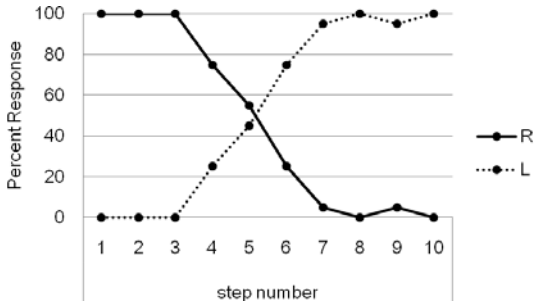

FIGURE 4. Averaged percent response (%) of /ra/ (solid line) and /la/ (dashed line) for identification of syllables presented in isolation.

Discrimination Task

Figure 5 indicates the discrimination function obtained from nine listeners of AXB-exp1. As the figure indicates, the function has a peak at the pair 4－6.

The peak at the pair 4－6 was well predictable from the identification function where categorical boundary was shown to locate at Step 5. In order to assess that the correct rate at the pair 4－6 was significantly higher than those at other pairs, we compared the correct rate at the pair 4－6 and those at the other pairs. For the statistical analysis, we assumed that the pair 4－6 was the only pair that crossed categorical boundary, and named the pair as the cross-boundary pair; the other pairs were named as within-category pairs. The within-category pairs that were perceived as /r/ were the pairs 1－3, 2－4, and 3－5 (within-/r/). The within-category pairs that were perceived as /l/ were the pairs 5－7, 6－8, 7－9, and 8－10 (within-/l/). If the correct rate at the cross-boundary pair was significantly higher than the rates at the within-category pairs, we would be able to call the increase of the correct rate at the pair 4－6 as a discrimination peak. The ANOVA with repeated measures revealed the main effect of category types ($F(2, 16)$ =

7.57, p = .005). The difference between the cross-boundary pair and the within-/r/ was significant (p = .009), and the difference between the cross-boundary pair and the within-/l/ was also significant (p = .028). Thus, the discrimination peak obtained for the pair 4－6 matches to the categorical boundary emerging from the identification function.

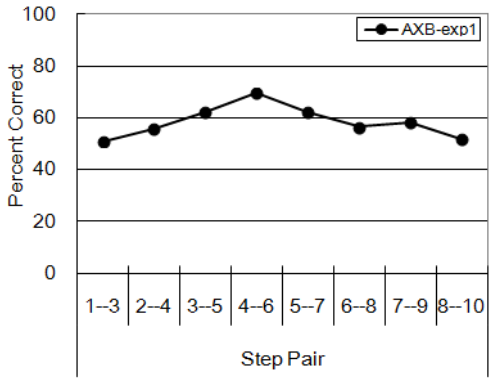

FIGURE 5. Averaged correct rate (%) for discrimination of each pair in the isolation condition.

Discussion of Experiment 1

In Experiment 1, two groups of native speakers of English were recruited to identify or discriminate synthetic syllables presented in isolation. Although listeners for the identification task were different from those for the discrimination task, results of the experiment appeared to be consistent. That is, the peak of the discrimination function of one group of listeners reflected the categorical boundary emerging from the identification function of another group of listeners. The discrimination peak at the pair 4－6 was well predicted from the identification function. That is, the identification function indicated that Step 4 and Step 6 belonged to different categories, i.e. /ra/ or /la/. In this sense, the two-step comparison was suitable for the current experiment. The present study interprets this "match" between the identification and the discrimination functions as evidence of a categorical perception. Thus, in the isolation condition, synthetic syllables were shown to indicate the categorical perception.

EXPERIMENT 2

In Experiment 2, we investigated whether or not a categorical perception is observed in perception of /ra/-/la/ syllables presented in sentences. In the experiment, the synthetic syllables created in MATERIALS were inserted into a sentence context, i.e. "Clear __ is appreciated." The stimulus sentences with the target syllables looked like this: "Clear /ra/-/la/ is appreciated." We conducted identification and discrimination tasks. Two groups of native speakers of English participated in either task.

Stimuli

The stimuli sentences (Sentence-Step 1 through Sentence-Step 10) created in MATERIALS were used as stimulus sentences.

Listeners

Two groups of native speakers of English were recruited for the experiment. One group consisted of two native speakers of English (1 male, 1 female), who also had participated in the identification task in Experiment 1 (ID-exp1&2). The listeners had the two experimental sessions, i.e. identification task in the isolation and in the sentence conditions, on the same day.

Another consisted of nine native speakers of English (5 males, 4 females) to participate in an AXB discrimination task (AXB-exp2). Eight of them were from the United States, and one of them was from United Kingdom. Ages ranged from 20 to 29 years (mean 21.9 years). They were exchange students at Sophia University, and have resided in Japan for 3 months to 11 months at the time of the experiment (mean 5.1 months). Eight of the

nine listeners had also participated in Experiment 1. For these listeners, experiments were held on different days, and they took sessions for Experiment 1 first.

None of the listeners reported any hearing problems at the time of the experiment.

Procedure

We conducted an identification task and a discrimination task. Listeners of ID-exp1&2 participated in the identification task and listeners of AXB-exp2 participated in the discrimination task. Stimuli presentations were made by using the Praat software [5]. All stimuli were presented diotically via Sennheiser HDA 200 headphones at participants' comfortable listening level.

Identification Task

Listeners of ID-exp1&2 were instructed to identify the target stimuli with surrounding contexts. During the experiment, the listeners were asked to choose if the presented syllables with surrounding contexts were "ra" or "la." They heard ten repetitions of stimulus sentences (10 stimulus sentences × 10 repetitions = 100 judgments). Stimuli were presented once, and listeners were not able to replay.

Discrimination Task

Listeners of AXB-exp2 were asked to discriminate syllables presented with sentence context. The stimulus sentences were paired such that each pair (AB) differed by two steps in the continuum, i.e. Sentence-Step 1－3, Sentence-Step 2－4, … , Sentence-Step 8－10. Two-step pair was preferred because we wanted to set the comparison condition equal to Experiment 1. In this task, the listeners were told that they would hear a man read a sentence *Clear **xx** is appreciated* three times, and that they were to judge if the second sentence sounded more similar to the first, or to the third. Listeners were informed that they would hear a syllable in the ***xx*** part. Paired stimuli that were 2-steps apart were presented as AAB, ABB, BAA, and BBA. There were three repetitions for each presentation, so listeners made 12 judgments for each pair. Thus, this makes total of 96 judgments (8 pairs × 4 presentations × 3 repetitions = 96 judgments). AXB presentations were made randomly. For purposes of familiarizing the listeners with the tasks, a short practice session was held prior to the experimental session.

Results

Identification Task

Results of the identification task are shown in Fig. 6. In identification of syllables in the sentence condition, the /ra/ responses dropped from 65%, which is above chance, at Step 5 to 30% at Step 6. On the other hand, the /la/ responses increased to 70% at Step 6. Thus, assuming that Step 5 and Step 6 belong to different categories, i.e. /ra/ and /la/, a discrimination peak, if there would be any, is predicted to appear at the pair at the pair 4－6 or/and the pair 5－7, which cross the categorical boundary.

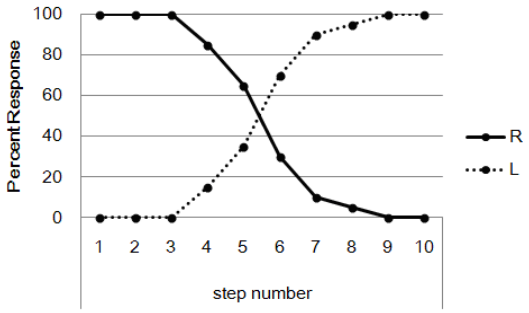

FIGURE 6. Averaged percent response (%) of /ra/ (solid line) and /la/ (dashed line) for identification of syllables presented in the sentence condition.

Discrimination Task

Results of the discrimination task are shown in Fig. 7 as a discrimination function. As indicated in the figure, any peaks were not observed in the discrimination function. If listeners took the advantage of the categorical boundary indicated in the identification function, discrimination accuracy should have been high at the pairs that crossed the categorical boundary, i.e. the pair 4－6, and the pair 5－7. However, it does not seem to be the case here. In fact, listeners' performance was consistently low, i.e. only slightly above chance. Thus, the categorical boundary was not reflected as any peaks in the discrimination function.

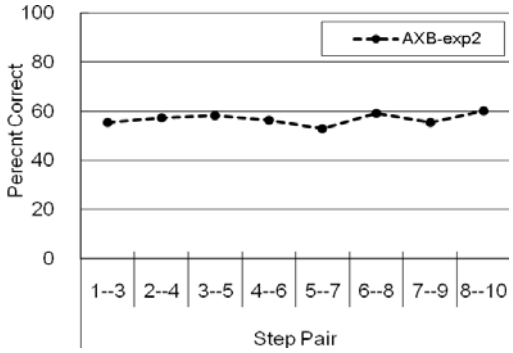

FIGURE 7. Averaged correct rate (%) for discrimination of each pair in the sentence condition.

Discussion of Experiment 2

Experiment 2 investigated whether or not a categorical perception is observed in perception of syllables presented within sentence context. In the current experiment, Step 5 and Step 6 were assumed belong to different categories based on the % response of /ra/ or /la/ at each step indicated from results of the identification task; thus, we predicted that a discrimination peak would appear at the pair 4－6 or/and at the pair 5－7. However, we did not obtain any discrimination peaks predicted from the identification function. Thus, it should be concluded that a categorical perception was not observed in perception of syllables in sentences. However, it is possible that the two-step comparison was not suitable for a discrimination task in the sentence condition for some reasons. How many steps listeners would need to get a discrimination peak in the sentence condition will be discussed in the future experiments.

DISCUSSION

The present report consisted of two experiments. In Experiment 1, we investigated perception of synthetic /ra/-/la/ syllables presented in isolation (isolation condition), and demonstrated that the perception in the isolation condition indicated a categorical perception. In Experiment 2, we next examined perception of the same syllables presented within a sentence (sentence condition). Results of the experiment revealed that the categorical perception was not evident in perception of syllables in the sentence condition. More concretely, native speakers of English seemed to be able to identify the /ra/-/la/ syllables in sentences, but they did not able to discriminate the difference even at the pair which was supposed to cross the categorical boundary.

In this section, we are interested to consider the reason(s) why listeners' discrimination accuracy dropped in the sentence condition. As one possibility, it can be related to contribution of vowel versus consonant information to sentence perception. In Kewely-Port *et al.* [6], it was demonstrated that vowel information is more important than consonant information for sentence intelligibility, as opposed to general consensus that consonants are more informative than vowels are for isolated words (or syllable) recognition [7]. The difference between the claims is assumed to come from linguistic processing of sentence versus words (or syllables) [6]. That is, it is assumed that sentence perception incorporates top-down processing, whereas word (or syllable) perception is more likely to rely on bottom-up information. We do not go deep into which kind of linguistic processing, i.e. top-down or bottom-up, is responsible for sentence or word (or syllable) recognition in this discussion. What we are interested in here is whether or not the discrimination accuracy drop attributes to little contribution of consonants to sentence

intelligibility. That is, we suspect that listeners' performance of discrimination was low because consonants are less important than vowels for sentence perception. In fact, Kewley-Port *et al.* [8] reported that discrimination accuracy of vowels presented in syllables did not greatly differ from that of vowels presented in sentences. Thus, it is reasonable to consider that listeners' performance depends on what kind of speech sounds, i.e. vowels or consonants, they are to discriminate. These issues will be dealt with in further experiments.

CONCLUSION

The present research reported that synthetic /ra/-/la/ syllables are perceived categorically (categorical perception) when they are presented in isolation (Experiment 1). However, such categorical perception was not evident in perception of the same syllables presented within a sentence. That is, native speakers of English were able to properly identify the syllables in the sentence condition, but the categorical boundary was not reflected as a peak in the discrimination function. We consider such results have something to do with little contribution of consonant information to sentence perception. The issue will be investigated in the further researches.

REFERENCES

1. 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. Psych. **54**, 358-368 (1957).
2. K. S. MacKain, C. T. Best and W. Strange, "Categorical perception of English /r/ and /l/ by Japanese bilinguals," Appl. Psycholing. **2**, 369-390 (1981).
3. 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).
4. V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," Sp. Comm. **9**, 351-356 (1990).
5. P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]", ver.5.3.23, http://www.praat.org/. (Last viewed 7 Aug. 2012.)
6. D. Kewley-Port, T. Z. Burkle and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 2365–2375 (2007).
7. M. J. Owren and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," J. Acoust. Soc. Am. **119**, 1727–1739 (2006).
8. D. Kewley-Port and Y. Zheng, " Vowel formant discrimination: towards more ordinary listening conditions," J. Acoust. Soc. Am. **106**, 2945–2958 (1999).