Abstract: For Japanese speech processing, being able to automatically recognize between geminate and singleton consonants can have many benefits. In standard recognition methods, hidden Markov Models (HMMs) are used. However, HMMs are not good at differentiating between items that are distinguished primarily by temporal differences rather than spectral differences. Also, gemination depends on the length of the sounds surrounding the consonant. Because of this, we propose the construction of a method that automatically distinguishes geminates from singletons and takes these factors into account. In order to do this, it is necessary to determine which surrounding sounds are cues and what the mechanism of human recognition is. For this, we conduct perceptual experiments to examine the relationship between surrounding sounds and primary cues. Then, using these results, we design a method that can automatically recognize gemination. We test this method on two datasets including a speaking rate database. The results attained well-outperform the HMM-based method and overall outperform the case when only the primary cue is used for recognition as well as show more robustness against speaking rate.

Keywords: Gemination, Automatic recognition, Japanese, Perception

PACS number: 43.72.–Pq, 43.71.–Km

1. INTRODUCTION

In Japanese, for voiceless obstruents and nasals there are two lengths: singletons and geminates. This contrast is a distinctive feature for lexical items in Japanese which makes it important for word recognition. In written Japanese the contrast is realized by preceding non-nasal consonants with a ‘small-tsu’ /Q/ and nasals with a moraic nasal /N/ and differentiates words such as /kaki/ (fence) or /kaQki/ (liveliness) or /ishi/ (will) vs /iQshi/ (touch) and for nasal consonants with the moraic nasal /N/ such as in /kona/ (powder) vs /koNna/ (this kind). A method that can accurately recognize this contrast could have many benefits in speech processing.

One area it may be able to benefit is automatic speech recognition (ASR). In ASR, automatic recognition of this feature could help with the rescoring an N-best list. While present ASR systems for Japanese do attempt to recognize this feature, the way of carrying out recognition is not suitable for automatic recognition of gemination. The reasons for this will be discussed below.

Automatic recognition of gemination could also assist in the improvement of computer-assisted language learning (CALL) systems for Japanese pronunciation. It is said that Japanese language learners often struggle with acquiring the gemination singleton contrast [1,2]. Therefore, if this feature could automatically be recognized, it would be possible to inform learners if they are accurately producing this timing feature.

To recognize gemination it is important to consider its acoustics. Acoustically, gemination is realized differently depending on the acoustics for that particular type of consonant. In the case of fricatives, the frication is longer for geminate fricatives than for non-geminate fricatives. For nasal sounds, the length of the nasality is longer for geminates than for singletons. As for stops and affricates, the duration of the closure is longer for geminates than for singletons [3,4]. We will refer to these features as the primary cues.
However, it has been shown that there can be an overlap between the durations of the primary cues of singletons and geminates [5]. This means that the primary cue of a singleton produced at a slow speaking rate may exceed the duration of a geminate produced at a fast rate. Because of this, a lot of research has been done to investigate secondary cues with the main candidates up to this point being the durations of the immediate surrounding sounds. It is still debated as to what the degree to which each secondary cue affects perception and how they affect perception (discussed in Sect. 3), but it would be desirable to exploit them in an automatic recognition task.

The standard method of recognizing gemination is through the use of HMMs. To discriminate between singletons and geminates, a separate phoneme HMM is used. For ASR, this method has the benefit that the ASR system can follow the standard speech recognition system flow. However, more ideal improvements can possibly be obtained by a recognition method.

There are two major reasons for this. First, it is well-known that HMMs are not very well-suited for differentiating sounds that are spectrally similar, but temporally different [6]. Because geminate consonants differ from singleton consonants primarily in duration, HMMs will often fail when trying to recognize gemination. Also, HMMs will not make use of information pertaining to the durations of the surrounding sounds, since they operate on a frame-by-frame basis and the state at the current frame depends only on the observation and the state of the previous frame. Because of this, a method that makes better use of temporal features in gemination recognition is desirable.

To do this, a post-processing on the initial alignment generated by HMMs can be used as in Fig. 1 to produce the phoneme sequence with gemination recognized. There have been methods proposed that take such an approach, but these methods either do not take secondary cues into account or cannot easily work if the other consonants in the utterance have not been classified as geminate or singleton. Therefore, a new method is necessary. For construction of such a method, ideally the features should have a causal relationship with the contrast between geminates and singletons. Additionally, knowledge of how each feature affects perception can be beneficial.

In order to investigate these aspects, perceptual experiments consisting of stimuli that make up speech continua can be employed. Speech continua have long been used in phonetics research to inspect what variables different categorization processes depend on. Advances in speech synthesis technology have made the process of generating speech samples for such controlled experiments easier, and with speech analysis software such as Praat [7], the process can largely be automated with its PSOLA [8] based resynthesis. Praat makes it possible to easily manipulate the duration of one sound in isolation while leaving the duration of the other sounds constant.

Through this manipulation, stimuli constituting a continua from singleton to geminate can be created. Then listening tests can be conducted in which the subjects classify the target sound of each generated stimulus as a geminate or singleton. From the results of such tests, it is possible to determine what features the perceptual boundary of the contrast depends on and how those features affect the perceptual boundary, with the perceptual boundary being the hyperplane which satisfies \( P(G = 1) = 50\% \), and from this it should be possible to develop an automatic recognition method.

We have conducted listening tests as described in the previous paragraph to investigate these issues. Based on these investigations, we have created a method to automatically classify a consonant as geminate or singleton incorporating the lengths of the surrounding vowel sounds and in addition to those, for affricates and stops, frication and aspiration respectively. This method is based primarily on a logistic equation which was trained on the responses for the listening tests. This equation calculates the probability of a sound being a geminate depending on the durations of the surrounding sounds and classifies based on these probabilities. Since voiced obstruent gemination occurs primarily in the case of loan words [9], in this study we focus on the perception and recognition of voiceless geminates and voiced nasal geminates.

In Sect. 2, previous methods proposed for automatically classifying gemination will be discussed. In Sect. 3, previous research conducted on gemination cues will be discussed. In Sect. 4, we describe the perceptual experiments that we carried out to determine the relationship between the surrounding sounds and geminate perception. In Sect. 5, the proposed recognition method based on our inspection into perception will be described, and then the results of a recognition experiment will be discussed. For this experiment, we compare three methods: an HMM-based, primary cue duration-based, and the proposed method based on the primary cue duration plus the
secondary cue durations that we investigated. Lastly, in Sect. 6, the conclusion to this paper will be given.

2. PREVIOUS METHODS FOR AUTOMATIC RECOGNITION

2.1. Overview of Previous Methods

There have been several methods proposed before to automatically recognize geminate consonants. As mentioned in the introduction, the conventional way to recognize gemination is to use an HMM. This method is used in speech recognition. Geminate consonants other than nasals will be preceded by the HMM ‘q,’ while nasal geminates will be preceded by the geminate ‘N.’ If the likelihood of these HMMs occurring before the consonant is higher than not occurring, the consonant is recognized as a geminate. As mentioned above, HMMs are not good at handling temporal differences or long term dependencies so this method will often produce erroneous outputs.

In research for creating a CALL system, Kawai et al. [10] employed listening tests for creating a method to automatically classify a vowel as long or short. In these tests, they used minimal pairs differentiated by the length of one consonant such as /hata/ (flag) and /haQta/ (put up). The stimulus set was created by lengthening and shortening the duration of the consonant that distinguished the words of the minimal pair. Thus, in this case samples making up the continuum from /hata/ to /haQta/ were created. Then, they played these stimuli to native speakers prompting the speaker for which word of the minimal pair the stimulus was. After obtaining the results from these tests, the consonant duration/selection rate graph was fit to a logistic curve to calculate $P(S = 1|CL)$ or $P(S = 1|WL)$ where $CL$ is the consonant length, $S$ signifies singleton, $G$ indicates geminate. With this logistic equation, the duration of the consonant to classify was used to determine whether or not it was a geminate. This method, however, did not take the durations of the surrounding sounds which will change due to speaking rate and other factors into account so there will be times when it misclassifies.

Because Kawai’s method did not consider the effects due to speaking rate, another method was proposed by Yamamoto et al. [11]. The method also made use of listening tests to obtain classification equations. The goal of this method was to take into account the dependencies due to the durations of the surrounding sounds. This was done by first lengthening the duration of the entire word and then for each word duration, resynthesizing the target consonant to have 10 different lengths. Then, the perceptual boundaries were obtained. After obtaining all of the perceptual boundaries, a linear fit of the change in perceptual boundaries due to vowel length was obtained. With this equation, the perceptual boundary of the consonant for an arbitrary word length could be calculated. After estimating this perceptual boundary, the consonant was classified as a geminate if its duration exceeded this perceptual boundary and a singleton if it was less than this boundary.

This solution has a problem, though. The goal of the method is to calculate $P(S|WL, CL)$ ($S$ is singleton, $WL$ is word duration, $CL$ is consonant duration) or $P(G|WL, CL)$. The problem is that $WL$ is dependent not only on speaking rate, but also on the lengths of the other vowels in the word. If another consonant length is mispronounced, the method may break down. Also, it does not provide clues for how to recognize errors for arbitrary words.

Lastly, Ishi et al. tackled this issue using the inverse speaking rate (ISR). In that work [12], first ISR is calculated with the formula

$$ISR = \frac{T}{N}$$

where $T$ is the total amount of time that all of the morae span and $N$ is the number of morae. The durations of geminates and the durations of singletons for the samples in the corpus were calculated for each ISR value. After this, linear fits were obtained for the relationship between geminate and singleton durations and the inverse speaking rate. These lines were used to predict the duration of an arbitrary singleton or geminate for an arbitrary ISR value. If the duration of the consonant to recognize was closer to the predicted duration for a geminate it was labeled as geminate, and if it was closer to the predicted duration of a singleton it was labeled as a singleton.

This method has the same problem, though, that Yamamoto et al.’s method has. In order to calculate the speaking rate used for classification the number of morae must be determined and to determine the number of morae all of the other consonants must be classified. However, in order to classify the consonants the speaking rate is necessary.

Another problem with the two methods is that it is not apparent whether or not they agree with human perception. There have been investigations into human perception which might provide clues for how humans recognize gemination. We believe that with further investigation into human perception more accurate methods can be developed.

3. PREVIOUS RESEARCH ON PRIMARY AND SECONDARY CUES FOR GEMINATION

There have been many works delving into the acoustics and perception of gemination looking at the primary and secondary cues, especially for the case of stops. There have been a number of works that have acoustically investigated the durations of the primary cues [9,13–16] finding that typically a geminate consonant is over twice the length of a singleton consonant. However, because with different
speaking rates the duration of singletons and geminates can overlap [5], a lot of research has been devoted to looking at secondary cues as well.

Much research has attributed one secondary cue to being the duration of the preceding vowel. Acoustically it has been found by several researchers that the preceding vowel is longer before geminate consonants and the following vowel is shorter than for singleton consonants [14–20]. If perception agrees with this then, it would be expected that a longer preceding vowel duration would result in a consonant more likely being perceived as a geminate and the opposite for the duration of the following vowel.

However, some research done by conducting perceptual experiments has claimed a positive correlation between the duration of the primary cue and the duration of the preceding vowel, while other research has claimed a negative correlation.

In Fukui’s research [17], he suggested that a secondary cue for the perception of geminate stops was the duration of the vowel preceding the closure of the stop. In Watanabe and Hirato’s research [21], a positive linear relationship was found between the duration of the preceding vowel and the closure duration at the perceptual boundary.

Since then there has been a wide variety of research done on this topic. Several researchers have reported a positive correlation for the duration of the preceding vowel with the duration of the primary cue at the perceptual boundary [5,16,22,23]. Other researchers on the other hand have reported, though, that the preceding vowel duration has a negative correlation with the primary cue duration at the perceptual boundary [24–27].

With regards to the following vowel, some perceptual research has claimed little or no effect due to the duration of the vowel following the consonant (following vowel) [28]. On the other hand, a positive correlation with the following vowel has since been reported by [23,26], though this was claimed to be smaller than the effects due to the duration of the preceding vowel in the work of Idemaru et al.

Another issue is voice onset time (VOT) following the release of closure. In [4,5] it has been reported that VOT does not significantly differ between geminates and singletons, but in [13,14], it was reported that the duration of VOT was slightly shorter for geminates than for singletons. In the research of Oba et al. [4], they also looked at the variations in perception of affricate gemination due to the duration of the frication. In the results, there appears to be a slight negative correlation.

Thus, it is still debated what the effects of the durations of the immediate context sounds are on the perceptual boundary and the extent of their effects.

4. LISTENING TESTS TO BETTER UNDERSTAND PERCEPTION

4.1. Proposal for Listening Test

In previous research, it has still not been settled what the effects of the secondary cues on perception are and the extent of their effects. In our study, we would like to develop a method to automatically recognize gemination, though. To do such, it is necessary to further examine the effects of the durations of the surrounding vowels on the primary cue. Also, it is important to look at their effects on other consonant types as well. Though Oba et al. [4] investigated the perception of affricates and fricatives, an in-depth look at the effects of the surrounding vowels on their perception was not conducted. Also, for our research we would like to derive an equation that can factor in the effects of the surrounding sounds. Thus, we will make some changes to the experiment design and stimuli in order to reach this objective.

4.1.1. Experiment design changes

To do this, we make some changes to the design of experiments that have previously been carried out. We investigate further the effects of the durations of the sounds surrounding the primary cue for stops and also look into their effects on nasals, affricates, and fricatives. We also, look into perception of aspiration/frication for stops and affricates respectively.

In the listening tests for previous research, it appears only options containing a short vowel for the preceding vowel were displayed in much of the research. If the preceding or following vowel duration is lengthened, though, it may become a long vowel. In this case, an option should be given to the subject that reflects this. If this is not reflected in the options, the subject may use both phonemic vowel and consonant lengths to make their selection. We would like the subject only to select based on whether a geminate was heard or not. Because of this, we mask all of the katakana surrounding the mora containing the target consonant with asterisks. By doing this, the problem of how to represent the preceding and following vowels can be avoided. Also, it will be more clear to the subject which part of the word he or she is to compare.

4.1.2. Stimulus set and subset creation

For this investigation, we first resynthesize a set of stimuli in which the duration of a surrounding sound (context sound) is manipulated in several steps to be several durations. Each duration of the context sound makes up a subset. Then for each subset, we resynthesize stimuli by manipulating the primary cue so that the primary cue durations span the singleton/geminate continuum.

After conducting a listening test where the subject classifies each stimulus, the primary cue duration/selection
rate graph for each subset is fit to a logistic curve and the perceptual boundary is determined similar to the research of Amano et al. [22]. In order to determine duration of the perceptual boundary for each length of the context sound, the logistic equation

\[ P(G = 1) = \frac{1}{1 + e^{\alpha (pcd - \beta)}} \]  

(2)
can be employed, where \( pcd \) is the duration of the primary cue, \( \beta \) is an estimate of the mean duration of the primary cue at the perceptual boundary and \( \alpha \) is the slope at the perceptual boundary. \( \alpha \) can be thought of as a measure of inverse spread, since the larger \( \alpha \) is, the smaller the spread of the primary cue at the perceptual boundary is.

To estimate the standard deviation for the perceptual boundary given a context sound, it can be noted that because the \( \alpha \) parameter is a measure of inverse spread (it is inversely proportional to the variance), the inverse of the sigmoid can be taken

\[ x = \frac{1}{\alpha} \ln \left( \frac{1 - y}{y} \right) \]  

(3)
where \( x \) is the length of the target consonant (primary cue) in this case and \( y \) is \( P(G = 1) \). Since \( \beta \) is an offset from 0 and we would like to know the deviation from the mean, it is not necessary to consider \( \beta \) in this function. Then by setting \( y \) equal to 0.341 (1 standard deviation), an approximation of the sample standard deviation of the decision boundary can be known. Considering that \( \beta \) should be an approximation of the mean, statistical significance can be calculated with summary data as given in [29].

### 4.2. Stimulus Creation for this Experiment

For duration resynthesis we make use of the speech software Praat [7] using PSOLA [8] to carry out the manipulation. First, a set is created by choosing an utterance and manipulating the sounds in the utterance to base lengths. A context sound is chosen for each set and its duration is manipulated 5 times to form 5 subsets. Then, the primary cue of the consonant of syllable 2 is manipulated 9 times for each subset to determine the primary cue duration at the perceptual boundary.

This analysis is divided into three groups based on the context sound and its location: Group 1 (preceding vowel), Group 2 (following vowel), Group 3 (aspiration/frication). In this experiment, only nonsense words are used. The nonsense words were chosen to contain affricates, fricatives, nasals, and stops to analyze the effects on different consonant types. They were also selected to have the consonants that can geminate except for the palatalized versions and to contain all 5 Japanese vowels for the sounds surrounding the manipulated consonant. The words forming the corpus were read by native Japanese speakers in a soundproof room. After recording, all of the vowels were equalized to be of approximately the same length, 100 ms, and all of the consonants roughly 80 ms, except for a few cases. The sets and durations of the sounds can be found in Table 1. The symbol ‘x’ indicates that the sound is a manipulated sound. An ‘*’ indicates that the sound does not exist for that particular word. ‘Cons’ indicates consonant, ‘Asp’ for aspiration, and ‘Vow’ for vowels. Aspiration duration is only given for the second consonant. Group 1 sets are suffixed by a ‘P,’ Group 2 sets by a ‘F,’ and Group 3 sets by an ‘A.’

For analysis, the vowels were segmented by locating the abrupt drop or rise in energy and presence of formants for the vowel. In the case of fricatives, frication was considered to go from one vowel to the next vowel. The closure was considered the area between where there is a sudden decline in F2 energy for the vowel and the burst associated with affricates and stops. Aspiration was

| Word      | Cons1 | Vow1 | Asp2 | Vow2 | Cons3 | Vow3 |
|-----------|-------|------|------|------|-------|------|
| bepisaP   | 0.04  | x    | 0.02 | 0.091| 0.072 | 0.098|
| ratogiP   | 0.031 | x    | 0.019| 0.083| 0.08  | 0.08 |
| dafeseP   | 0.043 | x    | *    | 0.091| 0.0723| 0.092|
| pishazuP  | 0.035 | x    | *    | 0.102| 0.06  | 0.11 |
| jinesP    | 0.031 | x    | *    | 0.103| 0.063 | 0.091|
| pimigaP   | 0.033 | x    | *    | 0.092| 0.067 | 0.086|
| gotsumeP  | 0.038 | x    | 0.037| 0.098| 0.075 | 0.085|
| sechishoP | 0.047 | x    | 0.037| 0.098| 0.065 | 0.089|
| bisokuP   | 0.033 | x    | *    | 0.098| 0.08  | 0.097|
| takeseP   | 0.033 | x    | 0.026| 0.091| 0.083 | 0.071|
| ganepiP   | 0.041 | x    | *    | 0.098| 0.08  | 0.096|
| binesaF   | 0.039 | 0.099| *    | x    | 0.071 | 0.090|
| chimoF    | 0.042 | 0.085| *    | x    | *    | *    |
| chimokuF  | 0.039 | 0.093| *    | x    | 0.091 | 0.091|
| dafeseF   | 0.041 | 0.098| *    | x    | 0.079 | 0.09 |
| dashoF    | 0.027 | 0.101| *    | x    | *    | *    |
| dashochiF | 0.033 | 0.097| *    | x    | 0.075 | 0.086|
| gechikaAF | 0.029 | 0.099| 0.097| x    | 0.07  | 0.091|
| gechikaBF | 0.029 | 0.04  | 0.037| x    | 0.062 | 0.079|
| petsubiF  | 0.043 | 0.096| 0.045| x    | 0.073 | 0.093|
| repasuF   | 0.041 | 0.098| 0.019| x    | 0.073 | 0.103|
| tafheF    | 0.03  | 0.107| *    | x    | 0.065 | 0.088|
| zatogiAF  | 0.04  | 0.094| 0.014| x    | 0.069 | 0.084|
| zatogiBF  | 0.04  | 0.04  | 0.014| x    | 0.069 | 0.084|
| bisokuF  | 0.035 | 0.1  | *    | x    | 0.087 | 0.09 |
| gamepiF  | 0.044 | 0.1  | *    | x    | 0.080 | 0.10 |
| takeseF  | 0.032 | 0.095| 0.028| x    | 0.085 | 0.083|
| bepisA    | 0.040 | 0.097| x    | 0.048| 0.069 | 0.106|
| ratogiA   | 0.040 | 0.101| x    | 0.044 | 0.075 | 0.089|
| petsubiA  | 0.041 | 0.097| x    | 0.094 | 0.072 | 0.094|
| sechishoA | 0.057 | 0.096| x    | 0.094 | 0.066 | 0.091|
considered the point from the burst to the following vowel. When resynthesizing, the middle 45 ms were resynthesized for each vocalization to manipulate. If the duration was 45 ms or less for that particular sound, a smaller range was chosen with that range made to be shorter than the duration of the sound and approximately in the middle of that sound. If there were not 45 ms, a smaller interval was chosen, but this interval was made sure to be approximately in the middle, not equal to the length of that vocalization, and not extend to the boundaries. For the primary cue duration, the manipulation region was decreased to 0.05 times (roughly 0.03 s for most) its base duration for the stimulus with the minimum duration and increased to 7 times its size for the stimulus with the maximum duration (roughly 0.24 s for most). The step size was set to equal lengths with 7 intervals between the minimum and maximum durations. The resulting durations for the manipulated secondary cue can be found in the graphs of Sect. 4.4.

4.3. Listening Test Environment

A total of 90 individuals participated in this experiment, with between 20–33 people listening to each stimulus. The subjects were all natives of Japanese. These experiments were all conducted with an Internet-based Java program that plays samples at random. It also gave the subject the option to return to the previous stimulus by clicking on ‘modoru’ (go back) and replaying a stimulus by clicking ‘saisei’ (play). He or she was also informed to wear headphones for the experiment. There were two choices given in two buttons for the subject to select for each stimulus. The top button (singleton), showed the katakana for the mora that contained the target consonants and all of the sounds surrounding that mora were masked with asterisks. The bottom button (geminate) was the same as the top except the target consonant mora was preceded by the katakana for a small ‘tsu’ for non-nasal consonants and a moraic /N/ for nasal consonants. The subject was told to select the button the consonant was heard as.

4.4. Listening Test Results and Discussion

The results for stops, fricatives, nasals and affricates where the manipulated context sound was the previous vowel, are given in the top four graphs of Fig. 2, where it was the the following vowel in the middle four graphs of Fig. 2, and where it was the aspiration/frication in the bottom two graphs of Fig. 2. The x-axes indicate the duration of the context sound and the y-axes indicate the duration of the primary cue at the perceptual boundary. The correlations for linear fits and the p values (calculated with one-way ANOVA) of the data shown in Table 2. The context vowel for the subset is indicated by the suffix on the word: P (preceding vowel), F (following vowel), A (aspiration).

For Group 1, the results are given in the top 4 graphs of Fig. 2. It can be seen that there is a negative correlation for the perceptual boundary duration with the preceding vowel duration for all cases up to around 150 ms. After reaching 150 ms, it can be seen that the variance grows larger and for some samples such as /bisoku/, the correlation goes from being negative to positive and the variance in the responses increases. This is possibly because the previous vowel has changed from being a short vowel to a long vowel at around this point. In Japanese, geminates are not favored after long vowels [27] so this is possibly a reason for this phenomenon. These results agree with the results of Ofuka et al., Arai et al., Takeyasu, and Kingston et al. [24–27], in which it was reported that there was a negative correlation with the duration of the previous vowel for stops. Overall, the results did not agree with some of the other research such as [22], but some of the correlations did become positive after the previous vowel duration passed around 150 ms in duration. The cause for this is possibly because these previous experiments only displayed two options for which the entire word was displayed, one option which had a geminate and one which had a singleton. The preceding vowel was only displayed as being short. To determine the exact reason for why this would be, more experiments are needed.

For many sets the results are not statistically significant, but it would probably be significant if more stimuli had been created for each subset. In almost all cases the correlation is negative, which would be extremely unlikely if there was not a relationship.

For Group 2, the results are given in the middle 4 graphs of Fig. 2. It can be seen that there is a positive correlation between the duration of the primary cue at the perceptual boundary and the duration of the following vowel. This is the case for every consonant type. Because of this, it is safe to say that with our experiment setup, the following vowel has a positive correlation with the duration of the primary cue at the perceptual boundary. These results show that the following vowel duration does have a significant effect for all consonant types. The results agree with the findings in [23,26], in which a positive correlation was found for stop consonants.

For Group 3, the results are shown in the bottom two graphs of Fig. 2. A negative correlation can be observed for the duration of the aspiration/frication for stops/affricates respectively with the duration of the primary cue at the perceptual boundary. Since these results are significant, it is important to consider these durations for a classification equation as well.

Thus, from these results it can be said that the primary cue, preceding vowel, following vowel, and aspiration/frication durations should all be used as features. In [30],
they proposed the use of the disyllable length as a feature, but we do not make use of this. The reason for this is because the previous vowel duration had a negative correlation and the following vowel length had a positive correlation with the perceptual boundary of the primary cue. Thus, lengthening the following vowel by some value will decrease the probability, and lengthening the previous vowel by the same value may increase the probability. Because of this, it can be thought disyllable duration is not causally related to the perception of gemination. For the same reason, the methods employed by Yamamoto, word length [11], and Ishi, inverse speaking rate [12], also do not appear appropriate for solving this problem.

5. PROPOSED METHOD FOR AUTOMATIC GEMINATE RECOGNITION

5.1. Recognition Flow

In the introduction, we gave an overview of the system...
flow, which is illustrated in Fig. 1. First, the alignment is carried out using speech recognition. Then gemination recognition is carried out.

5.1.1. Phoneme/acoustic alignment

In our work, because our goal is to compare gemination recognition accuracy, the alignment will be carried out with a forced alignment. This is done with phoneme HMMs to obtain the initial phoneme boundaries. Closure and aspiration/frication need to be segmented in the case of stops and affricates respectively. However, because it is not straightforward how to segment aspiration from closure with standard HMMs, we use the zero crossings for each 10 ms frame for the stop consonant and affricate consonant to carry out this segmentation. Zero crossings are calculated as follows

\[
Z_n = \sum_{m=-\infty}^{\infty} 0.5 \left| \text{sign}(x[m]) - \text{sign}(x[m - 1]) \right| w[\hat{n} - m]
\]  

where

\[
\text{sign}(x) = \begin{cases} 
1, & \text{if } x \geq 0 \\
-1, & \text{if } x < 0
\end{cases}
\]

where the window is a 25 ms Hamming window and the frameshift is 10 ms. The position where the number of zero-crossings/s exceeds 1,000 and stays above 1,000 was chosen to be the boundary between aspiration/frication and closure.

In Japanese there are also palatalized consonants such as /kʲ/ which results in a higher F2 for the following vowel. Typically in a speech recognizer these will be recognized with one HMM such as 'ky.' This HMM will typically cover the voiced formant transition also, so for this experiment, we recognized these consonants with two HMMs, 'k' and 'y,' and added the duration of the 'y' HMM to the duration of the following vowel.

5.1.2. Geminate recognition

In the next stage, the consonant is classified as a geminate or a singleton. If the consonant sound follows an /N/, is the first consonant, or it follows a long vowel it is automatically classified as a singleton. The reason to classify it as a singleton after a long vowel is because a geminate consonant does not typically follow a long vowel in native Japanese words [27]. In addition, for the listening tests the variance became quite large for the perceptual boundary for the stimuli where the previous vowel was long so we do not consider this issue at this point.

The consonants were classified based on the logistic equations below, similar to how Kawai [10] performed classification, but modified to account for the primary and secondary cues rather than solely calculate the probability by consonant duration. They were calculated by

\[
P(G) = \frac{1}{1 + e^{\alpha(\text{pcd} - (\gamma\text{aspd} + \beta) + \delta \text{pd})}}
\]

\[
P(G) = \frac{1}{1 + e^{\alpha(\text{pd} - (\gamma\text{aspd} + \beta) + \delta \text{fd})}}
\]

where the bottom equation is for affricates and stops, and the top equation is for fricatives and nasals, and pcd is the primary cue duration, pd is the preceding vowel duration, fd is the following vowel duration, aspd is the aspiration or fricative duration for stops and affricates respectively. The parameters for these equations are trained based on the responses from the subjects in the listening test experiments. We classify all of the consonants that make up a /N/, is the first consonant, or it follows a long vowel it is automatically classified as a singleton. The reason to classify it as a singleton after a long vowel is because a geminate consonant does not typically follow a long vowel in native Japanese words [27]. In addition, for the listening tests the variance became quite large for the perceptual boundary for the stimuli where the previous vowel was long so we do not consider this issue at this point.

The consonants were classified based on the logistic equations below, similar to how Kawai [10] performed classification, but modified to account for the primary and secondary cues rather than solely calculate the probability by consonant duration. They were calculated by

\[
P(G) = \frac{1}{1 + e^{\alpha(\text{pcd} - (\gamma\text{aspd} + \beta) + \delta \text{pd})}}
\]

\[
P(G) = \frac{1}{1 + e^{\alpha(\text{pd} - (\gamma\text{aspd} + \beta) + \delta \text{fd})}}
\]
perceptual experiments with a greater number of samples for each subset are needed to confirm this.

For comparison, we also train an equation that only takes account of the primary cue for each consonant type. Then we conduct recognition experiments with this method as well. This equation is essentially the same as the above equation except that the parameters for \( \gamma, \phi, \) and \( \delta \) are set to 0 for the parameter training.

5.2. Primary Cue Duration Adjustment

Because the phoneme boundaries that result from the forced alignment often do not agree with hand labeled boundaries, the duration of the primary cue are adjusted by

\[
pcd = pcd_{auto} + pcd_{adj}
\]

(8)
to account for consistent errors in the automatic alignment, where \( pcd_{auto} \) is the primary cue duration given by the automatic alignment and \( pcd_{adj} \) is the adjustment value.

The reason for only adjusting the duration of the primary cue is that it has a much larger influence on the perception of gemination than the other cues. The adjustment value to use for the test set, \( pcd_{adj0} \), is chosen so that it minimizes the difference between the recognition rate for geminates and singletons for each consonant types

\[
 pc_{adj0} = \arg \min_{pc_{adj}} \left| \frac{GRR}{SRR} - pc_{adj} \right|
\]

(9)
where \( GRR \) is the geminate recognition rate and \( SRR \) is the singleton recognition rate and \( pc_{adj} \) is the adjustment value used in the recognition function. The adjustment value is calculated by the algorithm given in Algorithm 1.

Require: \( MAX IT \geq 0 \)

\[
N \leftarrow 0
\]

\[
 pcd_{adj} \leftarrow 0
\]

\[
 incr_val \leftarrow 0.1
\]

\[
 incr_{adj} \leftarrow \frac{incr_{val} - 0.001}{MAX IT}
\]

while \( N < MAX IT \) do

\[
GRR, SRR \leftarrow \text{recognize_all}(pcd_{adj})
\]

if \( GRR < SRR \) then

\[
 pcd_{adj} \leftarrow pcd_{adj} - incr_{val}
\]

else if \( SRR < GRR \) then

\[
 pcd_{adj} \leftarrow pcd_{adj} + incr_{val}
\]

else

\[
 breakloop
\]
end if

\[
 incr_{val} \leftarrow incr_{val} - incr_{adj}
\]

\[
 N \leftarrow N + 1
\]
end while

Algorithm 1: Calculate \( pcd_{adj}. \) recognize_all() recognizes the gemination for all utterances and returns the recognition rate for geminates and singletons for a particular consonant type. \( GRR \) is geminate recognition rate and \( SRR \) is singleton recognition rate.

5.3. HMM-Based Method

In addition to testing this method against the case where only the primary cue is used, we also test the proposed method against an HMM-based method. To test the HMM-based method, we conduct a semi-forced alignment on all of the utterances in the test sets. For a standard forced-alignment all of the phonemes are specified for the speech recognizer to align to. We modify the alignment stage in Sect. 5.1.1 to allow both singletons and geminates for every voiceless fricative, stop, and affricate and voiced nasal when carrying out alignment. Thus, in the alignment each non-nasal consonant could be recognized as being preceded by a ‘q’ (name of HMM) or not preceded by a ‘q’ and each nasal consonant preceded by a ‘N’ or not preceded by an ‘N.’ However, consonants that begin a word, followed a long vowel, or followed a /N/ are not permitted to be geminates.

5.4. Datasets for Evaluation

We evaluate these methods on two different databases. The primary database is a speaking rate database using nonsense words. The secondary dataset that we tested it on contains native speech with (meaningful) words.

5.4.1. Speaking rate database

Since speaking rate is one aspect that affects the duration of surrounding sounds, we have constructed a speaking rate corpus to test our method. At faster speaking rates, geminates and singletons are shorter than they are for slow speaking rates. It is important to have a set that has a sufficient number of geminates and singletons and consonant types. In normal speech singleton consonants are more prevalent than geminate consonants so the balance of singletons and geminates will be poor. Therefore, for our database we made use of nonsense words and had them read at three speaking rates, slow, medium, and fast similar to as was done in the work of Hirata [31].

This database was constructed of nonsense words in isolation and ones inserted into the carrier sentence “kokoga_da” (this place is _). The nonsense words were automatically generated by a computer program we created. We generated words between 2 and 5 syllables with this program. First, the structure of each syllable was randomly selected from the list V, CV, V:, CV:, VN, CVN, V:N, VQ, CVQ, and CV:N. Then a phoneme was randomly selected for each sound of each syllable. We created five different word lists with this program. Each speaker read a different word list. For each word list, one third of the carrier sentences and isolated words were to be read at a slow speaking rate, one third at a medium rate, and one third at a fast rate. The speakers consisted of two female and three male native Japanese speakers.

5.4.2. Native dataset

It is also important to test the method on meaningful
speech as well, so we also test this method on native read text. This dataset consists of both sentences and isolated words. The native speech used for testing can be found at [32].

5.5. Alignment

We assume that the productions would be perceived as the speaker intended so we use the text the speaker read for the automatic alignment. The utterances are all aligned with the speech recognizer Julius [33]. For the acoustic model, a monophone three state-model trained on Japanese speech is used. Alignment is carried out in two different manners. For the proposed method, the alignment is carried out with the adjustments to the alignment process as mentioned in the Sect. 5.1 in order to perform recognition with the proposed method. Also, for the HMM-based method semi-forced alignment is done as discussed in Sect. 5.3.

5.6. Training Duration Adjustment and Testing

To train the duration adjustment parameters for each consonant type, we use the data of 4 speakers from the speaking rate corpus as training data and the remaining speaker as test data. Recognition on this database is repeated 25 times to obtain the adjustment parameters for each consonant type given the method described in Sect. 5.2. The data from each speaker except for the Speaker 5 are used as test data. The reason Speaker 5’s data are not used for test data was because the list that he read were made not to contain many geminate consonants. For the native test data, these parameters are trained on the data from all five speakers contained in the speaking rate corpus.

The speaking rate corpus consisted of 2,044 randomly generated isolated words and 522 carrier sentences containing a subset of those randomly generated words. For the speaking rate corpus, 1,644 of the isolated words and 450 of the carrier sentences are used for testing (a total of 4 speakers). The remaining speaker’s data are solely used for training. For recognition of gemination for the native speech, we train adjustment parameters using data from all 5 speakers of the speaking rate database. Gemination recognition is carried out for a total of 832 isolated words and 29 sentences.

5.7. Experiment Results and Discussion

We conducted experiments on all of the test sets based on the above conditions for the three different recognition methods. The overall results for the three methods are given in Table 3. In this table, SR refers to the Speaking Rate corpus, Native refers to the native dataset. The consonant types are given in the columns of the first row: Fric (fricative), Stop, Affric (affricate), and Nasal.

The recognition rates given in this table only contain the consonants which are not automatically labeled as singletons based on our algorithm. AC is the proposed method which makes use of all the cues that we found in this work, and PC is the method which makes use of solely the primary cue. Thus, consonants that follow long vowels, follow moraic nasals, and start words are not included in these results since they will be automatically labeled as short based on our algorithm. The recognition rates in bold are for the pairs of singleton/geminate recognition rates that produced the highest average value for that consonant type.

For the SR corpus, from these results it can be seen that the proposed method outperforms the other methods except for in the case of stops. In this case, the performance is roughly equal. We conducted a McNemar test comparing the overall results of the proposed method to PC and found a significant improvement ($\chi^2 = 13.315, df = 1, p = 0.0003$). Both methods far outperform the HMM-based method. This can be attributed to the fact that HMMs are not very capable when it comes to classifying based on temporal differences.

Since speaking rate can influence the length of the surrounding sounds which have an influence on the perception of gemination, we also compared the results of the method which solely makes use of the primary cue with the method that makes use of the primary cue and secondary cues at different speaking rate levels for the SR database. The results are shown in Fig. 3. In this figure, the y-axis gives the recognition rate and the x-axis lists five different speaking rate levels and each consonant type. To determine the speaking rate range for each speaking rate level, we first fit the speaking rates of all the utterances to a single Gaussian distribution. The speaking rate levels were chosen to correspond to each 20th percentile on the

| Consonant Type | Fric | Stop | Affric | Nasal |
|----------------|------|------|--------|-------|
| HMM            | S    | 0.96 | 0.51   | 0.68  | 0.39  |
|                | G    | 0.18 | 0.97   | 0.89  | 0.84  |
| SR             | S    | 0.81 | 0.91   | 0.81  | 0.84  |
|                | G    | 0.88 | 0.90   | 0.90  | 0.90  |
| PC             | S    | 0.88 | 0.90   | 0.84  | 0.90  |
|                | G    | 0.89 | 0.91   | 0.88  | 0.93  |
| Native         | HMM  | 1.00 | 0.76   | 0.81  | 0.56  |
|                | S    | 0.91 | 0.97   | 0.96  | 0.99  |
|                | G    | 0.88 | 0.90   | 0.85  | 0.89  |
| AC-Proposed    | S    | 0.88 | 0.97   | 0.90  | 0.98  |
|                | G    | 1.00 | 0.94   | 0.92  | 0.95  |
Gaussian curve. In this graph, the higher the percentile is, the faster the speaking rate is. This figure does not contain the results for the HMM-based method because the results for that were much lower than for the other two methods.

In the figure, it can be seen that AC performs better than PC. Overall, a certain degree of robustness can be seen against speaking rate for the proposed method. It can be seen that the results are overall more equal at slower speaking rates for AC and the average results are higher for it as well. Thus, it is stronger at handling a variety of speaking rates due to the use of secondary cues. However, both perform poorly at the fastest speaking rate level.

In Table 4, the average of the singleton and geminate recognition rates for each phoneme for the proposed method are given. In this figure, it can be seen that the proposed method gives roughly balanced results for the different phonemes.

One of the major issues with this method is that it relies on the automatic alignment generated by HMMs. The boundaries measured by HMMs often do not coincide with the actual phoneme/vocalization boundaries. Most of the time the error is small, but sometimes it is quite large. Because of this, it is believed that alignment errors are the biggest problem that the proposed method faces. For faster speaking rates, the relative amount of error will be larger, since the average length of a vocalization will be shorter. This is likely part of the cause in the drop-off in recognition rates for faster speaking rates.

### Table 4

Recognition rates for gemination length recognition experiment. CR is the rate of correctly recognizing the consonant length.

| Consonant | CR |
|-----------|----|
| p         | 0.90 |
| t         | 0.94 |
| k         | 0.88 |
| ky        | 0.89 |
| py        | 0.94 |
| ch        | 0.86 |
| ts        | 0.83 |
| s         | 0.91 |
| sh        | 0.88 |
| f         | 0.87 |
| n         | 0.94 |
| m         | 0.90 |
| ny        | 0.83 |
| my        | 0.90 |

### 6. CONCLUSION

In this paper we have proposed a method to automatically discriminate between geminate and singleton consonants. Such discrimination can have benefits in different applications of speech processing such as CALL systems and ASR. This method is motivated by perceptual experiments which we carried out. They lead to a better understanding of gemination perception and contributed to the construction of an automatic gemination recognition algorithm. From the listening tests we found that the length of the following vowel, the aspiration/frication, and the
length of the previous vowel all had a causal relationship with the perception of gemination for all types of consonants. Also, we found that aspiration/friction duration and preceding vowel duration up to around 150 ms were negatively correlated with the length of the primary cue similar to what has been reported before by some researchers for stop consonants [24–27]. Also the following vowel duration was positively correlated with it in all cases as has been reported by [23,26]. However, the correlation in many cases was positive after the preceding vowel exceeded around 150 ms.

The results of these listening tests helped us design a method to automatically classify a consonant as a geminate or singleton. To do this we used a logistic equation that had been trained on the subject responses for the listening test. We then incorporated this equation in the overall classification algorithm.

Overall, we were able to obtain noticeable improvements with our method. We compared this algorithm to the HMM-based method and a method that solely makes use of the primary cue. Of the three, the proposed method performed the best and the HMM-based method performed the worst showing the importance of making proper use of temporal features. Also, because it outperformed the case where only the primary cue is used, it shows the value of conducting perceptual experiments to understand the mechanism of perception. This method appears ready for application.

In order to improve the results further, better alignment is crucial. All of the features that we used are measurements of duration. Thus, accurate measurements in duration are essential for automatic classification. Also, it is possible that the incorporation of other possible cues such as the intensity contour and pitch features is important in classification. It has been claimed that fundamental frequency and intensity have an effect on human classification of gemination [16,34]. In addition, we would like to look into the effects of more global features as mentioned in [3] and try to understand the cause of the discrepancy between our work on gemination perception and that of [22,23].

For the future, to improve the results of this method we plan to work on improving the alignment and investigating the effects from the previous vowel more in depth as well as the intensity contour. We also intend to use this method in ASR and CALL systems and conduct experiments on non-native speech.

REFERENCES

[1] T. Toda, “Perceptual categorization of the durational contrasts by Japanese learners,” Tsukuba Univ. Linguist. Repos., 33, 65–82 (1998).
[2] M. S. Han, “The timing control of geminate and single stop consonants in Japanese: A challenge for nonnative speakers,” Phonettica, 49, 102–127 (1992).
[3] S. Kawahara, “The acoustics and perception of Japanese special moras,” J. Acoust. Soc. Jpn. (J), 69, 191–196 (2013) (in Japanese).
[4] R. Oba, A. Braun and J. Handke, “The perception of Japanese geminates by native and non-native listeners,” ISPhS: Int. Soc. Phonet. Sci., 92, 9–29 (2005).
[5] Y. Hirata and J. Whitton, “Effects of speaking rate on the single/geminate stop distinction in Japanese,” J. Acoust. Soc. Am., 117, 1647–1660 (2005).
[6] H. Strik, K. Truong, F. de Wet and C. Cucchiarini, “Comparing different approaches for automatic pronunciation error detection,” Speech Commun., 51, 845–852 (2009).
[7] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer (Version 5.1.05) [Computer program],” http://www.praat.org/ (2012).
[8] H. Valbret, E. Moulines and J. Tubach, “Voice transformation using PSOLA technique,” ICASSP-92, Vol. 1, pp. 145–148 (1992).
[9] M. Beckman, “Segment duration and the ‘mora’ in Japanese,” Phonettica, 39, 113–135 (1982).
[10] G. Kawai, “Spoken language processing applied to non-native language pronunciation learning,” PhD Dissertation, Vol. 1, 02 (1999).
[11] M. Yamamoto and J. Miwa, “Computer assisted learning system for Japanese special mora and its evaluation,” Tech. Rep. Hear. Acoust. Soc. Jpn., pp. 1–8 (2000).
[12] C. T. Ishi, K. Fujimoto and K. Hirose, “Identification of Japanese “tokushuhaku” regarding the influence of speaking rate,” Tech. Rep. IEICE, Vol. 100, No. 97, pp. 17–24 (2000).
[13] Y. Homma, “Durational relationship between Japanese stops and vowels,” J. Phonet., 9, 273–281 (1981).
[14] M. S. Han, “Acoustic manifestations of mora timing in Japanese,” J. Acoust. Soc. Am., 96, 73–82 (1994).
[15] R. Port, J. Dalby and M. O’Dell, “Evidence for mora timing in Japanese,” J. Acoust. Soc. Am., 81, 1574–1585 (1987).
[16] K. Idemaru and S. G. Guion, “Durational relationship between Japanese stops and vowels,” J. Int. Phonet. Assoc., 38, 167–186 (2008).
[17] S. Fukui, “Perception for the Japanese stop consonants with reduced and extended durations,” Onsei Gakkai, 159, 9–12 (1978) (in Japanese).
[18] Y. Hirata, “Durational variability and invariance in Japanese stop quantity distinction: Roles of adjacent vowels,” J. Phonet. Soc. Jpn., 11, 9–22 (2007).
[19] S. Kawahara, “A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese,” Language, 82, 536–574 (2006).
[20] E. Ofuka, “Perception of a Japanese geminate stop [t̚]: The effect of pitch type and acoustic characteristics of preceding/following vowels,” J. Phonet. Soc. Jpn., 7, 70–76 (2003).
[21] S. Watanabe and N. Hiroto, “The relation between the perceptual boundary of voiceless plosives and their moraic counterparts and the duration of the previous vowel,” Onsei-Gengo, 1, 1–8 (1985) (in Japanese).
[22] S. Amano and Y. Hirata, “Perception and production boundaries between single and geminate stops in Japanese,” J. Acoust. Soc. Am., 128, 2049–2058 (2010).
[23] K. Idemaru and S. Guion-Anderson, “Relational timing in the production and perception of Japanese singleton and geminate stops,” Phonetta, 67, 25–46 (2010).
[24] J. Kingston, S. Kawahara, D. Chamблess, D. Mash and E. Brenner-Alsop, “Contextual effects of the perception of duration,” J. Phonet., 37, 297–320 (2009).
[25] M. Arai and I. Kawagoe, “Consonant gemination and syllable
types in English: A perceptual test of nonsense words,” J. Phonet. Soc., 2, 87–92 (1998) (in Japanese).

[26] E. Ofuka, Y. Mori and S. Kiritani, “Perception of a Japanese geminate stop: The effect of the duration of the preceding/following vowel,” J. Phonet. Soc. Jpn., 9, 59–65 (2005) (in Japanese).

[27] H. Takeyasu, “Effects of the consonant and vowel durations in the preceding syllable on the perception of geminate stops in Japanese,” Phonol. Stud., 15, 67–78 (2012).

[28] N. Hirato and S. Watanabe, “The relation between the perceptual boundary of voiceless plosives and their moraic counterparts and the length of vowels following closure,” Onsei-Gengo, 2, 99–106 (1987) (in Japanese).

[29] N. S. Butt, “ANOVA with summary statistics: A stata macro,” Pak. J. Stat. Oper. Res., 2, 57–62 (2006).

[30] S. Amano and Y. Hirata, “The perception boundary between single and geminate stops in 3- and 4-mora Japanese words,” Proc. Interspeech 2011, pp. 2233–2236 (2011).

[31] Y. Hirata, “Effects of speaking rate on the vowel length distinction in Japanese,” J. Phonet., 32, 565–589 (2004).

[32] G. Short, “Non-native Japanese language corpus” (2012). http://www.gavo.t.u-tokyo.ac.jp/~short/corpus/.

[33] A. Lee, T. Kawahara and K. Shikano, “Julius—An open source real-time large vocabulary recognition engine,” Proc. Eur. Conf. Speech Communication and Technology (Eurospeech 2001), pp. 1691–1694 (2001).

[34] H. Kubozono, H. Takeyasu, M. Giriko and M. Hirayama, “Pitch cues to the perception of consonant length in Japanese,” Proc. 17th Int. Congr. Phonet. Sci., pp. 1150–1153 (2011).