A Pronunciation Training System for Japanese Lexical Accents with Corrective Feedback in Learner's Voice

Keikichi Hirose, Frédéric Gendrin** and Nobuaki Minematsu****

*Graduate School of Frontier Sciences, **Graduate School of Engineering, ***Graduate School of Information Science and Technology
University of Tokyo, Japan
{hirose, fred, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A system was developed for teaching non-Japanese learners pronunciation of Japanese lexical accents. The system first identifies word accent types in a learner's utterance using F_0 change between two adjacent morae as the feature parameter. As for the representative F_0 value for a mora, we defined one with a good match to the perceived pitch. The system notices the user if his/her pronunciation is good or not, and, then, generates audio and visual corrective feedbacks. Using TD-PSOLA technique, the learner's utterance is modified in its prosodic features by referring to teacher's features, and offered to the learner as the audio corrective feedback. The visual feedback is also offered to enhance the modifications that occurred. Accent type pronunciation training experiments were conducted for 8 non-Japanese speakers, and the results showed that the training process could be facilitated by the feedbacks especially when they were asked to pronounce sentences.

1. Introduction

The ideal way of teaching a language to non-native learners is to have one teacher to each single student. However, this ideal is hardly reachable because of teachers' availability and matters of money. The growing speed and memory size of personal computers with their decreasing price makes it possible to create automatic language trainers. Owing to recent advancements in written and spoken language processing technologies, these Computer-Aided Language Learning (CALL) systems have already been developed for writing and speaking skills, and become an alternative to such a lack of private lessons in nowadays classes.

As for the speaking skill, most systems are those for evaluating learner's pronunciation quality using HMM-based speech recognizers. Since learners have no idea on how to correct wrong pronunciations, an appropriate corrective feedback is essential for the efficient training. Most systems, however, repeat teacher's utterances with a visual display showing how the acoustic features of learner's utterances differ from those of teacher's utterances. Also, systems for training intonation and accent are rather few, though learning process concerning prosodic features requires a long practice of the language.

The Japanese language has the property of possessing a large number of homonyms, which can only be distinguished from each other by their respective *Kanji*'s (Chinese characters) in written communication, and their pitch patterns in oral communication. For instance, "hashi" can be "chopsticks," "bridge," or "edge," depending on the pitch pattern. Japanese pitch accent pronunciation, thus, becomes a great deal for non-natives who have no possibilities of distinguishing one homonym from the others.

From these points of view, we have developed a CALL system for Japanese lexical accent pronunciation, and realized a corrective audio feedback in learner's own voice [1]. The prosodic features of teacher's speech were copied to the learner's speech through speech modification. The learner can hear his/her speech before and after modification, and thus can obtain a better idea on his/her pronunciation problems. The rest of the paper is constructed as follows: In section 2, a mora F_0 with a good match to the perceived pitch is defined and the result of accent type recognition using the value is shown. After explaining the outline of the developed CALL in section3, the method of speech modification is explained in section 4. Visual feedback is explained in section 5. In section 6, results of accent type pronunciation training are shown. Section 7 concludes the paper.

2. Accent type recognition

One of the key technologies in developing the system is the accent type recognition. In continuous speech of Japanese, a content word (or a compound word comprising a sequence of content words) followed by a particle (in some cases, null or more than a particle) comprises a *bunsetsu*, which is a grammatical unit also corresponding to an utterance unit. In most cases, a *bunsetsu* is uttered with one accent type and corresponds to an accentual phrase. It is said that Japanese perceive lexical accents as the relative high-low pitch pattern of consecutive *morae*, and, therefore, a binary description in Fig. 1 is often used to schematically show the pitch movements. Here, *mora* is again a basic unit of Japanese utterance mostly coinciding to a syllable. Therefore, the most natural way to represent fundamental frequency (F_0) movements for accent type recognition may be to view F_0 average value for each mora. In our previous work [2], input speech was first segmented into morae by a phoneme-HMM-based speech recognizer and the average F_0 value of each mora was used for accent type recognition rather successfully. The method was applied to a CALL system teaching Japanese accent type pronunciation. The method worked well for isolated words, but showed a serious degradation for continuous speech. In continuous speech, due to many factors affecting F_0 movements, such as phrasing, accent sandhi, etc., an accent type shows variations in its features and its correct recognition comes rather difficult.

To cope with the difficulty, we conducted perceptual experiments to relate perceived mora pitch values with F_0 values. It came clear that the target value of F_0 movement

could be a better presentation of mora pitch value. Here, the target value was calculated as the F_0 value of the linear regression approximation of the F_0 movement at the end of mora. The average F_0 value was found to be a good presentation also, if its value was viewed in VC unit. The ratio of the mora F_0 vales thus defined at frames *i*-1 and *i* was used as the recognition parameter. For each mora length and each accent type, distribution of the ratio was assumed as a Gaussian, and its center and deviation were calculated from the training data. An accent type recognition experiment was conducted for ATR continuous speech corpus with 503 sentence utterances [3]. Ninety percent of the corpus was used for training and the rest was used for the testing. As the result, 75.5 % of correct recognition was realized. The detail was reported elsewhere already [4]. The use of mora F_0 values with a good match to the perceived pitch values will also be beneficial in that the learner can obtain a better view on his/her pitch control from mora F_0 sketch on the display.

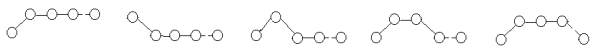

Figure 1: *Binary description of 4-mora Japanese pitch accent patterns. The fifth circle point in each pattern represents pitch level of the attached particle. Type 0 can be distinguished from type 4 by the particle's pitch level.*

3. System outline

The system first requests the learner to pronounce homonym pairs (with different meaning according to the accent types) either in isolation or in sentences. The readings of words and sentences are shown in Roman characters on display together with Japanese orthographic representation. The learner's utterances are recorded and segmented into mora by the forced alignment using mono-phone HMMs. Then the accent types of the homonyms (*bunsetsu*'s for sentence utterances) are identified using the method explained in the previous section. The system shows on display whether the accent type pronunciation is correct or not, together with some information useful for the training. An example is shown in Fig. 2. Also, the leaner can hear the teacher's (correct) sound pre-stored in the system by clicking a button on the display. This system shall be called the baseline system hereinafter.

Correct accent fall position:			
Ki ru (to wear)		Ki ＼ ru (to cut)	
Detected accent fall position:			
Ki ＼ ru	bad!	49.96	45.81
Ki ru	bad!	46.25	46.21

Figure 2: *An example of accent type pronunciation evaluation result. Symbol " ＼ " corresponds to the fall of the pitch accent from high to low at the accent nucleus. "bad!" indicates that the accent type is wrongly pronounced. If the accent type is correct, "good!" will appear instead. The four digit figures (such as "49.96") corresponds mora F_0 values in MIDI musical scale.*

The major point of this paper is that we added new audio and visual feedbacks to facilitate learning process. The audio feedback is the learner's utterance, whose prosodic features are modified to teacher's ones. The visual feedback is the waveforms before and after modification with schematic illustration of pitch movements. These are explained in the following sections. From now on, the system with these new feedbacks shall be called the new system.

4. Speech modification

Learner's speech was modified in its F_0, phoneme duration, and waveform amplitude to have prosodic features similar to teacher's speech through Time-Domain Pitch Synchronous OverLap Add (TD-PSOLA) [5]. The speech quality after TD-PSOLA process largely depends on the accuracy of pitch marking. Since the speech modification process should be done in online, no manual correction is allowed for detected pitch marks, which is different from the case of developing waveform concatenative speech synthesis systems. To solve this problem, we have utilized an automatic pitch marking method developed by the authors [6]. This method first locates the pitch marks based on the pitch extraction results, and then adjusts the locations so as to maximize a (total) cost through dynamic programming. The (local) cost is the signal amplitude of one pitch period which is hanning windowed centered at the pitch mark.

4.1. F_0 modification

The F_0 modification process stands in the pitch mapping function. This mapping function was realized with the teacher's speech pitch marks as the basis. First, the correspondence of phonemes in teacher's speech and learner's speech was obtained based on the result of forced alignment. Then each phoneme's pitch marks in the student's signal were replaced by the corresponding phoneme's pitch marks in the teacher's signal, after having changed in the overall F_0 level in order to keep student's original voice tone after modification. This was realized simply by multiplying the teacher's set of pitch mark delays (periods) by a ratio of the F_0 means of the two signals:

$$P_{ratio} = \frac{PmD_{ut.}(S)}{PmD_{ut.}(T)} \quad (1)$$

Where *PmD* stands for the mean of the pitch mark delays on the observation segment. *ut.* means that the calculation is done on the whole utterance, and *S* and *T* stand for Student and Teacher, respectively.

4.2. Duration modification

The previous process did not take into account the fact that the number of pitch marks should differ between the two signals, and did not manage the phoneme duration. These were handled in the duration modification process. The teacher's signal again served as basis during modification.

4.2.1. Duration adjustment

Here again, to respect the student's speaking rate, we first calculated the speaking rate ratio (SR_{ratio}) between the two signals, and multiplied the teacher's phoneme durations by the ratio:

$$SR_{ratio} = \frac{L_{ut.}(S)}{L_{ut.}(T)} \quad (2)$$

where the utterance length $L_{ut.}$ was calculated by adding all the phoneme durations that were obtained through the forced alignment process. Therefore, short pauses of both signals were not taken into account.

The number of pitch marks in a phoneme was modified in the following way:
The student's phoneme length has to be the same as the teacher's one after having modified according to the SR_{ratio}, i.e.,

$$L_{ph.}(T) \times SR_{ratio} = L_{ph.}^{new}(S) \quad (3)$$

where *ph.* stands for phoneme. We can get another expression of $L_{ph.}^{new}(S)$ and $L_{\text{ph.}}(T)$ by writing that the lengths are equal to the sum of pitch mark delays, or, in other words, their mean multiplied by the pitch mark numbers:

$$L_{ph.}^{new}(S) = \left(N_{pm}(S) + X_{new\_pm}(S)\right) PmD_{ph.}^{new}(S) \quad (4)$$

$$L_{ph.}(T) = N_{pm}(T) \times PmD_{ph.}(T) \quad (5)$$

where N_{pm} represents the number of pitch marks present in the original signal, and $X_{new\text{-}pm}(S)$ is the number of pitch marks we will have to add to (or delete from) the phoneme of student signal. Here, $PmD_{ph.}^{new}(S)$ can also be written as,

$$PmD_{ph.}^{new}(S) = PmD_{ph.}(T) \times P_{ratio} \quad (6)$$

From (3) to (6), the following relation is obtained:

$$N_{pm}(S) + X_{new\_pm}(S) = N_{pm}(T) \frac{SR_{ratio}}{P_{ratio}} \quad (7)$$

We have another constraint that the number of pitch marks in the teacher's speech should be match with the new number of pitch marks in student speech:

$$N_{pm}(T) + X_{new\_pm}(T) = N_{pm}(S) + X_{new\_pm}(S) \quad (8)$$

Combining (7) and (8), we finally obtain,

$$X_{new\_pm}(T) = N_{pm}(T)\left(1 - \frac{SR_{ratio}}{P_{ratio}}\right) \quad (9)$$

4.2.2. Pitch mark addition/deletion

The addition/deletion of pitch marks was realized in the central parts of vowels symmetrically around the central pitch mark position. For instance, if four pitch marks were to be added, two would be before the central pitch mark, and two after. The local pitch delays for these new pitch marks were chosen so that they would result in a smooth transition between original pitch marks; we opted for a mean of the pitch delays of the two surrounding pitch marks.

4.3. Power modification

Although power information is not essential for Japanese lexical accent realization, we decided to modify the power, as the learners could be tempted to use power as they might do in their mother tong. For instance, English native speakers may try to realize Japanese accent by placing a stress on a syllable.

Here again, the power reference was extracted from the teacher's speech. We modified the student signal intensity using TD-PSOLA by applying the teacher's signal intensity envelope (after the spline curve interpolation) on each phoneme of the student signal.

4.4. Short pause deletion

Non-native speakers tend to insert short pauses where native speakers do not. This insertion is a reason why non-native speaker's speech sounds unnatural. So we added a short pause (SP) deletion process to our speech modification scheme. In order not to delete useful SP's, we compared the resulting segmentation of the forced alignment process, and deleted the student signal SP's which were not present in the teacher's signal.

4.5. Evaluation

A listening experiment was conducted for native Japanese speakers to confirm that the modification was done properly. Ten sets of Japanese homonyms uttered by a non-native speaker with a good skill of realizing Japanese accent were first recorded. Then, each utterance was modified to have the opposite meaning of the homonym pair, which was offered to 8 male speakers, selected out of Japanese native speaker population of our laboratory, for evaluation. The modified signals were offered randomly to avoid any contextual effects. The evaluation was done in 5 scores: point 5 if the signal sounds completely with the meaning intended by the modification, and point 1 if its meaning is unchanged through modification. We got the average score 4.75 with variance 0.07, indicating that the modification was done mostly as we planned.

The modified signal presented some discontinuities due to some mis-placed pitch marks at phoneme extremities. These segments, having a low power, sometimes were recognized as unvoiced, and the pitch marking results came inaccurate. Another possible reason for speech quality degradation was that the spectral features were not taken into account during the modification. These are left for the future work.

5. Visual feedback

We combined our audio feedback with visual information to enhance the changes that occurred during the speech modification process. An example can be observed in Fig. 3, where a student was asked to utter two homonyms "kiru (to wear)" and "kiru (to cut)," but inversed pitch accent pattern of the two words. The visual feedback consists in displaying the original and the modified speech waveforms, and indicating the phoneme segmentation results. It also includes pitch movements as black arrows above the waveforms, showing the accent nucleus position within the word. The student can remark the differences in F_0, duration and power between the two signals.

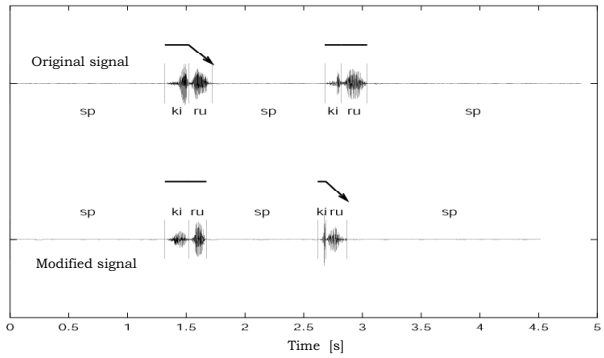

Figure 3: *An example of visual feedback for the couple of homonyms "kiru (to wear)" and "kiru (to cut)."*

6. Experiment

In order to evaluate the system, an experiment was conducted on the training of accent type pronunciation. The aim of the experiment is to observe if there is any difference in the learning process when using the baseline system and when using the new system. Eight non-Japanese male speakers, all foreign students at the University of Tokyo, were asked to use the both systems alternatively. They are 1 Bangladesh, 2 French, 1 Algerian, 1 Polish, 2 Chinese, and 1Urugaryan.

The experimental utterance consisted of the 10 homonym pairs and 4 sentences. Each sentence includes a homonym pair, and the system only evaluates the accent type pronunciation of the part. The experiment was done first using the new system for a pair and then using the baseline system for another pair so that habituation of using systems might not lead to a biased result, at least it did no work positively to the new system.

The learners were asked to utter the words/sentence without any accent type information at the first try. Then the feedbacks would be offered, and the further tries would be attended if necessary. An exercise would be considered as over when the student gets a "good" score, and a maximum of 5 trials were permitted. The number of tries before getting the "good" score was recorded and used as an index for the training efficiency. Even if the learner could not reach the "good" score after 5 trials, it was assumed that he got the "good" score. (The number of trials was counted as 5.) While experimenting the systems, the learners were also asked to fill in a form keeping their impressions using the new system. The form includes questions on:

1. Efficiency of the audio feedback in learner's voice. From 1 (inefficient) to 5 (very useful).
2. Efficiency of the visual feedback as shown in Fig. 2. From 1 (inefficient) to 5 (very useful).
3. Closeness of the modified speech to the learner's voice. From 1 (totally different) to 5 (the same voice).

It appeared that the first set of experiments based on isolated pairs of homonyms did not lead to striking differences between the two systems: the averaged number of trials being 2.4 for the baseline system and 2.5 for the new system. However, we could observe a difference in the continuous speech exercises, where the typical number of trials fell from 4.8 when used the baseline system to 3.8 when used the new system. The scores of the three questions were 4.4, 4.0 and 4.2, all indicating the positive effects of the feedbacks newly added.

Here are some comments the learners wrote after the experiment: In continuous speech, the modified signal would be easier to understand if there were a function to isolate words or parts in question from the sentence, and enhance the modifications. The function to slow down the modified speech would also do, they said. It also seemed that the new visual feedback was difficult to understand for beginners.

7. Conclusion

A CALL system to train Japanese pitch accent pronunciation was constructed. It corrects the prosodic features of the learner's utterance by the TD-PSOLA scheme, and outputs the corrected speech as an audio corrective feedback. Through the experiment of using the systems with and without such a feedback, the corrective feedback in learner's own voice was shown to be effective especially when the training task came complicated.

Further studies are planned on the following points:

1. To improve the speech quality by manipulating spectral features also.
2. To make it possible to choose the teacher's signal among a set of speakers to get the closest voice to the student's signal.
3. To improve the visual feedback to enhance important information.

8. Acknowledgements

The authors' appreciations are due to Dr. Carlos Toshinori Ishi for his useful discussions on writing this paper.

This work was partly supported by Grant in Aid for Scientific Research of Priority Areas (#120).

9. References

[1] Gendrin, F., Hirose, K., and Minematsu, N. " Corrective feedback for accent pattern CALL systems using speech modification," *Technical Report. IEICE Speech Committee*, SP2002-161, pp.1-6 (2003).

[2] Kawai G. and Ishi C. T., "A system for learning the pronunciation of Japanese pitch accent," *Proc. 6th European Conference on Speech Communication and Technology*, Vol.1, pp.177-180 (1999).

[3] Speech Corpus Set B. http://www.red.atr.co.jp/database_page/digdb.html

[4] Ishi, C., Hirose, K., and Minematsu, N., "Mora F_0 representation for accent type identification in continuous speech and considerations on its relation with perceived pitch values," *Speech Communication*, to be published (2003).

[5] Moulines, E. and Charpentier, F. "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, Vol.9, pp.453-467 (1990).

[6] Benjamin, N., Hirose, K., and Minematsu, N. "An experimental study on concatenative speech synthesis using a fusion technique and VCV/VV units," *Technical Report. IEICE Speech Committee*, SP2001-121, pp.53-60 (2002).