Rule-based Method for Pitch Level Classification for a Japanese Pitch Accent CALL System

Greg Short, Keikichi Hirose, Nobuaki Minematsu
Graduate School of Information Science and Technology, University of Tokyo
{short,hirose,mine}@gavo.t.u-tokyo.ac.jp

Abstract

In order to make a CALL system to detect lexical pitch accent errors, we are developing a multi-level pitch pattern recognition algorithm. With this method, it is possible to detect a wider range of pitch accent errors than with previous methods and it is possible to have more flexibility in displaying the pitch pattern to the learner. This rule-based method is based on perceptual experiments in which we made use of speech synthesis to manipulate the pitch of utterances to perform quantitative and qualitative analysis on accent perception. The recognition is based on two-mora units and the local pitch level pattern for those units. By combining the local pitch level patterns for a word it is possible to derive the two level pitch pattern for the word and also provide a variety of ways to display the word to the language learner. The average agreement rate with the labeling for these units was 75% in the case of the Japanese database and 60% in the case of the non-Japanese database.

Index Terms: Japanese pitch accent, classification, perception, CALL system

1. Introduction

In many Japanese language classes, there is very little time dedicated to learning pronunciation. Thus, CALL systems have been developed to supplement classroom instruction [3]. One problem that the systems have dealt with is the automatic error detection for the Japanese lexical pitch accent, a feature that distinguishes words by their mora pitch level pattern (high-low pitch pattern). Many learners have great trouble with correctly producing the Japanese pitch accent [2]. Even so, according to [1], the Japanese pitch accent receives little attention in the classroom so it is our goal to develop a system to automatically detect accent errors.

For the CALL system, we would like to make the error detection as reliable as possible. For error detection in previous pitch accent CALL systems, the accent type of the word was identified [4, 5]. However, as a result of language transfer, it is possible learners will produce pitch patterns outside the Tokyo accent type set. Errors for patterns outside the Tokyo accent type set will not be recognized correctly with those methods. Thus, we developed a method to recognize all possible pitch level patterns described in [9]. The performance of this system, though, was still unsatisfactory. Also, it was limited in that it could only abstract the word into two levels: high and low. However, some instructors of Japanese promote the use of multi-level representations [8]. It is our aim to be able to increase the error detection capabilities as well as provide a method that can be more flexible in how pitch is shown to the learner. In this paper, we will introduce a rule-based method based on equations derived from perceptual experiments and discuss the perceptual experiments carried out to make that algorithm. This algorithm classifies pitch level changes for each contiguous two mora pair throughout the word, and allows conversion to a two-level model as well as easy detection of errors. With this algorithm it is possible to detect errors that would be impossible to detect by using a method that attempts to recognize the accent type for a word.

2. Accent Identification and Proposal

2.1. Japanese Accent

In Tokyo Japanese, the pitch pattern for a word is defined by its accent type. The Tokyo accent types are differentiated by the position of the mora that resides before the pitch fall, the accent

Figure 1: Possible accent types for a four mora word

kernel. A word without a pitch fall is considered to be Type 0, and a word with a pitch fall is termed Type M with M being the position of the accent kernel. With the autosegmental-metrical (AM) model for labeling, each mora of a word is specified as being of high pitch (H), prosodically prominent, or low pitch (L), underspecified [6, 7]. Therefore, this model only has two levels. In Tokyo Japanese a word can have N+1 accent types, where N is the number of morae. Fig. 1 shows the possible accent types for a four mora word, a subset of possible patterns.

2.2. Recognition Algorithm

Since learners experience language transfer, it is necessary to be able to recognize all possible pitch patterns. For example, a learner could produce the pattern HLHL for a type 1 word, which is not in Fig. 1. This pattern has an extra rise, which is the signal for the start of an accent phrase, and an extra accent kernel. This type of error will not be recognized correctly by systems that try to identify the accent type. To deal with this problem, in previous research we proposed a method to identify all possible pitch level patterns, not just the subset of patterns that occur in Tokyo Japanese. This recognition method worked based on the two-level pitch pattern representation and ideally allowed for detection of a wider variety of errors. In this method, first the F0 was extracted, then the phonemes were aligned with the Julius speech recognition engine based on the text the learner read in order to perform mora alignment for the F0 contour. Following this, two mora pairs were trained based on their pitch level for the word, then recognized at the two mora level finding the likelihoods of the four pitch pattern combinations: LL, LH, HL, and HH. Then pitch level recognition was done on all contiguous two mora pairs for the word. Finally, the combination with the highest likelihood was then chosen to be the pattern for the word. However, satisfactory results were not obtained with this method. We discussed some reasons for this in [10]. One of which was that at the local level, there can be pitch level changes that are not accounted for by the two level model. To illustrate, in the top of Fig. 2, the pitch level drops between morae 4 and 5 and then it drops further between morae 5 and 6. However, the AM pattern for this word would be LLLHLL, indicating that morae 5 and 6 are at the same level (L). This made us think it might be better to perform recognition with two mora pitch level recognizers that recognize with these local pitch levels. Doing this, there will be less F0 overlap for the different two mora combinations. To illustrate, with the AM based recognition, in /shokubakarano/, the /shoku/ will be recognized as LL despite /shoku/ having a rising pattern and /rano/ having a falling pattern. Therefore, we would like to have /shoku/ recognized as LH and /rano/ recognized as HL. For this method, the two-mora level transition combinations to recognize will be HL, LH, and Level (no change).

Thus, while the old method will perform the recognition...
Conducted experiments to construct equations to deal with F, R-F, in Section 4, we will discuss the classification algorithm. Then, pair. In the next section, we discuss the experiments for coming rise (F), rise then fall (R-F), fall then fall, fall then rise, or monotone then fall pattern relative to the F0 at the end of the first mora, the selection rate was different than that of a word with solely a falling pattern, even if the de-

F-R, and M-F patterns. We broke this experiment into two parts: one to derive equations for the F pattern for the second mora of the word and another for other mora in the word, and the other to obtain equations for the R-F, F-R, and M-F patterns. For this experiment, we used 18 subjects who accessed a program via the Internet for the listening test. The subject had the option to playback the sample or return to the previous sample. The experiment conditions are summarized in Table 1.

For the first part of the experiment, we recorded one Japanese speaker pronouncing a word with an HL transition on the last two morae for a type N-1 accent with the correct accent and with an incorrect type 0 accent. Thus, if the original morphed samples were LHHL (correct) and LHHH (incorrect), we can obtain the F0 contours in between the two samples and determine the threshold for the correct pattern for the final two morae (HL) and incorrect pattern (HH), and thus get the probability for HL. We then morphed the sample that has the correct accent with the incorrect accent to obtain samples with pitch contours for the mora in between the two. Then the subjects, speakers of Tokyo Japanese, listened and judged whether or not the morphed sample had the correct or incorrect accent. As there are not a lot of accent minimal pairs for 4 mora or 5 mora words, we chose this method rather than using minimal pairs.

For the second part of the experiment, we created three sets of words, all of which were accent minimal pairs, to have F-R, R-F, and M-F patterns on the 2nd mora of the two mora pair respectively. For this, there will be two variables, whereas the equation obtained from the first part has only one. An example diagram illustrating how the R-F pattern was manipulated is in Fig. 3. The rise in that is variable among the different sets, while the fall is variable within the different sets. For each set, we varied the amount of fall for the F-R pattern, rise for the R-F pattern, and the length of the monotone for the M-F pattern across the set relative to the F0 at the end of the prior mora. Then for each word within the sets, we created multiple samples for that word varying the amount of rise for the words in the F-R set, fall for the words in the R-F and M-F sets incrementally relative to the pitch at the end of the previous mora similar to [10]. For this part of the experiment the learner selected which word of the minimal pair they had heard.

3.2. Results

For the first part of the experiment, the graph of the selection rate of the morphed data for cases with only a fall is shown in Table 1: Conditions for the perceptual experiments

<table>
<thead>
<tr>
<th>Subjects</th>
<th>18 Tokyo Japanese Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Part</td>
<td>Fall Equation</td>
</tr>
<tr>
<td>1 Japanese speaker</td>
<td></td>
</tr>
<tr>
<td>Synthesized by morphing with STRAIGHT</td>
<td></td>
</tr>
<tr>
<td>2nd Part</td>
<td>R-F, F-R, and M-F Equations</td>
</tr>
<tr>
<td>Synthesized by pitch manipulation with Praat</td>
<td></td>
</tr>
<tr>
<td>Selection</td>
<td>Which word of the minimal pair was heard</td>
</tr>
</tbody>
</table>
For the fall-rise-fall pattern, the system first checks the HL probability for the previous mora. If the probability is less than 50%, the system checks the probability for the fall-rise pattern, and performs recognition like in the rise-fall case above. The rise-fall-rise pattern classification was done similarly, first checking the HL probability for rise-fall and then the HL probability for the fall-rise.

4. Classification Experiment

4.1. Experiment Set-up

To test the rule-based method, classification was performed with the method outlined in Section 3. The testing was done on 100 sentences from two different corpora, JNAS (a continuous speech corpus with a wide variety of dialects and speaking styles) and JRF (a corpus containing speech by non-native Japanese speakers), since the purpose of this research is CALL system development.

For the labeling, two Japanese native speakers labeled each two mora pair as LH, HL, or Level. The 200 sentences were divided into two sets, each set composed of 50 sentences from JNAS and 50 sentences from JRF. Each labeler labeled one set. As the labeling was quite difficult and time-consuming, we were unable to have them label both sets. The experiment conditions are listed in Table 2. F0 was extracted by Praat and the forced alignment was done with Julius in order to align the F0 by mora.

4.2. Results

The tables showing the agreement for the recognition of the two mora pitch level units between the labeling data and the results
rule-based methods (RB) are given in Table 3. The agreement for LH was higher than the other two and Level had the lowest agreement rate for both the JNAS and JRF datasets. The major reason for this is Level can be easily confused with both LH and HL, whereas HL and LH are not easily confused. Also, for Level there is often a slightly falling F0 contour, which makes distinguishing between it and HL rather difficult. Also, as we did not have equations to determine LH probability based on perception, the classification may be slightly inaccurate. For HL, in the perceptual experiments, there was only near 100% agreement among subjects in the perceptual test for cases where there was a large fall, around .1 log10 Hz. Therefore, with data from only one labeler this level of agreement might be expected. For JRF, there was a large drop in the agreement rate for HL, this is likely because it being non-native speech, it is more difficult to distinguish between HL and Level. The results based on the AM model are shown in Table 4 with the rates being the agreement rate at the mora level, that is the rate of agreement of a particular mora being H or L. The results for the JRF corpus were slightly worse than those for JNAS. This can be attributed to the wider variety of F0 contours and increased difficulty of labeling and aligning non-native speech. The agreement rate for the JNAS corpus was also, slightly better than what we presented with the prior method [9].

4.3. Discussion

Though the results appear slightly insufficient to use in a CALL system, on comparing the results with the actual F0 contour using Praat, the results appeared to be a lot better for the method than it would seem from the agreement rate. The labeling for this was quite difficult especially because patterns such as LH are only an intonation difference. Also, HL and Level are often hard to distinguish, especially in cases of a long vowel (two mora vowel). For the conversion to the AM labeling, the results appeared only slightly better than for the previous research. However, on comparing those results with the F0 contour and these results with the F0 contour it appears as though these methods did slightly better for that as well. Hence, it is possible the results based on the perceptual experiment are more reliable than the hand-labeling. Upon implementing the CALL system, it will be necessary to have a subjective evaluation performed to confirm this.

The main weaknesses for this method were F0 estimation errors and alignment. The reason it did not do well for these is because as one of the targets, the average of the last two values for the mora was chosen, which are prone to F0 estimation errors. Also, from our perceptual experiments, we found the areas where the transition from vowel to consonant occurs either do not influence accent perception. To increase the robustness of this method, it will be necessary to include that in the calculations as well as include more processing to detect F0 estimation errors. Coming up with a method to perform forced alignment better will also likely improve results. Overall, though the level of agreement between the labeling and the results was only around 70%, it may be sufficient to use in a CALL system. Therefore, it is necessary to perform subjective evaluations and have the data labeled by more individuals to see if the performance is adequate.

5. Conclusion

In this paper, a rule-based algorithm for multi-level pitch pattern identification for a Japanese pitch accent acquisition CALL system was presented. The purpose is to provide flexible visual feedback to the learner and also detect errors in the pitch accent. From this pattern, it is possible to determine whether the pitch kernel and the accent tone are in the correct position, and if there is more than one kernel. It can also be used to represent the pitch for a word with multiple levels or with just the two levels for the word. Since learners may produce patterns that do not fall into the Tokyo Japanese accent type set, a robust method is necessary. The proposed method is more robust than previous methods and can handle errors that methods based on the accent types cannot handle. Also, it was based on perceptual tests in order to more accurately reflect Japanese perception. Though the agreement rate of the algorithms with the hand-labeling did not appear adequate from the results, on further inspection comparing the F0 contour to the pitch patterns identified with the algorithm, it appeared as though the results may actually be sufficient for a CALL system. To determine if the method is adequate, it will be necessary to perform more evaluations.

For future work, we aim to use a method to detect likely F0 errors so that they will not be used in the calculations. In addition, we plan to develop equations to recognize the probability of LH based on perceptual experiments. Also, we intend to conduct more perceptual experiments on the rise-fall and fall-rise by varying the position of the start of the rise for the former and the start of the fall for the latter to get more precise equations for them. We then plan to implement these into a CALL system and perform subjective evaluations to determine if the classification results are sufficient.

6. References