International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-10 Issue-2, July 2021

Automatic Geminate Insertion Algorithm for Japanese Audio Data

Hirofumi Maeda, Kenta Yamamoto

Abstract*: Generally, it is quite difficult for Japanese language learners to acquire Japanese special morae, namely, geminate, syllabic nasals and long vowels compared to independent morae. Among these three special morae, geminate is particularly difficult, and it takes much longer to fully acquire both production and perception of it. Especially for learners of Chinese native speakers, previous studies has shown that both production and perception of geminate are difficult in terms of the fact that not only no geminate is found in Chinese language, but also the phonological interaction between Japanese accent and Chinese tones. However, in the field of Japanese speech acquisition, research has not making progress because of a major problem, that is, researchers themselves manually create the acoustic experiment stimuli. Therefore, in this study, as a method to solve this problem, we propose an algorithm that automatically inserts geminate into the audio data used in Japanese speech acquisition research. This algorithm automates the insertion of geminate by performing three processes in order: mora extraction by noise removal, matching of original audio data and extracted mora, and insertion of soundless duration and geminate. The algorithm makes it possible to remove the noise, which is -50 dBFS and continues for 10ms or more, and replace it with soundless duration instead, allowing Japanese native speakers to percept it as geminate. The accuracy was equivalent as a result of comparing the data that was manually modified by a phonology researcher with the data that was generated by the algorithm. The result shows that the algorithm can be a practical solution for the automation of geminate insertion.

***Keywords*: Acoustic processing, Japanese language learning, geminate, noise reduction, algorithm**

I. INTRODUCTION

Generally, it is quite difficult for Japanese language learners to acquire Japanese special morae, namely, geminate, syllabic nasals and long vowels compared to independent morae [1]-[11]. Among these three special morae, geminate is particularly difficult, and it takes much longer to fully acquire both production and perception of it [12].

Manuscript received on July 17, 2021.
Revised Manuscript received on July 21, 2021.
Manuscript published on 30 July, 2021.
* Correspondence Author
Hirofumi Maeda*, Department of Information Science and Technology, National Institute of Technology (KOSEN), Yuge College, Ehime Prefecture, Japan. Email: maeda@info.yuge.ac.jp
Kenta Yamamoto, Department of General Education, National Institute of Technology (KOSEN), Yuge College, Ehime Prefecture, Japan. Email: yamamoto@gen.yuge.ac.jp

In the field of Japanese speech acquisition research, research has been carried out focusing on this difficulty for the learners. Kurihara conducted experiments on the perception and production of 30 native speakers of the northern Chinese dialect, with the aim of exploring the relationship between perception and production, using

Japanese long and short sounds as speech stimuli. In the experiment, two sets of minimal pairs were used, which can be distinguished only by the length of the vowels at the end of the words “Kato (transient)” and “Katoo (Katoo, a Japanese common surname)”, “Toko (floor)” and “Tokoo (travel).” Each was presented to the learner with a carrier sentence, “Read this as.” For long vowels, 14 levels of acoustically manipulated audio data were prepared. The learners were presented 15 different versions of a pseudo-word, each differs in terms of the length of the vowel, consisting of 14 long vowels and 1 short vowel in order from long vowels to short vowels and from short vowels to long vowels, and were measured their timing of the perceptual boundary. From the measurement results, it was reported that there was no difference in the perceptual boundaries of the long and short vowels at the end of the word, regardless of the learner’s Japanese proficiency [13]. On the other hand, Toda, who conducted the same experiment with English native speakers, reported that perceptual boundary becomes stable as their Japanese ability increases [14]. Considering that the stimulus words of these two studies were 3 morae, there is a possibility that the number of syllables or the position of special morae may affect categorical perception, which are not examined in these two studies. Furthermore, as the number of syllables increases, so does the accent patterns. In addition, many of the studies on the categorical perception for Japanese learners whose native language is Chinese are aimed at long vowels, and not many are focused on geminate.

Against this background, Yamamoto experimentally and exploratorily investigated how the difference in accent patterns and geminate positions affect the categorical perception. The result indicated that both accent patterns and the difference in the geminate position affect the categorical perception [15]. In addition to this result, although these speech acquisition research have great meaning and influence on improving the research efficiency in Japanese language acquisition, it was also found that the researcher need to modify the stimulus words data manually, which takes huge amount of time, resulting in the fact that the research in this field is not making so much progress. Furthermore, not only the number of sentences to be presented, but also the number of stimulus words to be prepared increases explosively given the combinations such as accent positions and intonation patterns.

Therefore, it is quite easy to imagine that the time required to create stimulus words data by manual operation directly leads to the stagnation of research.

This study focuses on developing an algorithm that automatically modifies the speech data and makes stimulus words without manual operation, aiming to reduce the workload of researchers in the field of speech acquisition research in Japanese language education. Also, the stimulus data that was manually modified mentioned in Yamamoto's study above is used in order to compare and verify the accuracy of the stimulus words data that was generated by the algorithm.

II. THE EFFECT OF ACCENT PATTERNS AND GEMINATE POSITIONS ON CATEGORICAL PERCEPTION

This chapter describes i) the outline of the study conducted by Yamamoto and ii) the stimulus data presented in the study. Yamamoto conducted two tasks with Japanese language learners of Chinese native speakers. First, accent identification task was performed, followed by categorical perception task. Yamamoto made a hypothesis that stimulus words with good accent pattern identification results would be more perceptually categorized, and conducted the accent pattern identification task as a preliminary experiment for the categorical perception experiment. The details of each experiment are shown below.

A. Accent Pattern Identification Task

There were a total of 23 intermediate and advanced Japanese learners whose mother language was Chinese, and all of the Japanese learners passed the Japanese Language Proficiency Test N1. Also, the participants were asked to take SPOT (Simple Performance-Oriented Test) prior to the first task in order for the researcher to understand their general language proficiency and performance of the phonological processing in detail. SPOT is a performance test that requires the test takers to fill in the blanks with one character. The test sentence is read in a natural speaking speed of Japanese native speakers. SPOT90 has three sections: beginner level, intermediate level, and advanced level, and test takers are required to go through 30 questions in 10 minutes for each session. Table I shows the results of SPOT90 for the participants in this experiment.

Table- I: Results of SPOT90

	Beginner level problems	Intermediate level problems	Advanced level problems
Average (n = 23)	29.6	26.3	23.7
Standard deviation	0.92	2.21	3.29

A total of 36 patterns of stimulus pseudo-words were created for this task, which consist of two minimal pairs: three syllables and three morae, and three syllables and four morae.

The ones with three morae were "takapa," "pataka," "petepa," and "potaka," and the ones with four morae include geminate, which is added to each of the first and second syllables. Also, for each stimulus word created by the combination of the number of syllables and geminate positions, they have three different accent patters: Low-High-High accent (LHH), High-Low-Low accent (HLL), and Low-High-Low accent (LHL). To create these pseudo-words, the consonants /p/, /t/, /k/ and the vowels /a/, /e/, /o/ were used. Japanese language has five distinctive vowels /a/, /e/, /i/, /o/, /u/ of which the devoiced vowels /i/ and /u/ were not used. As for the consonant from phonological perspective, voiced double consonant does not occur in Japanese, and Chinese does not have double consonant but tones, which appears on vowels. Also, to investigate the relationship between Chinese tones and Japanese geminate, fricative consonants were avoided in the tasks done by Yamamoto. Audacity is used for voice manipulation and acoustic editing of these audio data, and the live voice actually read by three women in their early twenties was used as the sound source for the audio data. The audio data was confirmed by two graduate students who specialize in Japanese language education. The audio data was manipulated as follows; the length of the stop in the words with three syllables and three morae was 80ms, and for the words with three syllables and four morae with a geminate in it, the length of the soundless stop was 220ms, both of which were acoustically adjusted to make them sound natural and appropriate for the purpose of the study. In the first task, accent identification task, the participants were presented the stimulus words three times randomly for each word and asked to identify the accent pattern. The participants were given 1 point when they identified the accent pattern correctly, and when they failed in identifying the accent pattern, they were given 0 point. The result shows that the accent patterns were correctly identified in order of LHH, HLL, and LHL. Also, there was no significant difference between the words with and without geminate in the word, indicating that whether there is a geminate in the word or not does not affect the identification of accent patterns (Table II).

Table- II: Results of Accent Pattern Identification Task

		Accent patterns			
		Low-High-High	High-Low-Low	Low-High-Low	Average
Geminate position	None	95.2%	88.9%	77.8%	87.3%
	First Syllable	95.2%	84.1%	76.2%	85.2%
	Second Syllable	92.1%	76.2%	88.9%	85.7%
	Average	94.2%	83.1%	81.0%	

B. Categorical Perception Task

A total of 23 learners and 23 Japanese undergraduate students, whose major is Japanese language education, participated in the second task. The learners were asked to take SPOT90 prior to participating in the task.

The stimulus pseudo-words, of which the length of the geminate stop was acoustically divided every 10ms, with the shortest length of 80ms to the longest 220ms, which adds up to 15 different versions for each pseudo-word, were used. The participants were presented a series of the manipulated pseudo-word with 10ms ascent/descent change of the length of the geminate stop, and were asked to react when they thought the length changed. There was a 1.5sec interval between each version of the word.

Retrieval Number: 100.1/ijrte.B62840710221
DOI: 10.35940/ijrte.B6284.0710221
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

As shown in Table III, for each group of native speakers and learners, a two-way ANOVA (Analysis of Variance) was performed with the categorical perception task as the dependent variable. As a result, regarding the accent pattern, the degree of categorical perception was higher in the order of HLL, LHH, and LHL. As for the geminate positions, the result shows that the categorical perception was not stable when the geminate was in the first syllable than when it was in the second syllable. Also, as for the LHL accent pattern, the degree of categorical perception was not stable irrespective of the groups of participants. This is because i) the LHL accent pattern has two undulations in it compared to the other two accent patterns, ii) there is not much LHL accent pattern vocabulary in Japanese language, iii) The Chinese intonation system does not have inverted V shaped tone. The numbers in the parentheses in Table III show the degree of categorical perception, indicating that the smaller the number is, the stabler the degree of categorical perception is. From the results of the first and the second task, it was clarified that both the difference of the accent patterns and the geminate position affect categorical perception.

Table- III: Results of Categorical Perception Task

Stimulus word			Chinese Native speaker (n = 23)		Japanese learner (n = 23)	
series	Accent pattern	Geminate position	Average threshold	Categorical perceptual	Average threshold	Categorical perceptual
Ascent	High-Low-Low	2	7.5 (2.0)	0.0	7.8 (2.6)	0.9
Descent	High-Low-Low	2	7.5 (1.8)		8.7 (3.5)	
Ascent	High-Low-Low	1	7.7 (1.9)	1.0	7.0 (2.1)	3.0
Descent	High-Low-Low	1	8.7 (1.5)		10.0 (2.1)	
Ascent	Low-High-High	2	6.9 (1.9)	1.4	9.1 (2.9)	2.8
Descent	Low-High-High	2	8.3 (2.0)		6.3 (2.4)	
Ascent	Low-High-High	1	6.9 (2.1)	2.6	7.4 (2.8)	2.3
Descent	Low-High-High	1	9.5 (1.7)		9.7 (2.7)	
Ascent	Low-High-Low	2	5.8 (1.5)	3.8	6.9 (2.9)	3.3
Descent	Low-High-Low	2	9.6 (2.0)		10.2 (3.4)	
Ascent	Low-High-Low	1	5.2 (1.6)	6.1	5.5 (2.2)	7.2
Descent	Low-High-Low	1	11.3 (2.2)		12.7 (3.3)	

III. AUTOMATIC GEMINATE INSERTION ALGORITHM

In this section, we describe how to automate the manual acoustic manipulation performed by Yamamoto in II, B. The automation process needs to be performed in order of "mora extraction by noise removal", "matching of original audio data and extracted mora", and "insertion of soundless duration and geminate."

A. Mora Extraction by Noise Removal

In Japanese, there is a unit called mora defined by a combination of vowels (V: vowel) and consonants (C: consonant). The basic mora has a V or CV structure, but some have a CCV structure. On the other hand, geminate, syllabic nasal, and long vowels called special mora are written in /Q/, /N/, and /R/, respectively.

The target is unvoiced plosives [p], [t], [k]. A plosive is produced by stopping the exhalation by using lips, palate, or teeth, and then suddenly releasing the exhalation. Figure 1 shows the schematic diagram of how plosives are produced. In unvoiced plosives, the vocal cord does not vibrate until the subsequent vowels are produced, whereas in voiced plosives, the vocal cord vibrates before the vowels are produced.

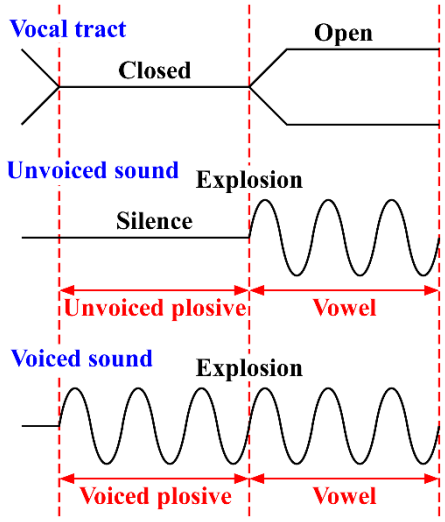

Fig. 1 Schematic of Plosives [t], [p], [k]

Fricatives and affricates are also described as supplements for the explanation of geminate. Figure 2 shows a schematic diagram of fricatives, and Figure 3 shows a schematic diagram of affricate.

A fricative is a sound produced by narrowing the vocal tract at the place of articulation and letting exhaled air pass through the gap. In unvoiced fricatives, the vocal cord does not vibrate until the subsequent vowel is produced, but in voiced fricatives, the vocal cord vibrates when articulated (Figure 2).

Affricate has the characteristics of both plosive and fricative because the air flow is momentarily closed at the place of articulation and becomes a fricative after it is released. For unvoiced affricate consonant, the vocal cord does not vibrate until the subsequent vowel is produced, whereas for voiced affricate consonant, the vocal cord vibrates before the subsequent vowel is produced (Figure 3).

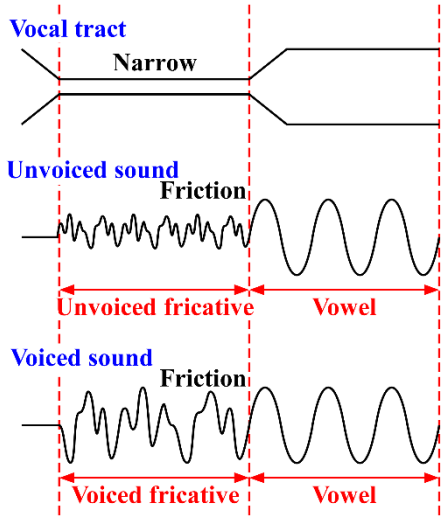

Fig. 2 Schematic of Fricatives [ʃ], [s], [z]

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

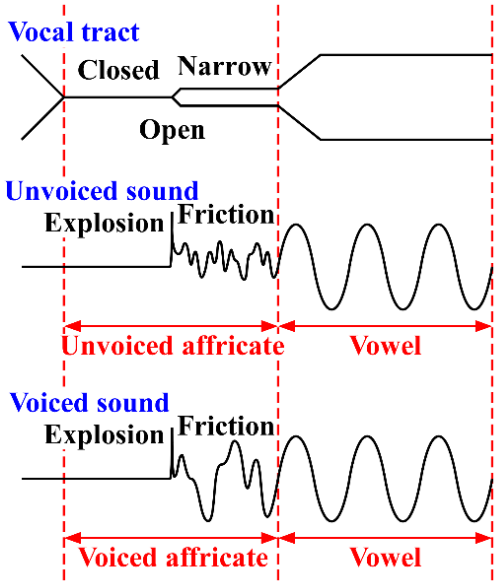

Fig. 3 Schematic of Affricates [t͡s], [t͡ʃ]

Also, the following figure shows how geminate is produced in Japanese language (Figure4). In Japanese phonology, geminate is produced differently depending on how the consonant which becomes the geminate is articulated. When the consonant is plosive or affricate, the geminate is pronounced like a long soundless stop, on the other hand, for fricative consonants, the geminate is pronounced by continuous pronunciation of the consonant itself, such as [ʃː]. For both cases, the geminate is counted as one mora.

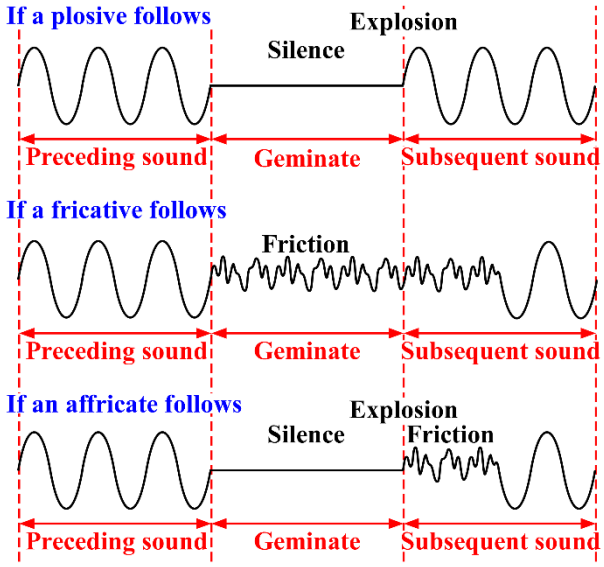

Fig. 4 Schematic of Geminate Variation

Therefore, theoretically, as shown in Figure 5, by inserting soundless duration into the first or second syllable of the pseudo-word consisting of three syllables and three morae, which is the original audio data, the pseudo-word that consists of three syllables and four morae including the geminate can be created.

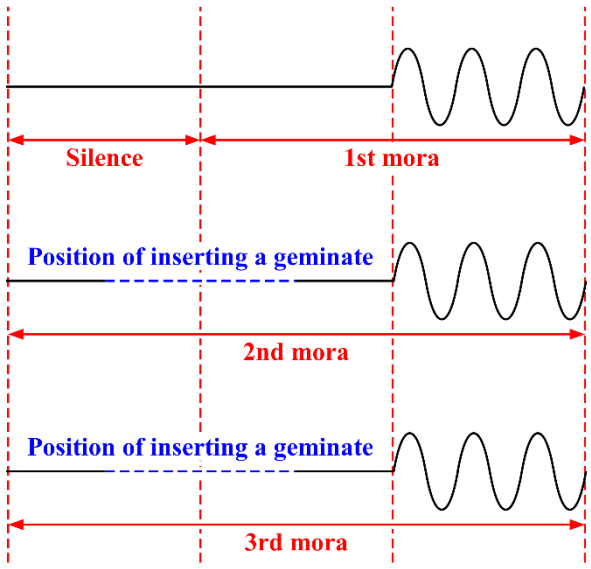

Fig. 5 Method of Creating Audio Data of 3 Syllables and 4 Morae

However, in reality, as shown in Figure 6, various noises that we do not hear, such as daily life noise, such as when clothes rubbed, or noise emitted from recording equipment, are recorded together in the original audio data.

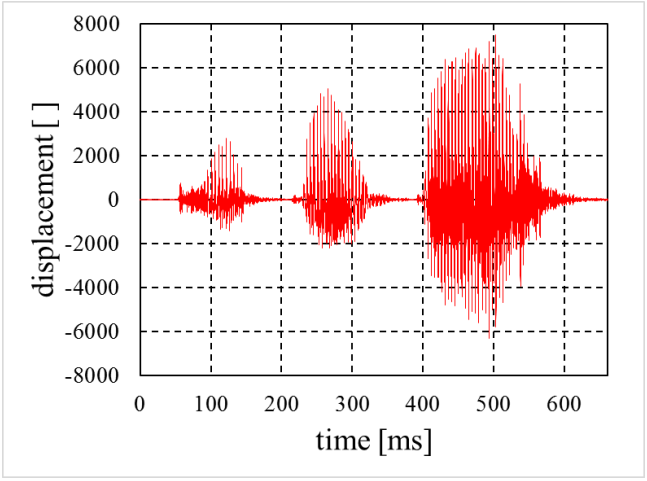

Fig. 6 Example of Noise Recorded in Audio Data

Therefore, by removing the noise that cannot be heard, only the mora part (CV) is extracted. Figures 7 to 9 are the extraction results from Fig. 6 when the noise is -50 dBFS and continues for 10ms or more.

Retrieval Number: 100.1/ijrte.B62840710221
DOI: 10.35940/ijrte.B6284.0710221
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

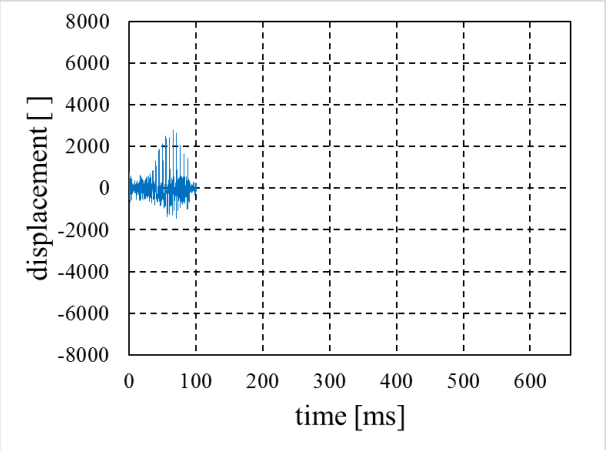

Fig. 7 Extraction of /ta/ by Denoising in Fig. 6

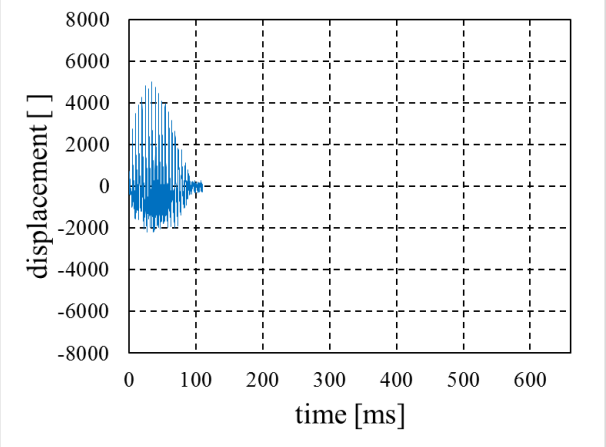

Fig. 8 Extraction of /ka/ by Denoising in Fig. 6

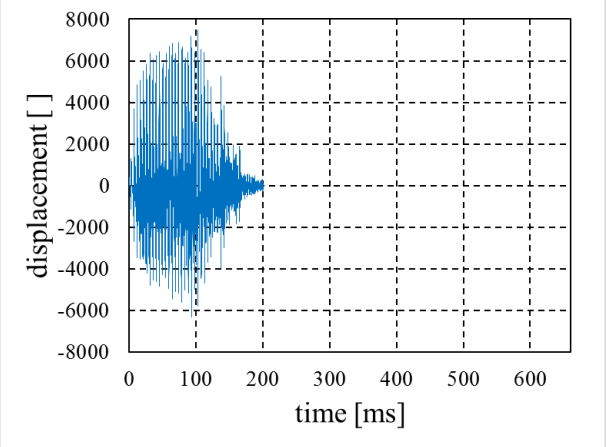

Fig. 9 Extraction of /pa/ by Denoising in Fig. 6

B. Matching of Original Audio Data and Extracted Mora

Each mora was extracted by noise removal in III.A, but it cannot be restored to one audio data again unless the extraction location is specified. Therefore, the place where the original audio data and the extracted audio data of each mora match is specified. Since the original audio data and the audio data of each mora are digital values, and each mora is extracted from the original audio data, there is always a place where these data match perfectly. Figure 10 shows a match between Fig. 6 of III.A and Fig. 7 to Fig. 9.

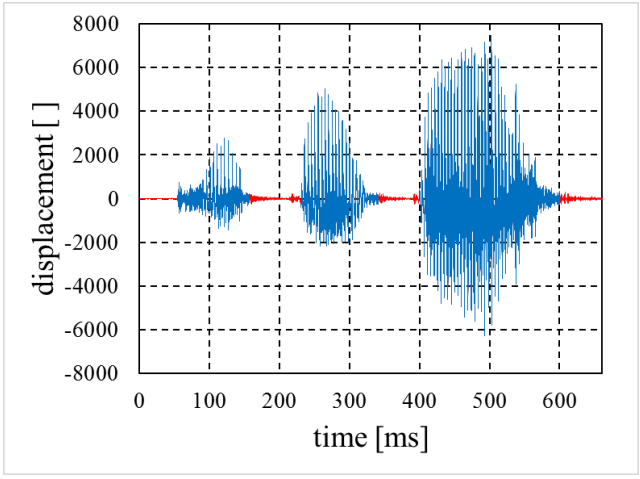

Fig. 10 Matching Example of Original Audio Data and Extracted Mora

C. Insertion of Soundless Duration and Geminate

Since the extraction location of the mora was identified by the matching of III.B, it is possible to restore the original audio data by filling this gap with soundless duration (0).

This is because the noise part of II.B is originally a sound that cannot be heard, meaning no problem will occur even if it is replaced with soundless duration by the algorithm. In addition, when replacing with soundless duration, geminate can be expressed by adding one mora of soundless duration to the first or second syllable. The reason why this is possible is that the mora is composed of three unvoiced plosives /p/, /t/, /k/ and vowels, and the geminate is equivalent to one mora, as already mentioned in III.A. Figure 11 shows the restoration of the audio data in Fig. 10 of III.B. Figure 12 shows the case where the 1.5 msec geminate is inserted in the first syllable, and Figure 13 shows when inserted in the second syllable.

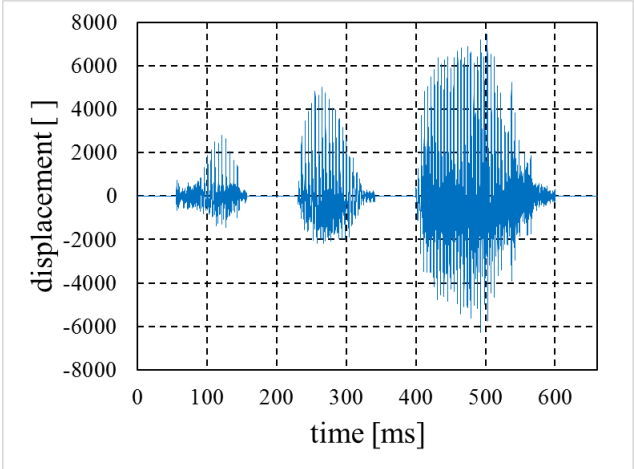

Fig. 11 Restoration Example of Audio Data

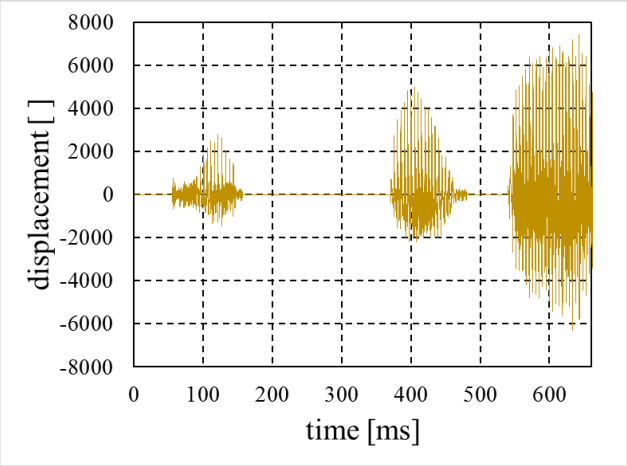

Fig. 12 Example of Inserting a Geminate into the First Syllable of Audio Data

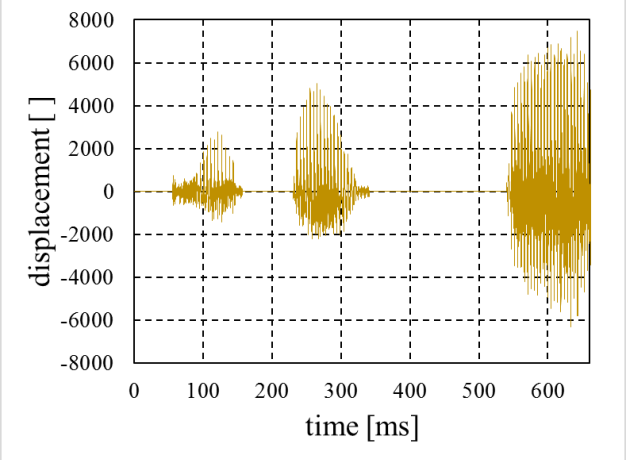

Fig. 13 Example of Inserting a Geminate into the Second Syllable of Audio Data

IV. TESTING OUT THE ALGORITHM

Using the automatic geminate insertion algorithm mentioned in III, the geminate was inserted into the audio data of "Takapa" used in the categorical perception task mentioned in II.B. Table IV shows the result when the geminate is inserted in the first syllable, and Table V shows when inserted in the second syllable. From Table IV and Table V, it can be seen that in all cases, the point of manually inserted geminate is within the automatically inserted soundless duration.

Table- IV: Comparison of Manual and Automatic Geminate Insertion into the 1st Syllable

Speaker	Accent	1st Syllable		
		Programmatic silence start time [ms]	Manual geminate insertion position [ms]	Programmatic silence end time [ms]
A	LHH	208.98	250.00	305.96
	LHL	171.00	220.00	278.96
	HLL	161.00	200.00	250.98
B	LHH	85.99	140.00	201.97
	LHL	110.00	150.00	221.97
	HLL	172.00	210.00	255.96
C	LHH	156.98	190.00	230.98
	LHL	214.99	256.00	290.98
	HLL	192.99	235.00	276.96

Table- V: Comparison of Manual and Automatic Geminate Insertion into the 2nd Syllable

Speaker	Accent	2nd Syllable		
		Programmatic silence start time [ms]	Manual geminate insertion position [ms]	Programmatic silence end time [ms]
A	LHH	415.99	475.00	543.97
	LHL	378.98	430.00	503.97
	HLL	342.00	400.00	477.96
B	LHH	304.99	365.00	433.97
	LHL	322.00	390.00	462.97
	HLL	333.99	390.00	449.98
C	LHH	341.00	370.00	399.98
	LHL	438.98	485.00	530.98
	HLL	431.00	470.00	520.98

Figures 14 and 15 show the ratio of the soundless duration before and after the geminate when the prompting position of the manual operation is used as the reference (0) for Table IV and Table V. The insertion point of the geminate by manual operation is the central part of the soundless section visually predicted by Yamamoto. Similarly, in Fig. 14 and Fig. 15, the geminate insertion position is located at almost the center of the soundless section, which is equivalent to that of manual operation, indicating that this algorithm can be utilized for audio data processing in voice acquisition study.

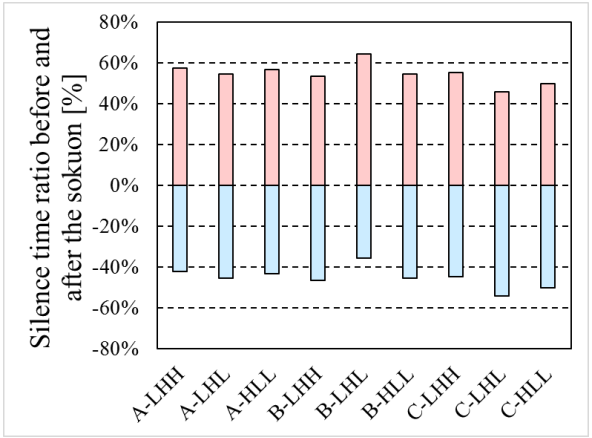

Fig. 14 Point of Geminate Inserted in the 1st Syllable

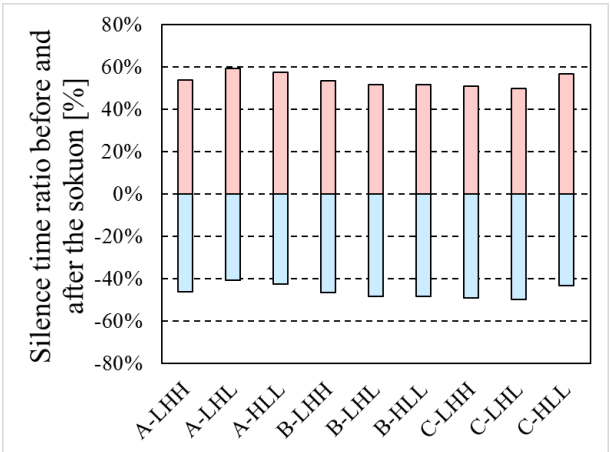

Fig. 15 Point of Geminate Inserted in the 2nd Syllable

V. CONCLUSION

In this study, we developed an algorithm that automatically detects the noise duration and inserts geminate in audio data. The algorithm automates the insertion of geminate by performing three processes in order: mora extraction by noise removal, matching of original audio data and extracted mora, and insertion of soundless duration and geminate. Also, as a result of a phonology researcher comparing the audio data generated by the algorithm with the data that was manually manipulated, the accuracy was equivalent irrespective to the variation of the stimulus words, indicating that the algorithm can be a practical solution for geminate insertion automatization (Fig. 14, Fig. 15).

REFERENCES

1. Toru I., "An Error Analysis on the Perception of Japanese Geminate Consonants by Native English Learners of Japanese: From the perspective of accent and located position", Bulletin of the Graduate School of Education, Hiroshima University. Part. II, Vol. 60, 2011, pp. 173-181.
2. Toru I., "The Effects of Heavy Syllable Position and Accentual Patternon the Perception of Japanese Special Moras by L2 Learners of Japanese", The journal of educational research, The Chugoku-Shikoku Society for the Study of Education, Vol. 8, 2011, pp. 21-30.
3. Teruhisa U., "CHARACTERISTICS OF AUDITORY COGNITION OF LONG VOWELS AND DOUBLE CONSONANTS FOR CHINESE STUDENTS IN LEARNING JAPANESE LANGUAGE", Japanese Journal of Educational Psychology, VOl. 41, No. 4, 1993, pp. 414-423.
4. Teruhisa U., "Evaluation of basic researches into perception of Japanese MORAsounds on teaching Japanese for foreigners", Bulletin of the School of Education, Nagoya University (Educational Psychology), No. 41, 1994, pp. 87-102.
5. Teruhisa U., "Categorical Perception of Relatively Steady-Static Speech Sound Duration in Japanese Morale Phonemes", Journal of the Phonetic Society of Japan, Vol.2, No.3, 1998, pp.76-86.
6. Rie O., "An Analysis of the Japanese Language Rhythm by Learners: Observed in the Vernacular Style Speech", Departmental Bulletin Paper, Ochanomizu University, No. 24, 2011, pp. 1-12.
7. Haruo K., "Geminate Obstruents and Accent in Japanese", NINJAL Project Review, No. 6, 2011, pp. 3-15.
8. Takako T., "Perceptual Categorizationn of the Durational Contrasts by Japanese Learners", Studies in language and literature. Language, Vol. 33, 1998, pp. 65-82.
9. Chikako M., "Effects of rhythm training for pronunciation of geminated stops by Chinese speakers", Konan Kokubun, Vol. 5, 2006, pp. 120-105.
10. Hiroshi M., "Preliminary Study of Syllabic Moras in Rhythm Training", Bulletin of the Department of Teaching Japanese as a Second Language, Hiroshima University, No.14, 2004, pp. 25-32.
11. Hiroshi M., "Native Speakers' Perception of Japanese Speech Segmentation", Journal of the Department of Japanese, Tohoku University, Vol. 6, 1996, pp. 81-92.
12. Makoto K., Eiji O., Steven S., "Phonetic features and notable differences between Japanese, English and Chinese: for understanding a second-language learning barrier", J. of Kyushu Univ. of Health and Wlfare, Vol. 8, 2007, pp. 133-138.
13. Michiyo K., "The Control and Perception of Durational Contrast of Vowels by Mandarin Chinese Speaking Learners of Japanese", Journal of the Phonetic Society of Japan, Vol. 10, No. 2, 2006, pp. 77-85.
14. Takako T., "Issues regarding Geminate Consonants in Japanese Language Education", Journal of the Phonetic Society of Japan, Vol. 11, No. 1, 2007, pp. 35-46.
15. Kenta Y., "The Effect of Geminate Positions and Different Accent Patterns on Categorical Perception of Geminate Sound", M.A. Thesis Hiroshima University, 2017.

AUTHORS PROFILE

Hirofumi Maeda, is Associate Professor in the Information Science and Technology Department at National Institute of Technology (KOSEN), Yuge College. Dr. Maeda's study focuses on practical application of mechanical engineering, namely, developing rescue robots, pipe inspection robots and natural language processing. Dr. Maeda previously served as a researcher at the NPO International Rescue System Institute. Dr. Maeda is currently a member of the Japan Society of Mechanical Engineers, the Robotics Society of Japan, the Japan Association for College of Technology, and the Japan Institute of Marine Engineering. Dr. Maeda has published nine peer-reviewed papers and presented 78 papers. In addition, Dr. Maeda received two awards at academic conferences and 12 external funds.

Kenta Yamamoto, is currently Assistant Professor in the Department of General Education at National Institute of Technology (KOSEN), Yuge College. His research interest is second language acquisition with a main focus on phonetic and phonology aspects of language learning. He served as Deputy Chief of Public Relations at the current institute and has taught Japanese and English at universities and colleges.