Mora Pitch Level Recognition for the Development of a Japanese Pitch Accent Acquisition System

Greg Short[1], Keikichi Hirose[1], Takeshi Yamada[2], Nobuaki Minematsu[1], Nobuhiko Kitawaki[2], Shoji Makino[2]
[1]Univesity of Tokyo, [2]University of Tsukuba
[short,hirose,mine]@gavo.t.u-tokyo.ac.jp, [takeshi,kitawaki]@cs.tsukuba.ac.jp, maki@tara.tsukuba.ac.jp

Abstract

A language learning system built for guiding a student on how to pronounce words in a second language must provide meaningful feedback while locating the learner's errors with a high accuracy. We propose a method to detect pitch accent errors in the speech of learners of Japanese. In previous methods, Tokyo accent type recognition of the learner utterance was focused on. However, the learner may produce pitch level patterns that do not exist in these accent types. This paper covers a proposed technique for identifying mora pitch level to recognize all patterns. Employing a 2 mora model trained on the continuous speech corpus JNAS, this method identifies pitch level for contiguous two mora unit sets. Then it determines the most likely combination of these units to find the pitch level for each mora of the word. Through this method, we achieved an 80.5% correct mora pitch level identification rate.

1. Introduction

In recent years, the number of international students studying in Japanese institutes of higher education has been rapidly increasing. This number is scheduled to rise tremendously in the coming years from its present number of around 140,000 up to 300,000 [1]. For those already residing in Japan and those intending to study in Japan, it is essential that there be adequate education in the Japanese language. It is often a difficult task for them to find employment and adapt well to life in Japan without a strong proficiency in the Japanese language, including Japanese pronunciation. That being said, there is little instruction on pronunciation in many Japanese language classes. Ideally, a language learner would have a native teacher well-trained in pronunciation teaching along with a lot of class-time for pronunciation practice. Due to class-time constraints and teaching priorities, this is almost always not the case, however [2]. To make up for this situation and also provide supplementary tools for language learners, a number of Computer Assisted Language Learning (CALL) tools have been developed for pronunciation [3][4].

CALL systems train learners in many different aspects of language learning, one of these being pronunciation. This has been made possible by the increased development in speech processing technology, which provides learners with tools to have his or her pronunciation errors located and give guidance on how to fix those errors. These tools have gained a lot of ground in terms of accurate detection of errors. In spite of this, a great deal of improvement is still necessary for accurate error detection.

In Tokyo Japanese, three perceptual pronunciation features are used by natives to differentiate words: phoneme, phoneme duration, and accent. Though pronunciation as a whole is generally given very little attention, there is especially little class time devoted to pitch accent education. Also, for many learners certain pitch patterns are difficult to produce without proper instruction [5]. Thus, it is necessary for them to have assistance in learning the pronunciation of the Japanese accent.

There has been research done on detecting accent errors but the results of this research are still not satisfactory for a CALL system. Therefore, in this research, we are focusing on the Japanese lexical accent error detection. In past research, errors were detected by identifying the accent type of the learner's utterance [6][7]. In those methods, features representing the accent types that occur in the Tokyo dialect of Japanese were compared with features extracted from the F0 contour of the learner and the best match was chosen to be the accent type. The accent types found in the Tokyo Japanese dialect, however, only comprise a subset of the possible mora pitch level patterns that a speaker may pronounce a Japanese word with. Due to language transfer, the learner may mispronounce words with pitch level patterns not usually found in the Tokyo dialect of Japanese [8]. Since it is essential to detect these errors, we have developed a pitch level recognition method that identifies the pitch level for each mora with templates consisting of two morae [9].

2. Japanese Accent Error Detection

2.1. Japanese Accent Overview

Japanese words are made up of timing units termed morae. One syllable is either made up of one or two morae, and a double morae syllable is roughly twice the length of a single morae syllable. Each mora in Japanese words has either a high pitch (H) or a low pitch (L) relative to the pitch of other morae in the word. The position of the drop from high pitch to low pitch in a word, the accent nucleus, is the distinguishing factor for the different accent types.

In the Tokyo Japanese dialect, a word with N morae can have up to N+1 accent types. The accent type is determined by the position of the accent nucleus in the word (i.e. the last mora perceived high before a drop to low pitch). The possible accent types a four mora word may have are listed below in Fig 1 with the number of the accent type being the location of the nucleus in the word. These accent types are further broken down into three type groups with type 0 being the "heiban" (flat) type because it lacks an HL transition, type 1 as the "atamadaka" (head-high) type due to the head mora being high and the rest being low, and accent types with a nucleus that is not the first mora as being the "nakadaka" (mid-high) types because there is a rise at the beginning and then a fall in pitch level. The difference between a type N word, where N is the total number of morae, and a type 0 word is the pitch level of a type N word drops with the addition of a particle at the end of the word and a type 0 word stays high in pitch level [10]. The word plus particle combination is called an accent phrase.

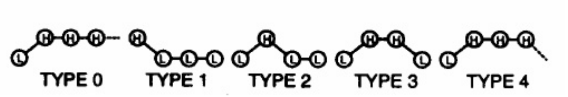

Fig 1– Potential accent types for a four mora word

2.2 Problem Overview

For a CALL system, it is said to be important for the system to accurately detect learner errors. Thus, it is our goal to develop a method to achieve accurate error detection in the pitch accent. In the past, methods have been proposed to detect learner accent type errors. These methods were developed for the recognition of Japanese accent types. However, because the pitch patterns common in the learner's native language may differ from those in the Tokyo Japanese dialect, the learner may produce pitch patterns that do not fall in the Tokyo accent type set. This phenomenon is called language transfer, and it leads learners to speak L2, Japanese, with features of L1, their native language [4]. Because of this phenomenon, previous methods are insufficient for successfully detecting learner accent errors. To provide an example of language transfer, the F0 contour of "kamikaze", a type 0 word, pronounced with the F0 contour shown in Fig 2 has the pitch level pattern in Fig 3. This will not be properly recognized by accent type recognizers because it is a pattern not in the Tokyo dialect accent type set. In the Tokyo dialect, an arbitrary word of N morae can potentially have N+1 different accent types. However, there are up to N^2 possible pitch level patterns it can have.

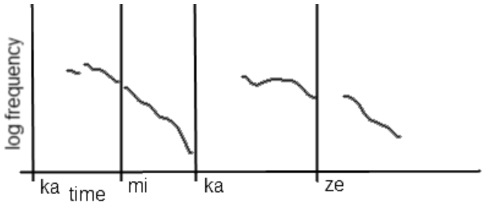

Fig 2– Pitch contour for "kamikaze"

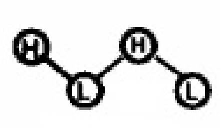

Fig 3– Pitch level pattern of "kamikaze"

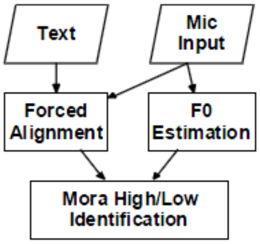

Fig 4– Process flow for recognition

2.3. Proposed Method

On account of the reasons outlined above, we have come up with a method to identify the pitch level of each mora in order to account for all N^2 possible pitch level patterns to accurately detect errors in the learner's pitch contour. In this section, a corpus based method making use of Japanese perception of pitch level for morae is proposed for recognizing all pitch level patterns will be discussed.

In order to perform the recognition, the process listed in the flow chart Fig 4 is carried out. Because the Japanese lexical accent is mora-based, first forced alignment is used in order to detect the boundaries for each mora. The F0 for the sound wave of the utterance is then extracted at short-period time frames. In Japanese, vowels sometimes undergo devoicing when lying between two unvoiced phonemes. Unvoiced morae have no pitch, but because it is possible to interpolate the perceived pitch level of the mora based on the pitch level of the surrounding morae, these morae were not accounted for in the feature extraction stage.

Following the segmentation of the F0 contour by mora, feature extraction is carried out. First each mora in the set is divided into two parts. Then for each mora F0mean and ΔF0mean are computed. F0 mean for each partition was calculated by

$$\mathrm{F0mean} = \frac{1}{n_p} \sum_{i=0}^{n_p} \log \mathrm{F0}_{i_y} \qquad (1)$$

where $\mathrm{F0}_{i_y}$ is the y value of the i^{th} value and n_p is the number of F0 values for partition j calculated by

$$n_p = \frac{n_T}{2} \qquad (2)$$

with n_T being the total number of F0 values, and ΔF0mean for each partition was found by

$$\Delta\,\mathrm{F0mean} = \frac{1}{n_p - 1} \sum_{i=0}^{n_p-1} \frac{\log \mathrm{F0}_{(i+1)_y} - \log \mathrm{F0}_{i_y}}{\log \mathrm{F0}_{(i+1)_x} - \log \mathrm{F0}_{i_x}} \qquad (3)$$

The F0mean values for each partition of each mora are all normalized by the F0mean of the second partition of the first mora. Also, the max mean log F0 value for all moras for the accent phrase is determined and the F0mean values for both morae are normalized by that value and used in the feature vector.

The recognition part of the process is then performed. In order to recognize all pitch level patterns for words, we have come up with a method based on the perception of high and low pitch levels that breaks a word into subunits containing two contiguous morae, whereby each set overlaps its neighboring set to the left by one mora. Thus, there will be N – 1 two mora units, where N is the number of morae in the word and 1 is the number of morae in each unit subtract one. The probability each two mora set is LL, LH, HH, or HL is then determined. Out of all the possible mora pitch level patterns for the word, it determines the combination with the highest probability to be the pitch level pattern of the utterance. This process carried out with two mora units is illustrated in the example in Fig. 5.

As mentioned above, unvoiced morae were removed from the recognition process so to handle the interpolation of unvoiced morae, we use rules based on observation. Assuming there are three contiguous morae, A, B, and C respectively, if mora A and mora C are the same level then mora B is also that level. If they are at different levels then B is low, unless B is the latter half of a long consonant and A is high, in which case it is high. If A is the start of the phrase, then A is low. If B is the end of the phrase it is the same level as A. An example of a word with an unvoiced mora is shown in Fig. 6.

For this, the third mora will be interpolated as H and the pattern will be LHHH.

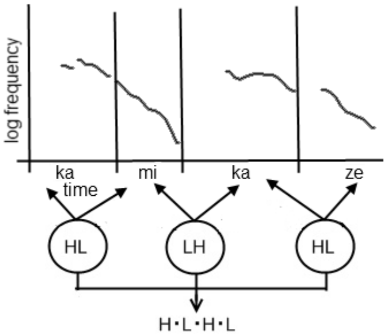

Fig 5– Example of recognition process

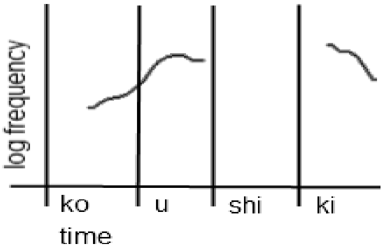

Fig 6– Example of word with unvoiced mora

3. Experiment

3.1. Experiment Overview

For training and testing this method, we made use of the JNAS corpus containing Japanese speech read out of newspapers by native speakers of various ages and regions. We had a native Japanese speaker do the high/low labeling for each mora in every word. This database contains speakers of several different dialects so a lot of the speech did not follow the pitch patterns listed in Fig. 1. In continuous speech, it is common for types other than type 1 to start with high pitch rather than low pitch as listed in Fig 1. Some of the accent phrases from the speech also contained two high/low transitions.

To perform the recognition, a 2 mora multi-dimensional Gaussian probability model was employed. We broke the data into the training data set and test data set at a ratio of 3:1 and because the amount of data was slightly insufficient we shuffled the two data sets ten times and averaged the results of the ten different sets.

To assess the performance of the proposed method, we analyzed the results based on the following criteria: overall results at the mora level, results for the three accent type groups (“heiban”, “atamadaka”, and “nakadaka”) and for the “other” patterns that do not follow the patterns listed in Fig 1. Also, we will look at the results of accent phrases consisting of two mora words up to accent phrases of words with nine morae. Then, we will discuss the changes being made to the LH transition and HL transitions in the accent phrase, since there is usually at most only one of each of these transitions in an accent phrase. Analyzing the displacement, deletion, and insertion of these transitions will help to understand the kinds of errors being made. The experiment conditions are shown in Table 1.

Table 1. Experiment conditions

Speech Data	JNAS 100 sentences
Training/Recognition	3:1 training to recognition ratio
Speakers	50 male/50 female
Sampling Rate	16kHz
F0 Extraction	Praat
Alignment	Julius 4.1.4

3.2. Experiment Results

The overall results and the results of each type group and other patterns not listed in Fig. 1 are shown in Table 2. Overall, there was only an 80.5% recognition rate at the morae level. For the different Tokyo accent types, the “heiban” type was clearly higher than the “atamadaka” and “nakadaka” types and the “other” patterns also showed a more accurate recognition rate than the “atamadaka” and “nakadaka” types. Many of the “atamadaka” words were recognized as having a high-pitched second mora rather than dropping in pitch level on the second mora. This is most likely largely due to there being several different patterns for a high/low transition. For the “atamadaka” type, the first mora often has a rise in pitch and the second mora has a fall in pitch. However, in the high-low transition for “nakadaka” types, the high-pitched mora often is perceptually monotone and the low-pitched mora has a fall. This is illustrated in Fig. 7, the F0 contour for “nyuusu” (type 1) and in Fig. 8, the F0 contour for “mazushii” (type 3). For the “other”

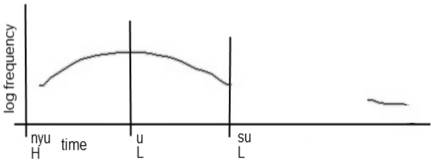

Fig 7– F0 contour of an “atamadaka” type

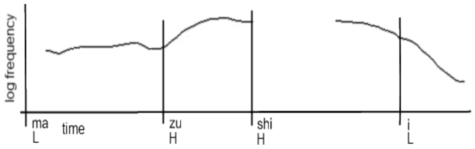

Fig 8– F0 contour of a “nakadaka” type

patterns, a large number of them did not include an “LH” transition, for which this method performed poorly so the recognition rate was higher. Many of these patterns were similar to the “heiban” and “nakadaka” types only without an LH transition.

The LH transition displacement rates are given in Table 3. Only around half of the accent phrases had the LH transition recognized in the correct position. Most of the errors came due to deletion. The first two morae of the accent phrase were often recognized as HH rather than LH. This is possibly due to an LH transition being difficult to define. It is said the accent nucleus (HL transition) is what differentiates two different Japanese words, but the position of the LH transition is more an intonational feature [11]. It can be seen that there was only a very small amount of insertions of LH to words that already contained an LH transition from Table 4. Also, around 6% of accent phrases without LH transitions had an LH insertion.

As seen in Table 5, for the HL transition or nucleus, only around 41% were labeled on the same two morae as the hand labels. Around 27% of them were displaced by one mora and 30% of the HL transitions were deleted. The displacement by one mora can be explained by the reason given above and illustrated in Fig 7 and Fig 8, that different pitch contour patterns can produce the same pitch level pattern. In Table 6, it can be seen for insertions, there were a very small number of accent phrases with HL transitions, 0.5%, which already had an HL transition. A few accent phrases, 6.9%, without an HL transition had an HL insertion.

Table 2. Total recognition rate and recognition rate for difference accent types

Type	Recognition rate (%)
Heiban	87.3
Atamadaka	69.3
Nakadaka	73.9
Others	84.8
Total	80.5

Table 3. LH transition displacement

Displacement (# morae)	LH Transition Displaced (%)
0	49.8
1	1.4
2	1.1
3	0.2
Deleted	47.5

Table 4. LH insertion

Word Type	LH Insertion Rate (%)
LH Word	0.9%
Non-LH Word	6.1%
Total	3.4%

Table 5. HL transition displacement

Displacement (# morae)	HL Transition Displaced (%)
0	41.1
1	27.0
2	1.4
3	0.2
Deleted	30.2

Table 6. HL insertion

Word type	HL Insertion Rate (%)
HL word	0.5
Non-HL word	6.9
Total	4.0

3.3. Discussion

Overall, the results appear insufficient to use in a CALL system for continuous speech. The results for the "heiban" type and the patterns listed under "other" were slightly inadequate and the results for "atamadaka" and "nakadaka" types were especially insufficient. The biggest problems were the deletion of HL and LH transitions and the displacement of HL transitions. Transition insertions only occurred around 4% of the time. As mentioned above, this is likely because a variety of pitch contour patterns can constitute the same pitch level pattern. Also, as this corpus contains continuous speech, it is harder to accurately detect the boundaries with a forced alignment so there may be some degradation in scores due to boundary detection. There were a wide variety of speakers of various regional dialects so the number of kinds of pitch contours in this corpus is much greater than a corpus of speech by announcers as well. F0 estimation errors were also likely responsible for some errors. For these reasons, a method based on the high/low perception of morae using a Gaussian vector model may be insufficient. It may produce better results look at the underlying rise and fall patterns and the rules for them to combine to form pitch level patterns and develop a method based rises and falls in pitch.

4. Conclusion

In this paper, a corpus-based method is proposed for the automatic detection of mora pitch level in Japanese words and accent phrases for the purpose of building a CALL system for training Japanese language learners in the Japanese pitch accent. In this method, Japanese words and accent phrases were broken down into two mora units, the best pitch level pattern for each pair was found, and then from these the pitch level pattern combination with the highest likelihood was obtained. This method achieved overall results of around 81% for correct mora pitch level identification. These results appear inadequate for developing a CALL system using this method. It is possible a method that is not solely based on perception of high and low pitch, but one that also accounts for the variety of pitch contour patterns that make up pitch level patterns will improve the results.

In the future, we plan to conduct pitch level identification experiments on non-native Japanese speech to do determine how well this method will work for identifying the pitch level patterns for language learning purposes. Also, we plan to further look into how to achieve better results. Because various rise and fall patterns can constitute the same pitch level patterns, we will look at developing a method to detect rises and falls and the rules for combining these to produce different pitch level patterns. From there, we plan to develop a pitch accent acquisition system for Japanese and have it tested from an educational standpoint.

10. References

[1] Ministry of Education, "300,000 International Student Plan," 2003.

[2] K. Hirose, "Accent Type Recognition of Japanese Using Perceived Mora Pitch," International Symposium on Tonal Aspects of Languages. 2003.

[3] G. Kawai, C.T. Ishi, "A System for Learning the Pronunciation of the Japanese Pitch Accent," Proc Eurospeech '99, 1999.

[4] A. Neri, C. Cucchiarini, H. Strik, "Feedback in Computer Assisted Pronunciation Training: When Technology Meets Pedagogy," Proceedings of CALL Conference, 2005.

[5] Isomura, Kazuhiro, "Kaigai ni okeru Nihongo Akusento Kyouiku no Genjou," Society for Teaching Japanese as a Foreign Language Fall Meetings, 2001.

[6] C. Ishi, N. Minematsu, K. Hirose, "Identification of Japanese accent in continuous speech considering pitch perception," The Institute of Electronics, Information and Communication Engineers, 2001.

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

[8] M. Sayora, "Roshiago wo bogoto suru nihongo gakushuusha no intoneeshon," Japanese Education and Speech Conference, 2004.

[9] G. Short, T. Yamada, N. Kitawaki, S. Makino, "Japanese Lexical Accent Recognition for Language Learning Purposes," Acoustic Society of Japan Spring Meetings, 2010.

[10] NHK Broadcasting Culture Research Institute, "Japanese Accent Dictionary," 2005.

[11] S. Tanaka, “Accent and Rhythm,” Kenkyuusha, 2005.