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 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 66% 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 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 and /rano/ will also need to 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

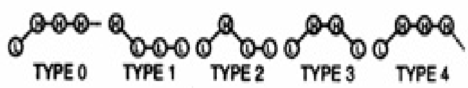

Figure 1: *Possible accent types for a four mora word*

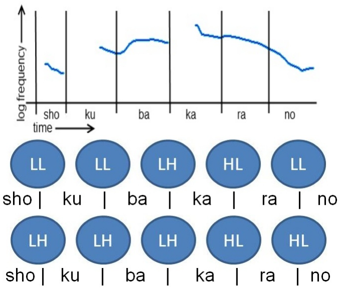

Figure 2: *Top: F0 Contour for "shokubakarano"*
Middle: AM Based Two Level Training/Recognition For "shokubakarano." Consists of the units LL, LH, HL, HH. In this case the pattern is LLLHLL.
Bottom:Local Pitch Transition based Recognition. Only consists of LH, Level, and HL transitions and represents multiple levels

based on the AM model, the new method will perform it with the multilevel method. These are in the middle and bottom of Fig. 2. From that figure, it is easy to see that multi-level transitions are represented. By noting that the end of the LH sequence at the beginning is H and the start of the drop in pitch level ('ka') is the transition to L, it should be apparent that it is possible to convert the pattern to the AM pattern as well, since the prominent morae will be high (in this case only 'ka') and the other morae will be L. The transition to high can be found at the end of a sequence of LHs and the transition to low can be found with an HL. It can be seen from this figure as well that it is easy to judge if the kernel is in the correct position or not and whether the start of the initial rise is in the correct position, which can be used for error detection. The proposed method will make it easy to determine pitch level relationships between words as well and, thus, detect the phrase tone, H-, which can be beneficial for teaching inter-word intonation. Also, by doing this, it will be possible to abstract the multilevel pitch pattern of the word. Generating a visible multilevel pitch pattern could also make it easier for the learner to understand the global pitch pattern.

2.3. Accent Identification Features

In the previous experiment we conducted in [9], it was unapparent which features should be used for recognition so we conducted a perceptual experiment in [10] to determine the features to use and develop a rule-based method for recognition. In those experiments, we adjusted the F0 of words that have accent minimal pairs using PRAAT to analyze the mechanism of accent perception quantitatively. An accent minimal pair for a two mora word will be a contrast between an HL word and a word that is either HH or LH.

From that perceptual experiment we conducted, it seemed as if only the F0 value at the end of the the first mora of the pair was important and the F0 values for the consonants were not. Also, it appeared as though the position of the mora in the word had an effect on how much of a pitch fall there should be to produce a pitch level change. If the second mora had a rise 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 degree of fall was the same. Hence, in order to create a rule-based method more studies into perception are needed. To do this it is necessary to come up with equations to deal with different cases that occur such as a fall (F), rise then fall (R-F), fall then rise (F-R), and monotone then fall (M-F) on the 2nd mora of the pair. In the next section, we discuss the experiments for coming up with those equations to develop a rule based method. Then, in Section 4, we will discuss the classification algorithm.

3. Rule-based Method Development

3.1. Perceptual experiment overview

To derive the equations for the cases mentioned above, we conducted experiments to construct equations to deal with F, R-F,

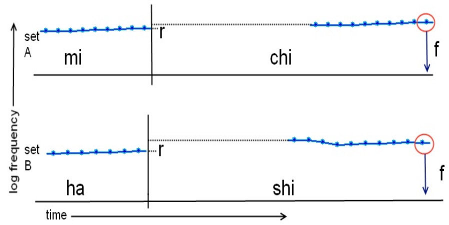

Figure 3: *Graph showing examples of the sets used for the derivation of the R-F equation. 'r' indicates the degree of rise for the rise variable and 'f', which was manipulated within each set, shows it for the fall variable. The F0 at the start of the second mora was fixed in the manipulation.*

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*

Subjects	18 Tokyo Japanese Speakers
1st Part	Fall Equation 1 Japanese speaker Synthesized by morphing with STRAIGHT
Selection	Accent Correct or Incorrect
2nd Part	R-F, F-R, and M-F Equations Synthesized by pitch manipulation with Praat
Selection	Which word of the minimal pair was heard

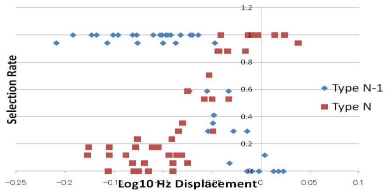

Figure 4: *Selection rate of the accent being correct for Type N-1 and Type 0 with varying displacements for the fall*

Fig. 4. The x-axis shows the displacement of the fall from the F0 at the end of the previous mora and the y-axis shows the selection rate for the accent being correct for words greater than two morae. From this the sigmoid approximating the selection rate is determined using the function

$$P_{HL}(x) = \frac{e^{a(x-IP)}}{1+e^{a(x-IP)}}SP \quad (1)$$

where x is the amount of fall, a is the rate of change at the inflection point, IP is the inflection point, and SP is the maximum value that is approached. The equation of the probability for two mora words being HL was obtained in the same way.

For the cases of R-F, F-R, and M-F, as the results had large ambiguities for small falls, we decided on an equation that could approximate the effect of the rising pitch on the perception of the fall would be sufficient. In the case of an R-F pattern being on the 2nd mora, To get the equations to approximate the selection rate for an arbitrary rise and fall, we first graphed (Fig. 5) the results for different rises where the displacement of the fall relative to the F0 at the end of the last mora were roughly equal for two displacements, -0.13 and -0.06 log10 Hz with the y-axis being the selection rate and the x-axis being the displacement of the rise from the end F0 value of the previous mora.

From this graph we were able to approximate functions for both displacements that can estimate what the selection rate for an arbitrary rise would be using linear regression. These equations are shown in the graph. Thus, to determine what the selection rate would be for an arbitrary rise and an arbitrary fall, first the selection rate would be in the case of that rise for the displacement of -.06 log10 Hz and -.13 log10 Hz is determined. This can be calculated with the linear equation

$$P_{HL}(d_i) = 1-(mx+b) \quad (2)$$

where in this case x is an arbitrary rise, m and b are calculated via linear regression, and i is either the displacement of -.06 log10 Hz and -.13 log10 Hz. This is subtracted from 1 because the graph shows the selection rate for non-HL. From there, assuming that at a rise of 0 log10 Hz and fall of 0 the selection rate would be 0% for HL, the linear equation between between -.06 and -.13 log10 Hz are calculated, if the fall is smaller than -.06 log10 Hz, the linear equation between 0 and -.06 log10 Hz was used. The linear equation can be represented by

$$P_{HL}(x) = \frac{(P(d_a)-P(d_b))}{(d_a-d_b)}(x-d_a)+P(d_a) \quad (3)$$

where x is an arbitrary fall and d_a is the fall displacement -.13 log10 Hz and d_b is -.06 log10 Hz. These linear equations allow for the approximation of the equations for HL for R-F. The graph for R-F shows a low correlation between the equation obtained by linear regression and the data. This is likely because we did not consider the start of the fall for the R-F pattern. To obtain a better equation, it will be necessary to take the start of the fall into account.

For the cases of the fall-rise, we derived equations in a similar fashion to method for obtaining them for the R-F pattern. For the M-F pattern, though, as there is no initial rise or fall, we based the linear equations on the percentage into the second mora at which the fall began and the amount of fall.

3.3. Rule-based Method Classification

For the rule-based method classification, we came up with seven different cases for the second mora F0 contour: rise, fall, rise-fall, fall-rise, rise-fall-rise, fall-rise-fall, monotone-fall, in which rise means the F0 rises above the last value of the F0 for the previous mora, fall means it falls below, and monotone means little to no change from last value of the previous mora. Since the LH transition does not differentiate words in Tokyo Japanese, we could not perform perceptual experiments with minimal pairs like we could for HL. Thus, we based the judgment of rise or not on the labeling data. From the labeling data it appeared as though a rise of 0.01 log10 Hz was sufficient to produce a rise in pitch level.

In order to do this, the local targets were extracted for the two morae. First the last two F0 values of the first mora of the two mora set were averaged, and then the targets for the second mora are extracted. The targets extracted for the second mora were the maximum and minimum values for the mora with two maxima being possible if there was a minimum in between the two and vice versa, and the start of the fall in the event that the beginning F0 values for the second mora are within 0.01 log10 Hz of the first mora. Also, the average of the first values and the average of the last F0 values for the second vowel of the pair were considered in the target candidate selection. All of the targets were subtracted from the average of the last two F0 values of the first mora of the pair. The F0 values for the consonants were disregarded except for the nasal /N/.

To determine the probability of HL we made use of the equations derived in Section 3. For the cases other than rise-fall-rise and fall-rise-fall, the system checks the HL probability. The pair is classified as HL if the HL probability is greater than 50%. Otherwise, it is classified as LH if the pitch rises above 0.01 log10 Hz of the previous mora or Level if it does not. Here, fall and rise are used to mean that the pitch rises above or falls below the previous target. Thus, we combined the equations we derived for the fall-rise and rise-fall patterns.

For the fall-rise-fall pattern, the system first checks the probability of HL for the fall-rise. If the probability is less than 50% the system checks the HL probability for the rise-fall 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.

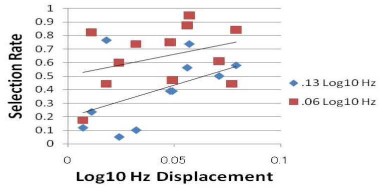

Figure 5: *Selection rate of the type 0 word (non-HL pattern) with the x-axis showing the displacement of the rise for the rise-fall pattern with fall displacements of -.06 log10 Hz and -.13 log10Hz.*

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

Table 2: *Conditions for the classification experiment*

Corpus (Native) Samples	JNAS 100 speakers, 100 sentences
Corpus (Non-native) Samples	JRF 34 speakers, 100 sentences
Forced Alignment	Julius (Japanese acoustic model)
F0 Extraction	Praat
Labelers	2 Japanese natives

Table 3: *Agreement rates for the local two mora pitch level*

Corpus	LH Rate	HL Rate	Level Rate
JNAS	0.85	0.77	0.62
JRF	0.81	0.6	0.58

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 that 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.

Table 4: *Agreement Rate for Mora Level after conversion to autosegmental-metrical (AM) notation*

JNAS Agreement Rate	JRF Agreement Rate
.81	.76

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

[1] Isomura, K, "The Present State of Japanese Accent Education Overseas", The Society For Teaching Japanese as a Foreign Language, 211-212, 2001.

[2] Nakagawa, C. "Tokyogo akusento shutoku junjo to gakushusha no ishiki", Waseda University Japanese Language Education Center, 2002. (Japanese)

[3] Kawai, G. and Ishi, C, "A System for Learning the Pronunciation of the Japanese Pitch Accent", Proc Eurospeech '99, 177-181, 1999.

[4] Ishi, C., Minematsu, N. and Hirose, K., "Identification of Japanese accent in continuous speech considering pitch perception", Technical Report of IEICE, SP2001-48, 23-30, 2001.

[5] Kumagai, Y., Yoshida, K. and Jouji, M., "On a Decision Method of Accent Type for Japanese Learning", IPSJ SIG Notes 99, 22-30, 1999.

[6] Pierrehumbert, J.B. and Beckman, M.E.,"Japanese tone structure", Cambridge, Massachessetts: MIT Press, 1988.

[7] Shport, I.A. and Guion, S.G., "The Effect of Segmental Structure on F0 Patterns of Words in Tokyo Japanese", Journal of the Phonetic Society of Japan, 4-16, 2008.

[8] Kanda, T. "Nihongo akusentono hyoukini kansuru kousatsu - Sansenshikihou" (Japanese), Bulletin of Gifu Women's College, 51-58, 2003.

[9] Short, G., Hirose, K., Yamada, T., Minematsu, N., Kitawaki, N., Makino, S., "Pitch pattern recognition of isolated words for the development of a Japanese language CALL system", Proc. Oriental COCOSDA, 2010.

[10] Short G., Hirose, K. and Minematsu, N., "An analysis on the perception of pitch level changes for Japanese words", Technical Report of IEICE, SP2010-128, 79-84, 2011.