On Why Japanese /r/ Sounds are Difficult for Children to Acquire

Takayuki Arai

Department of Information and Communication Sciences
Sophia University, Tokyo, Japan
arai@sophia.ac.jp

Abstract

Many studies have pointed out that the /r/ sounds in Japanese tend to be difficult for native children of Japanese to acquire. To verify this, we first investigated Japanese /r/ sounds uttered by two-year-old twins as a case study. The acoustic analysis of the recordings, which included several words with various /r/ sounds, revealed that certain /r/ sounds are difficult to produce and are often produced with speech errors. We also analyzed a set of utterances of Japanese /r/ spoken in a variety of phones pronounced by an adult male speaker. Then, for comparison, we synthesized Japanese /r/ sounds using four parameters. We conducted two perceptual experiments: one for the natural speech by the male speaker of Japanese, and another for the synthesized speech sounds based on the four parameters. The results showed that variation in pronunciation in adults was widely distributed. We discussed the reasons that it takes time for children to acquire /r/ sounds, and we concluded that it is possibly due to the combination of two factors: 1) some /r/ sounds themselves are difficult to produce, and 2) there is a wide distribution of pronunciation variation in adult speakers.

Index Terms: Japanese /r/ sounds, flap sounds, children's speech, allophones, speech production

1. Introduction

Japanese /r/ is often categorized as a flap [1-4], but actually, there are several phonetic (allophonic) variations of the single phoneme /r/ [1-6]. While intervocalic /r/ in Japanese is typically a flap, phrase-initial /r/ is frequently pronounced as a plosive [1-6]. Furthermore, retroflex and lateral approximants are not unusual variants of Japanese /r/ [2-6].

It is reported that acquisition of /r/ is delayed in native Japanese children, as compared with other consonants [7]. This is true for second language learners as well [8, 9]. In addition, elderly listeners of Japanese often misperceive /r/ sounds [10]. In a series of out studies, we first investigated Japanese /r/ uttered by two-year-old twins as a case study. The recordings were done in our pilot study [11] and included several words with various forms of /r/. In this pilot study, we observed variations including speech errors as listed below:

a) **Replacement with a plosive**: /onri/ -> /ongi/; /ringo/ -> /dingo/ or /gingo/. This happened commonly in word-initial position or following a syllable nasal [12].
b) **Deletion**: /beruto/ -> /beuto/; /arigato:/ -> /aigato:/.
c) **Palatalization**: /osora/ -> /osorya/, /kuru/ -> /kuryu/.
d) **Replacement with a glide**: /karappo/ -> /kayappo/ or /kawappo/; /senro/ -> /sen-yo/.
e) **Lateralization**: /oshiri/ -> /oshili/; /kuru/ -> /kulu/ or /kulju/.
f) **Lateral articulation**: /karikari/ and /oshiri/.

In the present study, we also analyzed the data acoustically. Additionally, for comparison, we analyzed a set of utterances of /r/ spoken by a native male speaker of Japanese. We then modeled /r/ sounds using four parameters and synthesized them from those parameters. A perceptual experiment was conducted to identify the important acoustic cues of Japanese /r/. Finally, we discussed possibilities for why it is difficult to acquire Japanese /r/.

2. Acoustic analysis

2.1. Children's speech

To observe children's speech errors pertaining to Japanese /r/,

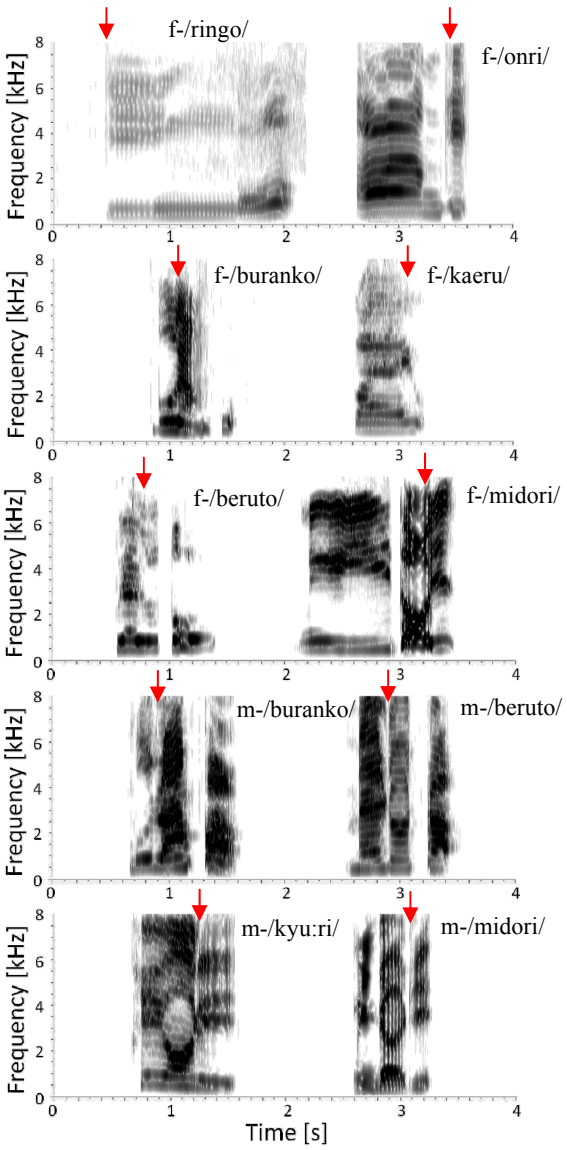

Figure 1: *Spectrograms of the utterances by children.*

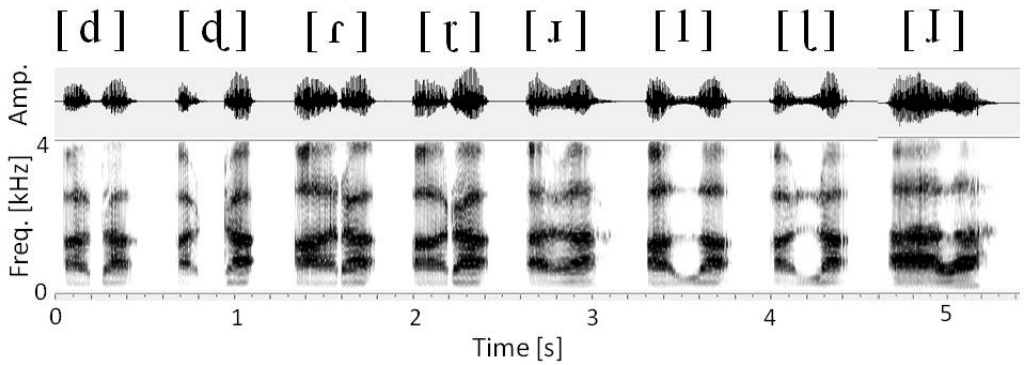

Figure 2: Waveforms and spectrograms of the eight consonants uttered in the /aCa/ context by an adult male.

we recorded several words with various /r/ sounds uttered by two-year-old twins as a case study [11]. In the recordings, a digital recorder (SONY, PCM-D1) with internal microphones were used. The original sampling frequency of 48 kHz was converted into 16 kHz for acoustic analysis.

Figure 1 shows sound spectrograms of several words uttered by the twins: m (a boy) and f (a girl). The first six utterances were uttered by a girl, and the rest four utterances were uttered by a boy. Examining the spectrograms we note that the children produced not only the expected flap /r/, but also plosives, approximants and other variants. We also noted that /r/ is often deleted. An arrow for each utterance indicates the location of /r/. The followings are the detailed observations:

f-/ringo/: The word-initial /r/ becomes a /g/-like plosive.
f-/onri/: The /r/ preceded by a syllable nasal becomes a /g/-like plosive.
f-/buranko/: The intervocalic /r/ becomes a glide [j].
f-/kaeru/: The intervocalic /r/ becomes a glide [j].
f-/beruto/: The intervocalic /r/ is deleted.
f-/midori/: The intervocalic /r/ is deleted.
m-/buranko/: The /r/ is pronounced as a flap.
m-/beruto/: The intervocalic /r/ is pronounced as a plosive-like sound.
m-/kyu:ri/: The /r/ is pronounced as a lateral approximant.
m-/midori/: The /r/ becomes so-called "lateral articulation."

From this figure and from speech errors obtained in the previous pilot study [11], it seems that some phones are difficult to pronounce while others are easy. For example, /r/ followed by /i/ is difficult to pronounce and often becomes a plosive or lateral (including "lateral articulation" [13]). This might be due to the height of /i/. In other words, the tongue tends to shift to the high position of /i/, which leaves only a small area for flapping the tongue. On the other hand, the alveolar lateral approximant [l] is relatively easy to produce.

2.2. Adult speech

For comparison, we also analyzed and modeled a set of utterances spoken by a male speaker. The original recordings were done in a VCV context with regular phonation as well as with an electrolarynx [11]. V is one of the five Japanese vowels and C is one of the following consonants:

d (alveolar plosive), **ɖ** (retroflex plosive), **ɾ** (alveolar flap), **ɽ** (retroflex flap), **ɹ** (alveolar approximant), **l** (alveolar lateral approximant), **ɭ** (retroflex lateral approximant), and **ɺ** (alveolar lateral flap).

In the recordings, a digital recorder (SONY, PCM-D1) with internal microphones was used. The original sampling frequency of 48 kHz was converted to 8 kHz for acoustic analysis.

Figure 2 shows sound spectrograms of the /aCa/ word set. For results, the following points were observed.

For [d] and [ɖ]:
- Before the burst, there are closure periods of varying lengths.
- Downward shift of the first formant (F1) frequency and upward shift of the second formant (F2) frequency towards C were observed.
- Downward shift of the third formant (F3) frequency towards C was observed for [ɖ].
- The sound [ɖ] is often observed in word-initial position, while the sound [d] is sometimes observed as a speech error.

For [ɾ] and [ɽ]:
- These sounds are typically observed inter-vocalically.
- The belief discontinuities, that reflect the short closure of the oral cavity were observed. The duration of these closures were approximately 20-30 ms.
- Downward F1 shift and upward F2 shift in frequency were also observed.
- Downward shift of F3 towards C was observed for [ɽ].

For [ɹ]:
- Because the tongue tip does not make any contact with the palate, there was no closure period, and the sound levels gradually varied in time.
- Because the tongue tip moves towards the alveolar ridge, downward F1 shift and upward F2 shift in frequency were also observed.
- There is a drop in F3 frequency during this consonant.
- The movement of the tongue can be sustained at its maximal height in English. When this sound appears in Japanese, however, the tongue tip returns immediately after it reaches the maximal point.

For [l], [ɭ], [ɺ]:
- Because the tongue tip touches the alveolar ridge along the midline and the air stream flows through the lateral regions of the tongue, there was no closure period, and the sound levels do not drop very much in time.
- For [l] and [ɭ], tongue contact with the palate can be sustained, and therefore, the durations of these consonants in this figure are relatively longer. On the other hand, because [ɺ] is articulated as the short version of the alveolar lateral approximant [l], the duration of this consonant is as short as other Japanese /r/ sounds.

Table 1: *Averaged scores for the likeliness of /r/ sounds (in %) for the /Ca/ and /aCa/ utterances uttered by an adult male speaker.*

	Consonant							
	[d]	[ɖ]	[ɾ]	[ɽ]	[ɹ]	[l]	[ɭ]	[ɻ]
/Ca/	60	82	86	83	21	81	91	38
/aCa/	44	43	93	96	26	62	75	35

- When the tongue contact is released, the downward tongue movement is rather fast, so rapid formant transitions are also observed in the figure.
- The F2 and F3 frequencies of [l] are almost constant during the whole utterance, while the F2 frequency of [ɭ] is high during this consonant.
- A pole was observed at a low frequency due to the two-branch acoustic tube for these consonants.
- One of the acoustic differences between [ɻ] and [ɾ] is the F2 transition: flat vs. upwards movement in frequency during the consonant.

In summary, we can conclude that Japanese /r/ sounds have the following acoustic characteristics:
- There are formant shifts in frequency. This is due to the tongue movement for the consonant. The F1 frequency shifts downward towards the consonant, whereas the F2 frequency shifts depending on the place of articulation.
- There is an amplitude drop during the consonant. This is due to either the tongue contact or tongue movement towards the palate.
- The duration of the consonant is rather short.

Based on these observations, we selected four acoustic parameters for modeling Japanese /r/ as shown in Fig. 3:
- **AVd**: the depth of the amplitude gap.
- **AVg**: the gap duration of the amplitude of voicing.
- **Fg**: the gap duration of the formant transitions.
- **F2p**: the peak of the F2 frequency.

3. Perceptual experiments

3.1. Natural speech by an adult male speaker

We conducted the first perceptual experiment based on the /aCa/ utterances recorded in Section 2.2. The first stimulus set was the original eight /aCa/ utterances. We also prepared a second stimulus set, which was the original eight /Ca/ utterances with the first vowel artificially removed. The experiment was conducted in a sound-treated room. Stimuli were presented monaurally through a loudspeaker (NAE NESmini) connected to an audio interface (RME Babyface) via an amplifier (FOSTEX AP1020). The five participants were seated 3-4 m from the loudspeaker. The sound level was approximately 70 dBA on average. There was a training session with eight stimuli prior to the main session. Twenty young listeners with normal-hearing (10 males and 10 females, ages 20 to 29 years) participated in the experiment. All were native speakers of Japanese and they were divided into four listener groups. Five participants from the listener group took part in the experiment simultaneously.

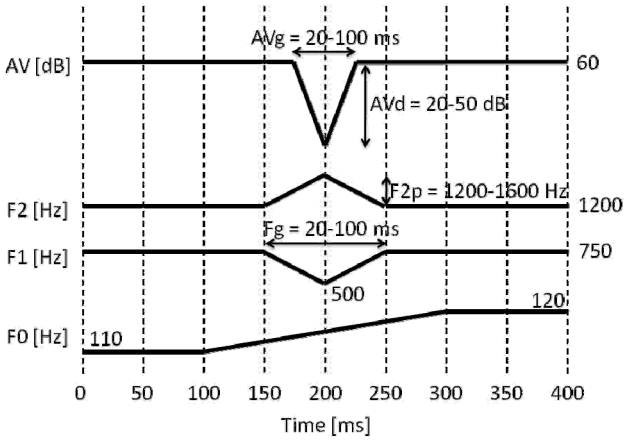

Figure 3: *A schematic plot of the four variables for synthesizing the stimuli: AVd from 20 through 50 dB in 10 dB step, AVg and Fg from 20 through 100 ms in 20 ms step under the condition of Fg >= AVg, and F2p from 1200 through 1600 Hz in 200 Hz step.*

In the main session, the eight utterances were repeatedly used three times in random order. There were 24 trials in total. A stimulus was presented in each trial, and the listeners were instructed to select one answer for the question displayed on a computer screen by means of a graphical user interface (GUI). The question was as follows:

How well suited is this sound to Japanese /r/:
100%, 75%, 50%, 25%, or 0%?

A similar experiment was repeated twice: one for the /Ca/ stimulus set and the other for the /aCa/ stimulus set. Table 1 shows the average scores by percentage as results for these stimulus sets. This table shows that there are definitely acceptable allophones for Japanese /r/ beyond the alveolar flap. The sound [ɖ] got a relatively low score in the /aCa/ context, but it did not have a low score in the /Ca/ context. This suggests that the sound [ɖ] is accepted as Japanese /r/ in word initial, but not word medial, position.

3.2. Synthesized speech

We conducted the second perceptual experiment based on the /aCa/ utterances synthesized from the four acoustic parameters described in Section 2.2. The stimuli were all synthesized by the XKL, which is a revision of the software package developed by Klatt [14]. Figure 3 shows a schematic plot of the variables for synthesizing the stimuli. AVd varied from 20 through 50 dB in 10 dB steps. AVg and Fg varied from 20 through 100 ms in 20 ms steps. The data that did not meet the following criterion were excluded: Fg >= AVg. This is because we assumed the closure (the amplitude gap) always occurs when the tongue is moving (formant transitions). Finally, F2p varied from 1200 through 1600 Hz in 200 Hz steps. "The no AV gap condition" was also included. The total duration of the stimuli were all 400 ms, and the default values were used for the rest of the parameters for synthesizing the stimuli. The experimental settings and listeners were exactly the same as in the first experiment.

Table 2: Averaged scores of the likeliness of /r/ sounds for the 195 synthesized / aCa/ utterances.

		1.2	1.4	1.6	1.2	1.4	1.6	1.2	1.4	1.6	1.2	1.4	1.6	1.2	1.4	1.6	← F2p [kHz]
		20	20	20	40	40	40	60	60	60	80	80	80	100	100	100	← Fg [ms]
20	20	34	40	35	48	51	55	53	73	70	46	65	76	54	74	81	
30	20	45	40	24	55	58	53	48	74	81	51	76	70	51	69	83	
40	20	43	46	36	60	56	64	55	73	64	59	71	71	59	73	64	
50	20	41	34	40	70	68	56	61	68	74	63	70	74	56	73	59	
20	40				29	33	34	41	45	44	43	48	59	34	68	73	
30	40				20	36	43	36	48	46	34	50	54	46	63	66	
40	40				40	33	48	33	59	45	43	58	55	40	55	63	
50	40				46	43	50	54	58	59	58	65	66	39	44	48	
20	60							14	23	16	20	25	35	18	35	33	
30	60							15	24	24	25	25	34	18	36	26	
40	60							16	14	23	18	29	28	16	38	28	
50	60							45	36	41	53	23	51	33	28	29	
20	80										14	5	14	11	16	24	
30	80										14	11	14	13	15	18	
40	80										14	13	16	18	18	14	
50	80										29	34	23	20	28	28	
20	100													6	15	19	
30	100													9	11	11	
40	100													10	14	15	
50	100													23	29	34	
60	—	48	43	43	66	65	76	64	64	73	61	81	68	48	71	49	
↑	↑																
Avd	Avg																
[dB]	[ms]																

In the main session, the 195 stimuli in total were randomly ordered. A stimulus was presented in each trial, and the listeners were instructed to select one answer for the question with the same GUI. The question was again as follows:

How well suited is this sound to Japanese /r/:
100%, 75%, 50%, 25%, or 0%?

Table 2 shows the experimental results for the 195 stimuli. Each number of this table is the averaged score of the likeliness (in percentage) of /r/ sounds for each stimulus among 20 listeners. The shaded numbers in this table refer to when the averaged value is greater than 75%. From this table, a certain set of stimuli got high scores for Japanese /r/ sounds.

4. Discussions and conclusion

In the present study, we focused on Japanese /r/, which is difficult for native children to acquire. There were two possibilities: 1) the /r/ sounds themselves are difficult to produce, and 2) there are several acceptable pronunciation variations of /r/ in adult speech. The analysis of the data by two-year-old twins supported the first point above. For the second point, we systematically analyzed the pronunciation variations for the speech samples uttered by an adult male speaker, then modeled the data, and conducted a perceptual experiment to find the acoustic cues for Japanese /r/ sounds. The results revealed the following:

- A short amplitude gap, such as 20-30 ms, is recognized as Japanese /r/. This gap should occur during formant transitions with a duration of 60-100 ms and an F2 peak of 1400-1600 Hz.
- Fast formant transitions, those that are 40 ms long, with no amplitude gap, are also recognized as Japanese /r/.

In the perceptual experiment, we looked only at the vowel /a/ as the context surrounding the target consonant /r/. However, other vowels should also be tested in the future.

Thus, adults utter several varieties of /r/ in Japanese. When young children hear such variations, since they are still developing their phonological system, they do not perfectly know which sounds map with which phoneme. In addition, children might misperceive each the sounds they hear. Finally, the difficulty of producing /r/ sounds in Japanese adds extra difficulties for acquisition.

These results lead one to consider the following application in speech pathology: When a child with articulation disorders mispronounces /r/ in Japanese, a speech pathologist might be able to train him/her by starting with an easy sound to produce, such as, [l], and then progressively shifting the articulation towards some of the other more difficult to acquire allophones of /r/.

5. Acknowledgements

I acknowledge the anonymous reviewers for their helpful comments. This work was partially supported by JSPS KAKENHI Grant Number 24501063. I would also like to thank Keiichi Yasu and Terri Lander for their supports.

6. References

[1] Hattori, S., *Onsei-gaku*, Iwanami, 1984.

[2] Kawakami, S., *Nihongo Onsei Gaisetsu*, Ohfu, 1977.

[3] Vance, T. J., *An Introduction to Japanese Phonology*, State University of New York Press, 1987.

[4] Saito, Y., *Nihongo Onsei-gaku Nyumon*, Sanseido, 1997.

[5] Arai, T., "A case study of spontaneous speech in Japanese," *Proc. of the International Congress of Phonetic Sciences (ICPhS)*, 1, 615-618, 1999.

[6] Arai, T., Warner, N. and Greenberg, S., "Analysis of spontaneous Japanese in a multi-language telephone-speech corpus," *Acoustical Science and Technology*, 28(1), 46-48, 2007.

[7] Funayama, M., Abe, M., Kato, M., "Kouon-kensa-hou ni kansuru tsuika-houkoku," *The Japan Journal of Logopedics and Phoniatrics*, 30, 285-292, 1989.

[8] Takubo, Y., *Onsei*, Iwanami, 1998.

[9] Kubozono, H., *Nihongo no Onsei*, Iwanami, 1999.

[10] Kobayashi, T., "Roujin-sei nanchou ni kansuru kenkyuu," *Nippon Jibiinkouka Gakkai Kaiho*, 61, 157-212, 1958.

[11] Arai, T., "Acoustic characteristics of Japanese /r/ sounds and errors in children's speech," *Proc. Spring Meet. Acoust. Soc. Jpn.*, 349-352, 2013.

[12] Ueda, I. and Davis, S., "Promotion and demotion of phonological constraints in the acquisition of the Japanese liquid," *Clinical Linguistics & Phonetics*, 15, 29-33, 2001.

[13] Kato, M., Okazaki, K., Suzuki, N. and Yamashita, Y., "Five cases with lateral articulation, " *The Japan Journal of Logopedics and Phoniatrics*, 22, 293-303, 1981.

[14] Klatt, D. H., "The new MIT speech VAX computer facility," *Speech Communication Group Working Papers IV*, Research Laboratory of Electronics, MIT, Cambridge, 73-82, 1984.