Automatic Recognition of Vowel Length in Japanese for a CALL System motivated by Perceptual Experiments

Greg Short¹, Keikichi Hirose¹, Nobuaki Minematsu²

¹Graduate School of Information Science and Technology, University of Tokyo, Japan
²Graduate School of Engineering, University of Tokyo, Japan
{short, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Acquisition of the Japanese vowel length contrast can be problematic for non-native speakers. For these speakers, a CALL system which can automatically recognize vowel length could be of great benefit for pointing out their errors and issuing corrective feedback. However, a method that can adequately do this has not been proposed yet. Vowel length recognition is made difficult because the vowel length distinction is dependent on the surrounding vowel durations which vary due to speaking rate and other factors. Hidden Markov Models (HMMs), the standard way of recognizing this distinction, do not make use of this information. Methods have been proposed to recognize this in the past, but they do not appear viable unless knowledge about the durations of other vowels is present. Thus, we carry out perceptual experiments to gain more knowledge about the vowel length contrast. From this analysis, we develop an automatic recognition algorithm for vowel length that uses support vector machines (SVMs). We tested this method on a speaking rate corpus, native speech, and non-native speech. The method produced recognition results that are overall superior to HMMs and also more robust against speaking rate differences with an average of a 0.83 correct recognition rate for the 3 datasets. The error and non-error classification rates on non-native speech for this are 0.86 and 0.84 respectively.

Index Terms: Japanese, vowel length, recognition, speaking rate, perception, CALL

1. Introduction

There are two phonemic lengths for each vowel in Japanese: long and short. This distinctive feature is important for human word recognition and mistakes in it can reduce intelligibility and naturalness [1]. This contrast can be difficult for learners of Japanese to acquire [2], though, so there is a need to develop a CALL system with an algorithm that can automatically recognize this contrast.

In speech recognition, the typical way of recognizing this contrast is through the use of HMMs. The phonemic vowel lengths are recognized by using different HMMs for short vowels and their long counterparts. Despite being the method commonly used in automatic speech recognition, HMMs are not an ideal way to automatically differentiate between the two lengths. One reason for this is that HMMs are said to not be good at differentiating between items that are temporally different but spectrally similar [3].

Another problem is that the decision boundary (perceptual boundary) between long and short vowels varies due to speaking rate. As speaking rate goes from fast to slow the durations of short vowels and long vowels increase and, thus, the vowel duration at the perceptual boundary increases as well. The frame-by-frame processing that HMMs employ will not take this information into account when classifying. Without integrating this into the model, if the learner does not speak at the same rate as the speech used for training the recognition model, there will likely be many vowels that are not errors misclassified as errors and vice-versa.

Thus, for a CALL system, a recognition method that can take these factors into account is desirable. Such a CALL system could have a flow like the one in Fig. 1. In this flow the phonemes for the text that the learner reads and the mic input are forced aligned, and from the phoneme alignments the vowel lengths are recognized. Based on these alignments, error classification and feedback generation can be carried out.

In this paper, we will focus on the vowel length recognition/error classification stages of this flow. Previous research for developing methods to carry out recognition for a CALL system has attempted to factor such features in, but the results do not appear satisfactory and cannot be applied in a simple manner. This will be further discussed in the Section 2. Because of these issues, more knowledge about perception of vowel length is necessary.

Taking the above into consideration, we have conducted perceptual experiments in order to better understand what vowel length classification depends on. After gaining more understanding of the mechanisms of human perception, we have developed an algorithm that is motivated by these perceptual experiments and makes use of SVMs with features motivated by the perceptual experiments.

In this paper, we discuss that algorithm and the basis for its development. We then test that algorithm on a speaking rate corpus using nonsense words and a corpus made up of native Japanese speech data, and non-native speech data. In Section 2, previous methods proposed for automatic recognition of vowel length will be discussed. In Section 3, the perceptual experiments we carried out to develop the method will be overviewed. In Section 4, the proposed method based on these listening tests will be overviewed. In Section 5 the recognition experiments will be discussed. In Section 6, the conclusion will be given.

2. Research on Automatic Classification of Vowel Length

There have been several methods proposed for automatically classifying vowels as short or long. The most common such method is the use of two HMMs to represent the two vowel lengths: a short vowel HMM and a long vowel HMM. This is the simplest and the most widely used method. The problems with this method, however, are that HMMs are not good at
In Kawai et al. [4], they employed listening tests for developing a method. In these tests, they used minimal pairs differentiated by the length of one vowel such as /toru/ (to pass) and /to:ru/ (to take). The stimulus sets were created by lengthening and shortening the vowel length that distinguishes the words of the minimal pair. Thus, in this case a continuum of stimuli from /toru/ to /to:ru/ was created. Then, they played each of these stimuli to native speakers having them select which word of the minimal pair the stimulus was. After obtaining the results from these tests, a logistic equation was obtained for each word by fitting the vowel duration/selection rate graph and this was used for classification. This method, however, did not take speaking rate into account when classifying.

Two other methods that attempted to take this into account were then proposed. One of these methods was carried out by Yamamoto et al [5]. In their research, they conducted listening tests like in Kawai’s method. The difference was that they resynthesized the word length as well as the vowel length. First, the entire word duration was resynthesized to different durations. Then for each word duration, the target vowel was resynthesized to different durations to create continua from short to long for various word durations. From the perceptual experiments they conducted with this data, they derived an equation that predicted how the perceptual boundary of the sound would change due to the duration of the word and used this equation to recognize vowel length.

This can lead to problems, though. For one, it is not apparent if word duration is how humans carry out speaking rate normalization. Another problem is that it assumes that other vowels in the word were correctly produced. It is possible that other vowel lengths may have been incorrectly pronounced. For example, assume learner uttered /oji:saN/ (grandfather). For determining whether the /o/ was mispronounced as /a/ the word duration is used. If the learner mispronounced the /o/ as /a/, though, the word duration will be longer than usual and thus misclassifications can arise due to this so the word duration normalization calculation must take this into account. This is not simple, though, because all of the vowels could potentially be pronounced with the incorrect length. Thus, this does not appear to be a viable solution when all of the vowel lengths are unknown.

The other method that has been developed to handle speaking rate was proposed by Ishi et al [6]. They looked into using the inverse speaking rate (ISR) calculated by dividing the number of seconds by the number of morae as a means to automatically classify vowel length. However, this method has the problem that in order to calculate the ISR it is necessary to determine how many morae there are. In order to determine how many morae there are, though, it is necessary to classify the lengths of all the vowels. Thus, the inputs of the function require the outputs and this method is not easy to carry out.

3. Perceptual Experiments

3.1. Overview

In the previous section, we introduced four methods for automatically recognizing vowel length, three of which were for CALL systems. Two of the methods do not take into account variations due to speaking rate. The other two do not appear easy to apply if all vowel lengths are unknown, unless changes are made to the algorithms. Thus, a new approach to automatic recognition of vowel length is necessary.

For such an approach, better understanding of the perception of vowel length is needed. For this, listening tests like the ones in Kawai et al [4] and Yamamoto et al’s [5] works can be used to further investigate the mechanism of perception.

Thus, we have carried out such perceptual experiments. In these experiments we have investigated how the duration at the perceptual boundary of a target vowel changes due to the durations of surrounding vowels. First, we recorded several nonsense words of the syllable structure CVCVCV in a soundproof room. Then, the durations of one or two context vowels or a context consonant is chosen to be manipulated. The context vowel(s) we manipulate is(are) manipulated to M durations creating M subsets. For the cases where two context vowels are manipulated, both of the context vowels have a different duration for each subset. Then for each duration of the context vowel(s) (each subset), the target vowel was manipulated N times, creating M x N stimuli.

After this listening test, the selection rates for the N stimuli of each of the M subsets are fit to a logistic curve to get M fits

\[ P(VowelLength = Long) = \frac{1}{1 + e^{\alpha(td - \beta)}} \]  

where \( td \) is the target vowel duration, \( \beta \) is the target vowel duration at the perceptual boundary and \( \alpha \) is the slope at the perceptual boundary. By analyzing the change in \( \beta \) due to changes in the context vowel durations and preceding consonant duration, it can be understood which sound durations are important for a classification algorithm.

An online program was used to carry out the experiments. When this program initiates, a sample is played to a subject at random and two buttons are displayed. The target sound is shown in the top button with only the katakana character for the mora with the target vowel and on the bottom button with the katakana character for that mora along with a dash to indicate that it is a long vowel. All of the phonemes except for the target vowel are masked with ‘*’ and displayed in both the top and bottom buttons. The subject was instructed to click the bottom button if he or she perceived the vowel as long and the top button if he or she perceived it as being short. A total of 90 native Japanese speakers participated in the experiments.

3.2. Stimulus Sets

In this paper, we discuss three different groups to see the effects of the durations of the surrounding vowels and preceding consonant. For the first group, we chose the middle vowel
of a three syllable word to be the target and the surrounding two vowels to be the manipulated context sounds to create the subsets. Both of these context sounds were lengthened for the creation of each subset and for each of the subsets the target vowel was manipulated 13 times. This was to see how manipulating the durations of two context vowels would affect perception. In the second group, one of the context sounds was manipulated and the other was held constant. This was to see how perception would change if only one surrounding vowel was manipulated for comparison with the case where two were manipulated. Lastly, in the third group, the preceding consonant was manipulated to various durations and the other sounds were held constant to determine if the consonant duration affected vowel length perception. For the second and third groups, for each length of the context sound, the target sound was manipulated 9 times. The details concerning the manipulations are given in Table 1. The context sounds that were not manipulated were set to roughly 100ms in duration for vowels and 80ms for consonants except for the first consonant which was set to be 40ms.

### Table 1: Manipulations for words for different sets.

<table>
<thead>
<tr>
<th>word</th>
<th>target</th>
<th>manipulated context</th>
<th>nonmanipulated vowel dur (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bepisa</td>
<td>V2</td>
<td>V1 &amp; V3</td>
<td>N/A</td>
</tr>
<tr>
<td>TakeSe</td>
<td>V2</td>
<td>V1 &amp; V3</td>
<td>N/A</td>
</tr>
<tr>
<td>Bepisa2</td>
<td>V2</td>
<td>V1</td>
<td>0.2s</td>
</tr>
<tr>
<td>Zatogi</td>
<td>V2</td>
<td>V3</td>
<td>0.2s</td>
</tr>
<tr>
<td>Bepisa3</td>
<td>V2</td>
<td>C2</td>
<td>0.15s</td>
</tr>
</tbody>
</table>

#### 3.3. Perceptual Experiment Results

The results for the sets where the surrounding vowels were manipulated are shown in Fig 2. The top graph shows the sets for which both context sounds were manipulated. For this graph, both the left and right contexts are plotted. The set names suffixed by ‘L’ indicate that that plot is for the duration of the left context vowel (V1) duration. The names for the plots appended by ‘R’ indicate that that the x-values for that plot are for the right context (V3) duration. For the case of ‘bepisa’ the right context sound was shorter and in the case of ‘takeSe’ the left context sound was shorter. Despite this, the plots for the shorter of the two context sounds and the plots for the longer of the two context sounds overlap.

In the bottom graph, the duration of the shorter of the two context sounds for both of the nonsense words used in the previous test are plotted with the sets with only one manipulated context vowel. For those words the other surrounding vowel was set to 200ms since we wanted this vowel to remain longer than the other surrounding vowel for most subsets. In this case as well, there is an overlap for the target vowel length at the perceptual boundaries of the four different sets. For the two sets in which one of the surrounding vowels was not manipulated, it appears there is a peak in the vowel duration at the perceptual boundary as the manipulated context duration gets longer than the non-manipulated context sound. This would also agree with the idea that the shorter vowel is what is being used for classification, since it should reach a peak as the manipulated surrounding vowel approaches the duration of the other surrounding vowel. We have conducted more experiments related to vowel length perception in [7] which also showed that as the manipulated context vowel becomes longer than the one that was not manipulated, the target vowel duration at the perceptual boundary peaked.

The results from the set where the consonant was the manipulated context sound can be seen in Fig. 3. In this figure, the x-axis indicates the duration of the consonant and the y-axis the duration of the vowel at the perceptual boundary. We carried out a linear fit on this data and found the coefficient of determination ($r^2$) between the consonant duration and the duration of the vowel at the perceptual boundary to be 0.06. This means that perhaps the consonant duration is not essential for classification.

### 4. Vowel Length Recognition Proposal

#### 4.1. Background of Proposal

In the results of the previous section, it appeared as though the duration of the shorter of the two context vowels is more important in determining whether or not a vowel is perceived as long or short. Also, it did not appear as though the preceding consonant played a role in its perception. Based on those results, we present a vowel length recognition algorithm in this section along with experiments conducted using it.
4.2. Details of Proposal

For considering how to develop an algorithm, the important thing to note is that it appeared as though the shorter of the two surrounding vowels was causally related with the changes in the perceptual boundary. We propose to perform recognition centered around this idea. The recognition flow we propose is given in Fig. 4. This takes the phoneme alignments for a sentence as the inputs and outputs the sentence with the recognized vowel lengths. The output is passed along to the error classification stage given in Fig. 1.

The first stage of this algorithm classifies all of the vowels as being longer than, shorter than, or the same length as the surrounding vowel it is compared to. In the second stage, all of the vowels classified as being the same length are reclassified as long or short based on how the surrounding vowels are classified. If all of the vowels for a particular word are classified as being the same length, in the third stage these vowel lengths will be classified based on the durations of the vowels of the surrounding words.

4.2.1. Initial Classification

In the first stage, each vowel is recognized as shorter than (S), longer than (L), or the same length as (Sa) the surrounding sounds used in classification. At this stage, these are classified with 3-class SVMs. Since vowel length is a lexical feature, this is performed at the word level so if the vowel to the left or right is a separate word, it is not used at this stage. Ideally, the vowels for the sentence ‘watashiwa/isogashi’ (I am busy) will be recognized as ‘Sa-Sa-Sa-Sa/Sa-S-S-L’ at this stage.

4.2.2. SA Vowel Classification

The above stage leaves us with three relative vowel lengths. In the second stage the vowel lengths labeled as Sa are classified as S or L for the words where one or more of the vowels were classified as S or L. If a word does not contain a vowel that has been labeled L or S, this stage is skipped for that word. In this stage, each vowel labeled as Sa next to a vowel that is not labeled as Sa is relabeled to be the same length as the shorter of the surrounding vowels. This process is repeated until all vowel lengths have been relabeled. After completing this process the vowels of ‘watashiwa/isogashi’ should ideally be labeled as ‘Sa-Sa-Sa-SA/S-S-S-L’.

4.2.3. SA Word Classification

The above stage leaves the vowels for words which only have Sa length vowels unchanged. At the third stage, words that contain only Sa vowels are relabeled. The vowels for the first word next to a word that does not contain any Sa vowels are recognized. If the word comes after the word without any Sa vowels (non-Sa word), the vowel duration of the last vowel of the non-Sa word is used as a feature to classify the length of the first vowel of the Sa-word and if it comes before, the duration of the first vowel of the non-Sa word is used to classify the length of the last vowel as in the first stage. After classifying, the length of the classified vowel will be Sa, L, or S. If it is Sa, it is relabeled to be the same length as the vowel that was used to classify it. Then, the lengths for the rest of the vowels for that word can be classified as in stage two. This process is repeated until there are no Sa words. After completing this stage, the vowel lengths should ideally be ‘S-S-S-S/S-S-S-L’ for ‘watashiwa/isogashi’. If all of the vowels for all of the words are Sa, the number of L vowels and S vowels for the word in the canonical spelling are counted. The vowels of the utterance are chosen to have the length which is more prevalent in the canonical spelling. The reason for doing this is because the perceptual vowel length is relative. Since this method uses a relative approach to recognition, differences in speaking rate should be accounted for.

4.2.4. Features Used

For the features, we try a variety of feature sets in addition to simply using the duration of the shorter surrounding vowel. These features are listed in 2. ‘Target’ indicates the vowel to recognize. The nasal duration used is the duration of the moraic nasal /N/ that sometimes follows vowels in Japanese. If it is not present, a duration of 0 is given. For ‘BothSides’ and ‘BothSides2’, if one of the vowel sounds is not present, a value of -10 was given for the duration.

Table 2: Feature sets used for recognition. If not otherwise specified, the vowel/nasal features are the duration of the syllables vowel/nasal.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Features Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shorter</td>
<td>shorter vowel, target</td>
</tr>
<tr>
<td>ShorterNorm</td>
<td>shorter normalized by target</td>
</tr>
<tr>
<td>ShorterNasal</td>
<td>shorter, target, shorter and target nasals</td>
</tr>
<tr>
<td>ShorterFormant</td>
<td>shorter, target, typical shorter and target formant values</td>
</tr>
<tr>
<td>BothSides</td>
<td>left vowel, right vowel, target</td>
</tr>
<tr>
<td>BothSides2</td>
<td>2 vowels to left, left vowel, right vowel target</td>
</tr>
</tbody>
</table>
4.3. Training Data

Since we would like to build a system that is fairly robust against speaking rate, for training and testing we have constructed a small corpus consisting of 5 native Japanese speakers. This corpus consists of a variety of nonsense words read at three different speaking rates: slow, medium, and fast. We created a program to automatically generate these nonsense words. With this program, first, the number of syllables for the word was chosen. Then, for each syllable, a syllable type was randomly chosen from the syllable types that exist in Japanese (i.e., V, CV, VN, VQ, CVN, CVQ and their counterparts with long vowels except for CV-Q). Then, for each sound in the syllable, a phoneme was randomly chosen. We created words with 2, 3, 4, and 5 syllables. Also, for half of the data we placed syllable type constraints allowing only CV, V, CV; and V: type syllables. The recording was conducted in a soundproof room.

5. Vowel Length Recognition Experiment

5.1. Experiment Conditions

For preprocessing, we conducted forced alignment on all of the utterances in the corpora using Julius [8]. For the initial recognition stage, we used SVMs trained with libsvm [9]. We used RBF kernels with parameters selected with 5-fold cross validation.

For our recognition experiments, we tested the method on three speech datasets. First we conducted speaker/vocabulary-item open tests training with 4 of the speakers from the speaking rate data and testing on the remaining one. The speaking rate corpus consists of 2044 randomly generated isolated words and 522 carrier sentences (‘kokogawā’ - ‘this place is ____’) containing a subset of those randomly generated words. For the speaking rate corpus, 1644 of the isolated words and 450 of the carrier sentences were used for testing (a total of 4 speakers). The remaining speaker’s data was solely used for training. Next, we conducted tests on the non-native Japanese database that we have constructed [10]. This database consists of the speech of 4 native speakers and 27 non-native speakers who speak 19 different languages as L1s (first languages). For recognition of the vowel lengths for the corpus, we trained SVMs using data from all 5 speakers of the speaking rate database. For the non-native speech, the vowel lengths for a total of 182 isolated words and 46 sentences were recognized and for the native speech, 832 isolated words and 29 sentences.

5.2. HMM-based Recognition for Comparison

For comparison, we also carried out HMM-based recognition. To do this we performed semi-forced alignment. For semi-forced alignment, all of the phonemes are fixed as in normal forced alignment, but the vowel lengths are not fixed. Thus, all vowels are permitted to be long or short in conducting alignment. We do not compare the proposed method to the other previous methods we introduced since it is not clear how to use them if all vowel lengths are unknown.

5.3. Labeling

For the speaking rate and native datasets we assumed that the vowel lengths produced by natives would be perceived at the length the speaker intended them to be. Thus, for alignment the transcript given for the native speaker to read was used for automatic alignment. For non-native speakers, there are errors in the pronunciation, though, so we had 7 native speakers of Japanese label each vowel length in every utterance as short or long. The vowel length label that made up the majority was chosen as the vowel length for that particular vowel.

5.4. Results and Discussion

The results for the experiment are given in Table 3. From these results it can be seen that the methods based on ‘Shorter’ overall perform better for all cases with ‘Shorter’ performing the best. This agrees with our results from the perceptual experiments from which it appeared that the shorter surrounding vowel duration is important for human classification. In all cases, the methods using the lengths of both sides performed lower. This can be thought to be due to it using information that is not pertinent to the recognition process. Also, the ‘ShorterNorm’ method did not perform as well as the method that used the actual lengths. This is probably either due to a) linear normalization not being appropriate, or b) errors in alignment.

The HMM-based method did not perform well on the speaking rate corpus and its performance was noticeably lower than the ‘Shorter’ based methods for the other two datasets. This can be thought to be HMMs performing recognition in a more absolute manner and not taking into account the duration information of the surrounding vowels. This means that if the speaking rate is near to what the average speaking rate was for corpus, this method should perform fairly well, but it will break down otherwise.

In Fig. 5 the results for the ‘Shorter’ method and HMM-based method are given at various speaking rates for the speaking rate corpus at different percentiles of the Gaussian distribution. To calculate the Gaussian distribution, the morae/second speaking rate was calculated for each utterance. The x-axis indicates the speaking rate percentiles for the sample Gaussian distribution, the different methods, and vowel lengths. The y-axis gives the recognition rate. From these results it can be seen that the proposed method overall performs better than the HMM-based method for a variety of speaking rates. The performance of the proposed method drops, however, for long vowel recognition at the fastest speaking rate, 81-100%. The short vowel recognition rate at slow speaking rates does not have this degradation, though. For the HMM-based method, it can be observed that as the speaking rate gets slower the recognition rate drops for short vowels and as it gets faster it drops for long vowels. This shows a higher-level of robustness for the proposed method.

We also analyzed the capabilities of ‘Shorter’ for error classification on the non-native corpus to ensure that errors and non-errors were being adequately classified so that it could be used for a CALL system. These results are given in Table 4. From these results, it can be seen that correct classification of errors is roughly equal to that of non-errors and both are fairly high, showing sufficient results for a CALL system.

This method recognizes short vowels robustly at all speaking rates, but this method does not perform as well for long vowels at a very fast speaking rate. At very fast speaking rates, alignment errors in time will be a larger percentage of the vowel duration and thus make automatic vowel length recognition more difficult. Thus, it is likely that mistakes in automatic alignment are a large contributing factor to the less robust performance at fast speaking rates for this method. It is also possible that other features should be accounted for as well such as intensity and pitch, which has been said to have some effects on vowel length perception [11].
Table 3: Recognition rates for vowel length experiment. The columns represent the different feature sets/algorithms that were used for conducting the recognition with SVMs plus SpeechRec, the semi-forced alignment results. Rates that are shown in bold indicate that the average of the S/L pair for that Feature Set/Method was the highest.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Shorter</th>
<th>Shorter</th>
<th>Shorter</th>
<th>Shorter</th>
<th>Both</th>
<th>Both</th>
<th>HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Norm</td>
<td>Nasal</td>
<td>Formant</td>
<td>Sides</td>
<td>Sides2</td>
<td>Sides2</td>
<td></td>
</tr>
<tr>
<td>SR</td>
<td>0.89</td>
<td>0.77</td>
<td>0.89</td>
<td>0.9</td>
<td>0.90</td>
<td>0.90</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>0.78</td>
<td>0.83</td>
<td>0.78</td>
<td>0.73</td>
<td>0.74</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>0.91</td>
<td>0.65</td>
<td>0.89</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td>0.85</td>
<td>0.8</td>
<td>0.76</td>
<td>0.72</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Nonnative</td>
<td>0.86</td>
<td>0.56</td>
<td>0.84</td>
<td>0.84</td>
<td>0.77</td>
<td>0.81</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>0.76</td>
<td>0.89</td>
<td>0.75</td>
<td>0.76</td>
<td>0.67</td>
<td>0.70</td>
<td>0.78</td>
</tr>
<tr>
<td>Average</td>
<td>0.83</td>
<td>0.76</td>
<td>0.83</td>
<td>0.82</td>
<td>0.79</td>
<td>0.8</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Figure 5: This figure shows the vowel length recognition for different speaking rates. The x-axis indicates the different methods used, the vowel length, and the speaking rate percentile.

Table 4: Confusion matrix for Non-error/Error Classification on non-native speech for proposed method

<table>
<thead>
<tr>
<th></th>
<th>Non-error</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-error</td>
<td>0.84</td>
<td>0.16</td>
</tr>
<tr>
<td>Error</td>
<td>0.14</td>
<td>0.86</td>
</tr>
</tbody>
</table>

6. Conclusion

Japanese language learners have a difficult time in acquiring the vowel length contrast in Japanese. To assist them, we have developed an algorithm that automatically recognizes vowel length. Because the perceptual boundary changes due to the length of the surrounding vowels, this is not a straightforward problem. Thus, in order to solve this problem we conducted perceptual experiments. In those experiments it appeared as though the shorter of the surrounding vowels was important for recognition. Motivated by this, we created a novel algorithm to automatically recognize vowel length which outperformed a standard HMM-based method on three different databases and showed a higher degree of robustness against speaking rate. The error/non-error classification capabilities for non-native speech were also good. It appears sufficient for use in a CALL system.

While recognition of short vowels was fairly robust across speaking rates, recognition of long vowels was not very robust at fast speaking rates. In the future we would like to find ways to remedy this problem with perceptual experiments and by improving alignment. We would also like to explore using features such as pitch. We also plan to integrate this method into a CALL System with a feedback generation interface.

7. References