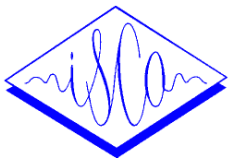

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese

Reiko A. Yamada, & Yoh'ichi Tohkura

ATR Auditory and Visual Perception Research Laboratories
Kyoto, 619-02, Japan

ABSTRACT

This study investigated how native speakers of Japanese, who are midway toward acquiring English phonemes, perceive and produce American English /r/, /l/, and /w/ sounds. From the perceptual experiments using synthesized stimuli, the following results are obtained. Japanese listeners perceive /r/ and /l/ sounds non-categorically, and they use both F2 and F3 frequencies as cues to identify /r/ and /l/. In contrast, American listeners perceive these sounds categorically using F3 as the primary cue. This result is consistent with the result obtained from the acoustic analysis of the /r/ and /l/ sounds produced by native Japanese. F3 frequency is a main feature for American subjects to produce /r/ differently from /l/. Such acoustic features, i.e. formant frequencies of the consonant part, are not well differentiated into phoneme categories when Japanese speakers produce these phonemes. However, the tendency to use both F2 and F3 frequencies, even as production cues, is observed in the sounds by Japanese speakers.

INTRODUCTION

The process of acquiring phonemes in a second language(L2) is much different from that in the first language(L1), because of the difference in age and the existence of the first language phonological system. When studying the process of acquiring phonemes in L1, it is seen that the ability to perform experimental tasks varies with age, and this causes a methodological problems. Although such methodological problems can be avoided when studying the process of acquiring phonemes in L2, many factors, such as age when learning L2, quantity and quality of learning, etc., affect the process. The modeling of acquiring phonemes in L2 can be realized by describing the effects of those factors on the process of language acquisition. For such purposes, the process, by which native speakers of Japanese (J. speakers) learn the phonemic distinction between English /r/ and /l/ sounds, provides a clue to solving the problem, because there are no sounds similar to English /r/ and /l/ in the Japanese phonological system, and most of the native speakers of Japanese have considerable difficulty in acquiring those sounds when learning English as a second language.

Previous studies on /r/ and /l/ perception by J. speakers revealed that J. speakers have considerable difficulty in discriminating these phonemes, and do not perceive a synthesized [r-l] continuum categorically. On the other hand, native speakers of American English (A.E. speakers) perceive the [r-l] continuum categorically (Liberman et al. 1973[4]; Miyawaki et al. 1973[7]; MacKain,K.S. et al. 1981[6]; Mochizuki 1981[8]). Acoustically, the distinctive feature that sets apart /r/ and /l/ sounds spoken by A.E. speakers is mainly the F3 frequency (Dalston,R. 1975[2]). This result is consistent with the results from perceptual experiments using A.E. speakers (Lisker,L.1957[5]; O'Connor,L.J. et al. 1957[10]), in which F3 frequency is an important perceptual cue when identifying /r/ and /l/. The authors (1989[11]) have reported that the main perceptual cue when identifing /r/ and /l/ sounds is the F3 frequency for A.E. speakers, and both the F3 and F2 frequencies for J. speakers. The production ability of /r/ and /l/ for J. speakers has been studied using two strategies: perceptual judgement by A.E. speakers (Goto,H.,1971[3]; Cochrane,R.M.,1980[1]) and acoustical analysis (Nakashima,H. et al. 1986[9]).

This paper discusses three points: (1) perceptual characteristics of /r/ and /l/ for J. speakers, (2) production features given by acoustic analysis of /r/ and /l/ sounds uttered by J. speakers, and (3) the relationship between perceptual cues and acoustic features of /r/ and /l/ sounds by J. speakers focusing on F2 and F3 frequencies.

Three types of perceptual experiments, i.e. identification tests and ABX discrimination tests of synthesized stimuli, and identification tests of naturally-spoken stimuli were assigned to subjects. After the perceptual experiments, the utterances of each subject were recorded.

1. Subjects

Five male A.E. speakers who were born and raised in the U.S.A. served as subjects in Group-A. Thirty-seven male J. speakers who have never resided abroad served as subjects in Group-J. Subjects in Group-J were senior high-school and university students whose ages varied from 15 to 23, and they had started learning English as L2 in junior high-school at about age 12. All subjects reported medical histories free of hearing or speaking disorder.

The subjects were the same all through the experiments, excluding two subjects in Group-A, one of whom participated only in the perceptual experiments, and the other only in producing the utterances.

2. Perceptual Experiment

2.1 Stimuli

We used a synthetic /rait-lait/ continuum generated by Klatt's cascade formant synthesizer (Klatt; 1980), and naturally spoken stimuli. Figure 1 provides a synthetic spectrographic representation of the initial

CV portion /rai-lai/ for the synthesized stimuli. The acoustic parameters for idealized "right" and "light" were derived from the naturally spoken /rait/ and /lait/ uttered by a native male speaker of American English. To construct the stimuli, three acoustic parameters, F2 and F3 onset frequencies and F1 transition, were varied. Two sets of stimuli, type P and type S, were used.

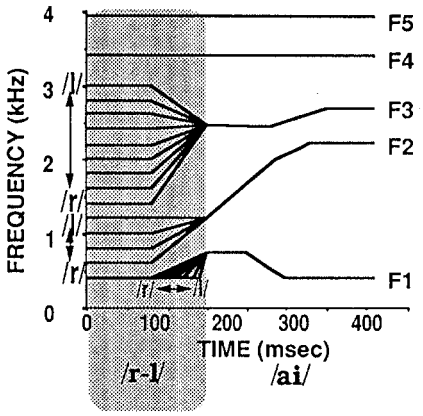

Fig.1 Schematic representation of spectrum for the [rai-lai] part of the stimuli.

In the type P set, a variety of F2 and F3 onset frequency combinations were used. The onset frequency of F2 was varied from 800Hz to 1400Hz in 200Hz steps, and that of F3 was varied from 1200Hz to 3000Hz in 200Hz steps. There were 37 combinations in total excluding some contradictory combinations in which the F2 frequency was equal to or higher than the F3 frequency. F1 transition duration was varied from 70ms to 16ms in 6ms steps as the F3 onset frequency was varied from 1200Hz to 3000Hz.

In the type S set, all three parameters, i.e. F2 and F3 onset frequencies and F1 transition duration, were varied dependently constructing a /rait-lait/ continuum, which consisted of 17 stimuli (St1 through St17). From St1 to St 17, the F2 and F3 onset frequencies varied from 960Hz to 1280Hz in 20Hz steps and 1400Hz to 3000Hz in 100Hz steps, respectively. F1 transition durations were varied from 61ms (St1) to 13ms (St17) in 3ms steps. As the F1 transition durations varied, the F1 steady state duration varied from 89 ms (ST1) to 137 ms (ST17) in 3 ms steps to keep constant the total duration from the onset to the end of the formant transition (i.e., 150 ms).

In all synthesized stimuli for both stimulus types, the acoustic parameters for the vowel part /ai/ were common, and the duration of the /rai-lai/ part was fixed at 360 ms. Formant frequencies of F1, F2 and F3 at the end of transitions (i.e., onset of the vowel part /ai/) were 750, 1220 and 2465 Hz, respectively. F4 and F5 were kept at constant values, 3400 and 3950 Hz, all through the stimuli.

Naturally spoken stimuli contained sixteen combinations of English words (List A in Appendix A). Each combination consisted of three words which were different from each otheronly in the initial consonant, i.e. /r/, /l/, or /w/. The forty-eight words (sixteen combinations each containing three words) were spoken by two native American English speakers (one female and one male) to produce a total of ninety-six stimuli. They were recorded and converted from analog to digital at a 20kHz sampling frequency with 16-bit accuracy.

2.2 Procedure

In the identification test of synthesized stimuli for set P, each of the 37 stimuli was presented three times in one experiment session. Two sessions were assigned each subject, with a short break between sessions.

In the identification test of synthesized stimuli for set S, 6 repetitions of each of the 17 stimuli were presented in one experiment session.

For identification tasks, listeners were asked to identify word initial consonants, and to make a forced choice among the given response categories even when it was difficult to choose. For the response categories, /w/ is added to /r/ and /l/ because Japanese listeners often perceive /w/ even when English-speaking listeners do not (Mochizuki, M., 1981[8]; Yamada,R., 1989[11]).

For ABX discrimination tests, the stimuli in set S were used. The 15 four-step comparison pairs (St1 vs St5, St2 vs St6, —, St15 vs St19) were arranged in triads in the four possible ABX permutations (i.e., ABA, ABB, BAA, and BAB) with an inter-stimulus interval (ISI) of 1 second. Two sessions were assigned to subjects, and each session consisted of a total of 52 trials (i.e., triads),

In the identification test of naturally-spoken stimuli, each of the 96 stimuli occurred one time without repetition to make one experiment session, and other conditions were common to the identification tests of synthesized stimuli.

All through the experiments, the stimuli were synthesized and reproduced with 16-bit accuracy at a sampling frequency of 20 kHz and low-pass filtering with a cutoff frequency of 10 kHz. For each experiment session, the stimuli were randomized and recorded on tape using a DAT recorder, SONY DTC-1000ES, with an inter-stimulus interval (ITI) of 2 seconds, and inter-block interval (IBI) of 6 seconds. A beep sound was recorded as a start signal 2 seconds prior to the beginning of each block of ten trials. These stimuli were presented to listeners binaurally over headphones, STAX SR Lambda Professional, in a soundproof room at a fixed level of about 85 dB SPL at peak intensity, which was a comfortable level for the listeners.

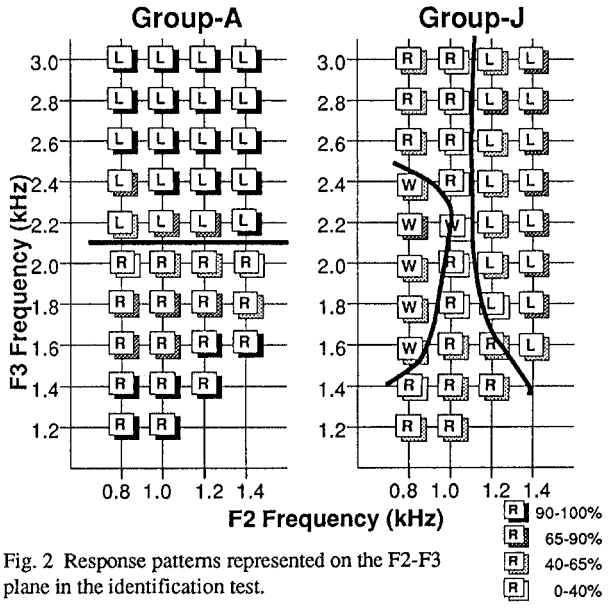

Fig. 2 Response patterns represented on the F2-F3 plane in the identification test.

2.3 Results

Identification patterns for the type P stimuli, averaged across subjects for Group-A and Group-J, are represented on the F2-F3 planes in Figure 2. The abscissa and ordinate represent F2 and F3 onset frequencies, respectively. The F2-F3 planes were divided into two or three regions corresponding to the three phoneme categories, whose boundaries are drawn based upon the most frequent response to each category. The degree of response concentration is also different between two groups, so the identification rate for the Type S stimuli differs between groups, with high consistency in Group-A and low consistency in Group-J. The peaks of the correct response rates in the ABX discrimination test are consistent with the identification rate in each group.

For Group-A subjects, the /r-l/ boundary is clear and is located at an F3 onset frequency of 2,000-2,200 Hz. Compared with Group-A's results, the responses are much less concentrated and the phoneme category boundaries are less clear in Group-J. The /r-l/ boundary seems to depend upon both F2 and F3. In addition, /w/ responses by Goup-J subjects are observed more frequently than those of the Group-A subjects.

In identification tests of naturally-spoken stimuli, the averaged correct response rates were 100% in Group-A, and 64.9%(varied from 50.0% to 76.4%) in Group-J.

3. Production Experiment

3.1 Speech Samples

The 16 combinations of words and 10 combinations of CV syllables (see Appendix A) were used as speech material. Each combination consisted of three words or syllables which differed only in the initial consonant, /r/, /l/, or /w/. Each combination was uttered once, excluding one combination, "right" "light", and "wite", which was repeated three times. 29 triads (87 samples) were recorded per subject. As the J. subjects had some difficulty in reading the lists, the material was presented in the order shown in List A in Appendix A. Furthermore, the guide speech spoken by the bilingual female speaker were presented through headphones. These guide speech stimuli are intended not only as a reading aid but also to keep the F0 pattern for each utterance consistent.

Subjects were asked to pronounce the words and syllables on the lists in order by repeating each word they had heard. They were also instructed that they need not imitate the guide speech, which was provided only to help identifying the material.

Recordings were made in an anechoic room, and speech samples were recorded on tape using a microphone, SONY C350, and a DAT recorder, SONY DTC-1000ES.

The recorded speech samples were digitized at a 20kHz sampling frequency with 16-bit amplitude accuracy.

3.2 Acoustic Analysis

After the 20kHz speech samples sampling frequency were down-sampled to 12kHz, the formant frequency trajectories were estimated by an LPC formant tracking method with the parameters shown in Table 1. The errors in formant estimation were corrected by visual inspection by referring to the soundspectrogram and bandwidth of each estimated peak.

Table 1
The parameters for LPC analysis.

ORDER OF LPC	16
FRAME PERIOD	2.5 ms
DFT POINTS	1024
WINDOW(HANNING)	30 ms
PRE-EMPHASIS FACTOR	0.98

The F2 and F3 frequencies were measured about 30ms after speech signal onset as the steady-state frequency of F2 and F3 for the initial consonant of each speech sample.

Table 2
Mean frequency values in Hz of F2 and F3 for Group A-subjects(native speakers of A.E.) and Group J-subjects (native speakers of J.).

	A.subjects		J.subjects	
	F2	F3	F2	F3
/r/	984	1,590	1,287	1,878
/l/	1,193	2,602	1,535	2,233
/w/	721	2,331	915	2,302

3.3 Result

From the visual inspection of the soundspectrograms, the acoustic distinctive features of the three phonemes, /r/, /l/ and /w/, seemed mainly the F2 and F3 frequencies for Group-A subjects. On the other hand, the acoustic distinctive features of these consonants for J. subjects were in addition to F2 and F3 frequencies. A release burst line is sometimes observed for the production of the /l/ sound by J. subjects, indicating that they are not classified into liquid but rather into the flaps or plosives. An extraneous formant between F2 and F3 for the initial consonant part was also observed for the production by a few J. subjects. Data with such release burst lines or extraneous formants were neglected when measuring the F2 and F3 frequencies. With respect to F2 and F3 frequencies, the production by J. subjects had the following characteristics when compared to Group-A's production. (1) Although the data are well differentiated with respect to the phoneme categories, /r/, /l/, and /w/ for Group-A subjects, the categories are not well differentiated for Group-J subjects. (2) Individual differences are small in Group-A, but much larger in Group-J. (3) As shown in Table 2, the mean values of both F2 and F3 frequencies for the consonant part are higher in /l/ than in /r/ in both groups. However the difference between /r/ and /l/ is 1.25z in F2 and 3.16z in F3 on the bark scale. This result suggests that the F3 frequency is the primary acoustic feature used to differentiate /r/ and /l/ sounds by Group-A. On the other hand, in Group-J, the differences between the mean frequency of /r/ and that of /l/, 1.28z in F2 and 1.12z in F3 on the bark scale, are small. Data from the /r/ and /l/ categories widely distribute much along the F3 axis, indicating that the F2 frequency is the primary acoustic feature used to differentiate /r/ and /l/. (4) In Group-A, the low frequency of F2 is the main feature used to differentiate /w/, but the F2 frequency of /w/ is higher than that in Group-J, even though it is more differentiated from other two categories. (5) The productions of the J. subjects with high correct response rates in the test of identifying naturally spoken stimuli are differentiated more than those with lower rates. However, even for the J. subjects with higher rates, perception and production are classified into so-called Japanese type, i.e. using F2 as a perceptual cue and as an acoustic distinctive feature in production, and with a wide /w/ area both in perception and production.

represents three typical cases for three subjects: subject (A17) in Group-A, subject (J197) in Group-J whose data are relatively well differentiated with respect to three phoneme categories, and subject (J182) whose data are not well differentiated. Even though there are the J. subjects with higher production ability, like J197 in figure 3, they are still in the process of acquiring those phonemes. The target direction, in which /r/ and /l/ are differentiated along the F3 axis, seems to be absent and /r/ and /l/ are differentiated along not only the F3 axis but also along the F2 axis.

4. Discussion

Results of two experiments are regarding the following points. (1) Distributions of both perceptual responses and acoustic features are much wider in J. speakers than in A.E. speakers. (2) Individual differences in both perception and production are larger in J. speakers than in A.E. speakers. (3) Although the F3 frequency is the primary cue used for differentiating /r/ and /l/ in A.E. speakers, the F2 frequency is also an important cue for J. speakers. (4) The area of the /w/ category on the F2-F3 plane expands more for the high F2 area in J. speakers than in A.E. speakers.

The first two points depend on the failure or difficulty in acquiring those sounds. The third point shows the possibility that not only are the J. subjects still in the process of acquiring /r/ and /l/, but also that the characteristics of the categories they are going to acquire might shift from the real characteristics of the categories that native English speakers have. Two possibilitis that cause this shift are considered. The first is that the method of teaching those phonemes to J. speakers has been inadequate or incorrect. In fact, J. speakers are sometimes taught that /r/ is heard much more like /w/ than /l/, and that for producing /r/ sounds, lips shoud be rounded as when producing the /w/ sound. The other possibility is that the strategy of using F2 when distinguishing /r/ and /l/ in both perception and production is easy for J. speakers. It is interesting to compare the process of acquiring /r/ and /l/ categories for J. speakers who are learning English in an English speaking environment to the present results. The fourth point might be caused by the exsistence of the /w/ category which is close to the other two in the Japanese phonological system.

From the results discussed here, it is obvious that, when acquiring phonemes in L2, perception and production have a strong relationship. In the future, we are going to investigate further the process of acquiring phonemes in L2 through training experiments: how sound articulation training affecst perception, and how listening training affects production.

References

[1]Cochrane, R.M. "The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language," Journal of Multilingual and multicultural development, 1, 331-360, 1980.

[2]Dalston, R.M. "Acoustic characteristics of English /w,r,l/ spoken correctly by young children and adults," J.Acoust. Soc. Am., 57, 462-469, 1975.

[3]Goto, H. "Auditory perception by normal Japanese of the sounds L and R," Neuropsychologia 9, 317-323, 1971.

[4]Liberman,AmM., Miyawaki,K., Jenkins,J.J., and Fujimura,O. "Cross-language study of the perception of the F3 cue for [r] vs. [l] in speech- and nonspeech- like patterens," J.Acoust.Soc.Jpn., 29, 315, 1973.

[5]Lisker,L. "Minimal cues for separating /w,r,l,y/ in intervocalic position," Word 13, 256-267, 1957.

[6]MacKain, K.S., Best,C.T., and Strange,W."Categorical perception of English /r/ and l/ by Japanese bilinguals," Appl. Psycholinguist., 2, 369-390, 1981.

[7]Miyawaki, K., Strange,W., Verbrugge,R., Liberman,A.M., Fujimura,O., and Jenkins,J. "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," Percept. Psychophy., 18, 331-340, 1975.

[8]Mochizuki,.M. "The identification of /r/ and /l/ in natural and synthesized speech," J.Phonet., 9, 283-303, 1981.

[9]Nakashima, H., Kobayashi,T., and Kakusho,O. "Perception of /r/ and / l/ sounds for Japanese subjects," Trans. Com. Elect. Speech Res. SP86-56, 39-44, 1986(in Japanese).

[10]O'Connor, J.D., Gerstman,L.J., Liberman,M.A., Delattre,P.C., and Cooper,F.S. "Acoustic cues for the perception of initial /w,j,r,l/ in English," Word 13, 25-43, 1957.

[11]Yamada,R., Tohkura,Y., and Kobayashi, N. "Perceptual characteristics of English syllable-initial /r,l/ for Japanese listeners," J.Acoust.Soc.Am. 86, Suppl.1, S102, 1989.

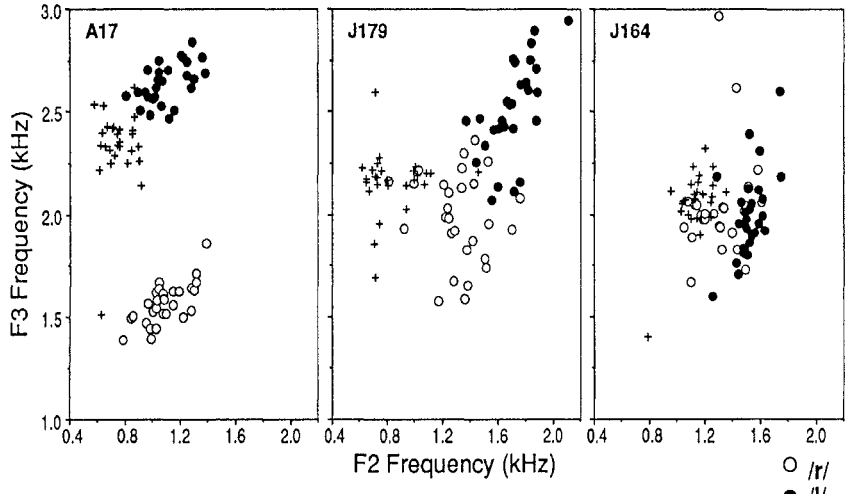

Fig. 3 Examples of formant frequency scatter on the F2-F3 plane. The three graphs represents the data from subject A17 (native speaker of A.E.), and one native speaker of J. whose data are well differentiated among Group-J (middle), and one native speaker of J. whose data are not differentiated. Circles represents F2 and F3 of a word or syllable initial /r/ sound in speech samples, dots represents /l/, and plus signs represent /w/.

Apeendix A

Speech materials used in the experiment.

List A		
read	lead	weed
rip	lip	wip
red	led	wed
rack	lack	wack
ra	la	wa
rock	lock	wock
raw	law	waw
rook	look	wook
root	loot	woot
rush	lush	wush
rate	late	wate
right	light	wite
royal	loyal	woyal
row	low	woe
rout	lout	wout
rear	lear	weer
List B		
ri	li	wi
re	le	we
ra	la	wa
ro	lo	wo
ru	lu	wu
reigh	leigh	weigh
righ	ligh	wigh
roy	loy	woy
row	low	woe
rou	lou	wow