PITCH PATTERN RECOGNITION OF ISOLATED WORDS FOR THE DEVELOPMENT OF A JAPANESE LANGUAGE CALL SYSTEM

Greg Short, Keikichi Hirose and Nobuaki Minematsu

University of Tokyo
short@gavo.t.u-tokyo.ac.jp

Abstract: A language learning system for guiding a student on how to pronounce words of 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 patterns outside of these accent types so it is necessary to be able to identify all possible pitch level combinations. This paper presents a technique to identify all patterns. Employing either a 2 mora model or a 3 mora model, this method identifies pitch level for contiguous two or three mora unit sets. Then it determines the most likely combination of these units to determine the pitch level pattern of the word. Through this method we achieved a 93.4% correct pitch level identification rate at the mora level.

1 Introduction

In recent years, Japan has undergone an increase in internationalization including the bringing in of more and more international students each year. The number of international students is scheduled to rise greatly in the coming years from its present state of around 140,000 up to 300,000 [1]. For these students, it is essential that there be adequate education in the Japanese language. It is often a difficult task to find employment for them without a strong proficiency in the Japanese language including Japanese pronunciation. That being said, in a large number of classes pronunciation is given very little class time. Ideally, a language learner would have a well-trained native teacher along with a lot of class time for pronunciation practice. This is almost always not the case, however [2]. To compensate 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 work with many different aspects of language learning, among them pronunciation. This has been made possible by the increased development in speech processing technology being able to provide learners with tools to locate his or her pronunciation errors and give guidance on how to fix those errors. While these tools have seen a great deal of progress in the past few decades, there is still a necessity for improvement in the area of accurate error detection of pronunciation features.

For pronunciation in Japanese, there are three features used to differentiate words: phoneme, phoneme duration, and pitch accent. Of these, surveys show that in many countries teachers cover the pitch accent very little. And for many learners certain pitch patterns are difficult to produce without proper instruction [5]. There has been research done on detecting accent errors but the results of this research are still not satisfactory for a CALL system. Thus, in this research, we are focusing on the Japanese lexical accent error detection. In the research carried out in this area in the past, errors were detected by identifying the accent type of the learner's utterance [6][7]. In those methods, the features representing the accent types that occur in the Tokyo dialect of Japanese were compared with features extracted from the learner's F0 wave and the best match was chosen to be the accent type of the learner's speech.

The accent types found in the Tokyo Japanese dialect, however, only comprise a subset of the possible mora pitch patterns that a speaker may possibly 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 bottom-up pitch level recognition method that identifies the pitch level for each mora with templates consisting of multiple morae.

2 Pitch Level Identification Method

2.1 Japanese Accent Overview

Japanese words are made up of timing units termed morae, units which are less than a syllable, but have roughly equal timing. 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. A drop from high pitch to low pitch in a word, the accent nucleus, determines the accent type of the word and different accent types differentiate words. 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 in pitch). The possible accent types a 4 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 types with type 0 being the "heiban" (flat) type because it lacks an HL transition, type 1 as the "atamadaka" (head-high) type because the head mora is high and the rest are 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 and fall in pitch level [9].

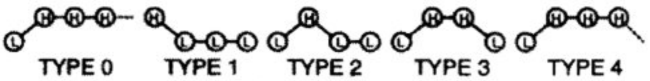

Fig 1 – Possible accent types for four mora words

2.2 Problem Overview

For the construction of a CALL system, it is said to be important for the system to accurately detect learner errors. So 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 focused were mainly developed for the recognition of Japanese accent types. However, it is not always the case that a learner will pronounce a word with one of the Tokyo Japanese dialect accent types because of interference of the features of the learner's first language [4].

The different spoken language features that exist between different languages result in language transfer when learning a foreign language. This causes the learner to speak L2, the language being learned, with features of his or her own language. On account of language transfer, the learner may possibly pronounce words with pitch level patterns outside of the Tokyo dialect of Japanese. For example, a speaker of American-English as a first language pronounced the word "kamikaze", a type 0 word, with the F0 wave found in Fig 2, represented in Fig 3 as a pitch level pattern. This will not be properly recognized by accent type recognizers because it is a pattern not in the Tokyo dialect accent type set. While in the Tokyo dialect, an arbitrary word of N morae can potentially have N+1 different accent types, there are up to N^2 possible pitch level patterns it can have and the learner might use.

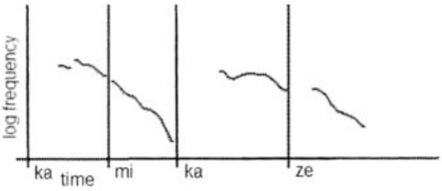

Fig 2– Pitch wave for "kamikaze"

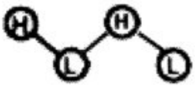

Fig 3– Pitch level pattern of "kamikaze"

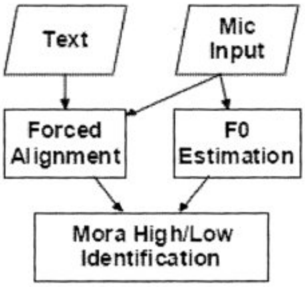

Fig 4– Process flow for recognition

2.3 Proposed Method

Due to the above to accurately detect errors in the learner's pitch wave, it is necessary to be able to identify all possible pitch level patterns. In this section, the method 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. After the above processing and the F0 segmented by mora boundaries is obtained, feature extraction is performed.

Next, the recognition stage is carried out. In order to recognize all pitch level patterns for words, we have come up with a method that breaks a word into subunits containing two or three contiguous morae, where each set overlaps by the neighboring sets by the number of morae in the set subtract one. Thus, there will be N – M units, where N is the number of morae in the word and M is the number of morae in each unit. This means that a four mora word will be divided into either two sets of three morae or three sets of two morae for the two mora model and three mora model respectively. Then, the pitch level pattern of each set is identified by labeling each mora low or high. Of 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.

The features used are illustrated in Fig 6. First, the mora is partitioned into four units. Then the log F0 mean of each partition is obtained and the end point of the log F0 linear regression line for the last partition is calculated. This process is repeated for every mora within the unit used (two or three mora unit) and inserted into a vector. Lastly, every value in the vector is subtracted by the end point of the linear regression line for the first mora to normalize the values in the vector.

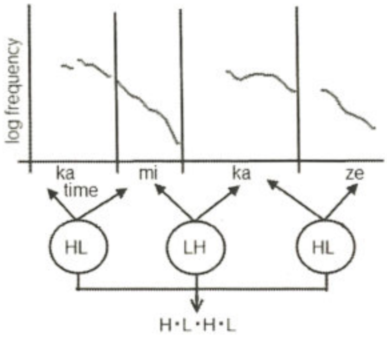

Fig 5– Example of recognition process

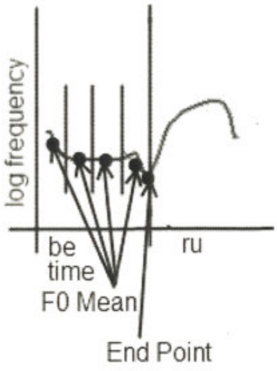

Fig 6– Log F0 based features

To identify the pitch pattern for words containing more than two morae we believed that it may produce better results to use three mora units. This is because some say there are actually three levels of pitch in the Tokyo dialect, even though in terms of accent, there are two [10]. However, in the Tokyo dialect the patterns HLH and LLH rarely occur within a word so training these two patterns is very difficult just with a Japanese corpus. Thus, to train these two templates, we resynthesized the pitch of around 80 samples of Japanese speech so that each of the samples would contain either or both of those two patterns.

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

Size	2 Mora Model Recognition Rate (%)	3 Mora Model Recognition Rate (%)
2	86.7	N/A
3	93.0	92.7
4	91.8	94.0
5	96.6	94.5
6	87.3	91.3

Table 4 – Mora recognition rates per size

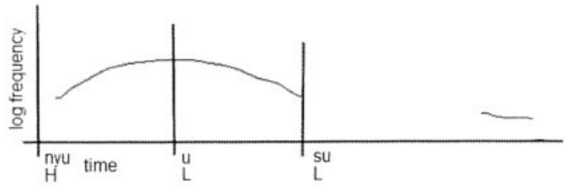

Fig 8 – F0 wave of an "atamadaka" type

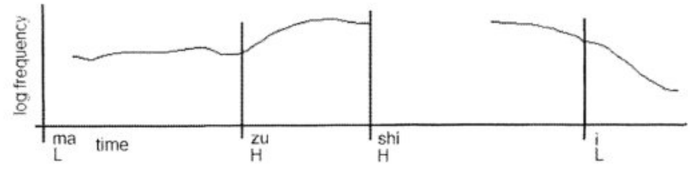

Fig 8 – F0 wave of a "nakadaka" type

3.3 Discussion

For the overall results, the 3 mora model performed only slightly better than the 2 mora model, but the 3 mora model had better accuracy when taking size and type into account so the 3 mora model presented is better suited for pitch pattern identification than the two mora model. Even so, the three mora model was trained on Japanese speech with resynthesized pitch so it may not be an accurate model in terms of the HLH and LLH pitch patterns. To test whether it is, it will be necessary to use natural speech that contain these types such as non-native speech,. Also, these experiments were performed on people trained in pronouncing the Japanese accent so this method needs to be tested on speech of people of the general public, as it is unnecessary for the learner to sound like announcers or narrators.

4 Conclusion

In this paper, a method for the automatic detection of pitch level patterns in Japanese word for the purpose of building a CALL system. This method was based on breaking Japanese words down into either two or three mora units, finding the best pitch level pattern for each unit and then from these the pitch level pattern combination with the highest likelihood. Though in Japanese, the combinations of LLH and HLH do not typically exist in native speech, we were able to train the three mora model by resynthesizing Japanese words to have those two combinations. These methods achieved overall results near 77% and 79% correct identification at the word level and around 91% and 93% at the mora level for the 2 mora model and 3 mora model respectively. The three mora model, though, performed better for the atamadaka type and nakadaka types so it could be more useful for a CALL system than the 2 mora model. These results may be sufficient for building language learning applications for word level accent practice.

In the future we plan to conduct experiments on non-native speech as well as speech by natives not trained in the Japanese accent. Also, we plan to further look into how to achieve better results. It may be necessary to use a different model making use of the underlying pitch wave features in the labeling process such as the rises and falls that occur within the word rather than just the perceptual features of high and low in order to determine the pitch level pattern. This is because a variety of pitch waves can represent the same pitch level pattern.

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

[10] T. Kanda, "Nihongo Akusento no Hyouki ni kansuru Kousatsu – Sanshiki Hyoukihou," Bulletin of Gifu Women's College, 2003, pp 51-58.

[11] K Tamaoka, T. Yasushi, "Mora or Syllable? Which unit do Japanese use in name visually presented stimuli?"Applied Psycholinguistics, 2004, pp 1-27.