Accent Type Recognition of Japanese Using Perceived *Mora* Pitch Values and Its Use for Pronunciation Training System

Keikichi Hirose

Graduate School of Frontier Sciences, University of Tokyo, Japan
hirose@gavo.t.u-tokyo.ac.jp

Abstract

Through perceptual experiments, the fundamental frequency (F_0) in *mora* unit was defined which well corresponded to the perceived pitch value. Several candidates, given as the combination of periods of observation and methods of F_0 calculation, were tested. Based on the definition, a method was developed for the accurate recognition of Japanese lexical accents, and was applied to a system for teaching non-Japanese learners pronunciation of Japanese accents. In the method, each accent type was represented as a multi-dimensional *Gaussian* model, where F_0 change between two adjacent *morae* was used as the feature parameter. The system first recognizes accent types of words in a learner's utterance, and then notices the learner if his/her pronunciation is good or not with audio and visual corrective feedbacks. Using TD-PSOLA technique, the learner's utterance is corrected in its prosodic features by referring to teacher's features, and offered to the learner. Since the learner can hear how his/her utterance should sound after correction, he/she can obtain a better idea on the correction as compared to the case where only teacher's utterance is offered as the audio 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

Recent development of internationalization largely increased the situation where a person should speak in his/her non-mother tongues. For many foreigners, it is not an easy task to reach an affordable level of making conversations with native speakers. When a person is looking for a job in foreign countries, the situation is serious; his/her mental ability is sometimes ranked in a lower level because of his/her accented pronunciation. The ideal way for learning a foreign language is to have a "good" native teacher, which is usually not the case. If a device enabling the self-training of pronunciation skill for learners of foreign languages is available, especially for their early stage of training, such a problem on the lack of private lessons in nowadays classes will be eased.

Technologies on spoken language processing, such as speech recognition and synthesis, have shown remarkable progresses in these several years. These technologies reached a level enough to be utilized in language learning systems. Speech recognition can be utilized to assess learner's pronunciation, while speech synthesis may be utilized to generate speech with correct pronunciation in learner's voice. Therefore, a rather large number of computer-assisted language learning (CALL) systems have already been developed.

Many systems for pronunciation training try to score the learner's speech depending on its distance from reference (teacher's) speech in an acoustic feature parameter space. However, this method includes a serious problem in that the score does not necessarily correspond to the learner's pronunciation skill. A scheme is necessary to avoid a learner to be forced to mimic teacher's utterance. Recently, HMM-based speech recognizers are used in several systems to tell if the learner's utterance is good or not. However, use of the state-of-the-art technology cannot solve the problem; speech misrecognized by the recognizer is not necessarily a wrong pronunciation. Furthermore, corrective feedback from the system is rather poor in most systems. Usually, teacher's (correct) utterances are repeatedly offered to a learner with a visual display showing how the acoustic features of learner's utterances differ from those of teacher's. No clear instruction on how learner's can correct their pronunciation is offered. If he/she can correct his/her pronunciation only with this corrective feedback, the pronunciation training will not be necessary from the first place.

From these considerations, the following three will arise as necessary functions, which a pronunciation training system should have:

1) Clarify where and how a learner makes a mispronunciation,
2) Indicate whether the mispronunciation is acceptable to natives or not,
3) Show how the learner can correct his/her pronunciation.

Also, systems for training intonation and accent are rather few, though learning process concerning prosodic features requires a long practice of the language.

From this point of view, we have already developed a system for training pronunciation of Japanese lexical accents, where learner's utterances are evaluated if they are perceived correctly to native Japanese speakers [1]. In the current paper, we present another system for Japanese lexical accents, where learner's pronunciation with wrong accent types is corrected by the time-domain pitch synchronous overlap add (TD-PSOLA) method [2] and is offered to the learner as the corrective audio feedback [3]. The learner can hear his/her

speech before and after correction, and thus can obtain a better idea on his/her pronunciation problems. The system also provides a visual feedback to facilitate the training process.

In the system, fundamental frequency (F_0) of each *mora* in learner's utterances is represented by a value corresponding to the perceived pitch and is used for the accent type recognition. Here, a *mora* is a basic pronunciation unit of Japanese, and is an important reference to keep a tempo of an utterance. It corresponds to a syllable, except for *tokushuhaku* syllables. The *mora* F_0 value has two merits: one to realize an accurate recognition of accent types, and the other to visualize the F_0 movement with a good match to the perceived pitch values. The latter will be beneficial in that the learner can obtain a better view on his/her pitch control from *mora* F_0 sketch on the display.

2. Japanese Lexical Accents

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.

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 serves as an utterance unit. In most cases, a *bunsetsu* is uttered with one accent type and corresponds to a prosodic word. 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. In Tokyo dialect, an *M-mora* word can have one of *M+1* accent types, as shown in Fig. 1 for $M = 4$.

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.*

3. *Mora* F_0 value

The binary description in *mora* units shown in Fig. 1 is widely used to explain Japanese accent types, since it agrees with Japanese native speaker's feeling of producing and perceiving accent types. Thus, even though F_0 may change within a *mora* and/or between *morae*, it would be possible to associate one representative F_0 value for each *mora* (henceforth, *F0mora*) of the utterance. The question is how we calculate it form the observed F_0's. First, all the F_0's are represented in a musical scale as defined as follows so that their movements are analyzed on a logarithmic scale.

$$F_0\ [semitone] = 12 * \log_2 (F_0\ [Hz]) \quad (1)$$

The most naive way is to take the average of F_0's in the *mora* unit. However, there arise two questions; what is the unit of *mora* and is the averaging the best way? As for the unit, it is natural to select *CV* (consonant-vowel) sequence. However, when pronouncing Japanese, it is widely said that a *mora* isochronism is perceived, and there are arguments through perceptual experiments showing that isochronism is more evident in ***VC*** units than in *CV* units [4]. Taking this into account, both *CV* and *VC* units are selected and compared as the unit of calculation.

Averaging cannot be the best way if we take the non-linearity of human perception process into account. For example, when listening to a word with accent type 1, the first *mora* is perceived as having higher pitch than the second *mora*. However, the average F_0 of the first *mora* is often equal to or lower than that of the second *mora*. This fact implies that the direction of F_0 movement within a *mora* may affect the perception of the *mora* pitch height (higher pitch for increasing F_0 and lower pitch for decreasing F_0). Based on this consideration, the target value of the F_0 movement within the *mora* is added as a candidate of *F0mora* besides the average value. Currently the target value is calculated as the value at the unit end of the linear regression line to the F_0 curve.

Portions with lager power affect more on human perception than other portions. Following to this assumption, an option of weighting F_0 values with the waveform power is also added to the *F0mora* calculation. When calculating the target value, the weighting is included in the process of obtaining the linear regression line; the error between observed F_0 and F_0 of the regression line is weighted. As the combination of these options explained above, we investigated 8 alternatives: *CV* vs. *VC*, average (*avg*) vs. target (*tgt*), and non-weighted (*nw*) vs. weighted (*w*).

The (perceived) pitch value of each *mora* (henceforth, *F0human*) was obtained through perceptual experiments using a device, whose MIDI sound pitch is adjustable by the subjects [5]. Through preliminary experiment, clarinet sounds were selected for the purpose among several instruments sound candidates. Six subjects with musical experience were asked to adjust the pitch of the MIDI sound so that it is perceived as having the same pitch as a *mora* (syllable) extracted from a natural utterance. Adjusted pitch values are averaged over the subjects to obtain *F0human*. One hundred syllables were manually segmented from 9 sentences and 2 isolated words uttered by one male and one female speakers. Table 1 shows root mean square errors (in semitone) between *F0human* and *F0mora* candidates averaged over the 100 syllables. The result indicates that *avg-VC* and *tgt-CV* give better estimates for *F0human*, and estimates by *w* are slightly better than those by *nw*.

Table 1: *Root mean square errors, in semitone, between F0human and F0mora (CV/VC; avg/tgt; nw/w).*

	nw		*w*	
	CV	*VC*	*CV*	*VC*
avg	1.81	1.61	1.61	1.45
tgt	1.68	3.11	1.58	2.74

4. Accent type recognition

The ratio of *F0mora*'s thus defined at frames *i*-1 and *i* was used as the recognition parameter [5]. For each *mora* length and each accent type, distribution of the ratio (henceforth, *F0ratio*) 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 [6]. Ninety percent of the corpus was used for training and the rest was used for the testing. As a baseline, we also built F_0-based HMM's, where F_0 contour was represented as a sequence of frames of F_0 and/or ΔF_0. Left-to-right duration-controlled HMM's with states equal to the number of *morae* of the prosodic word in question were used. As feature parameters, we used 2-dimensional vectors of F_0 and ΔF_0 for one model, called HMM_{ref1}, and ΔF_0 only for another model, called HMM_{ref2}. The latter was intended to correspond to the proposed *F0ratio*-based method, where only the differences between adjacent *F0mora* are counted. The results are shown in Table 2. The proposed method outperformed the baseline for all of the *F0mora* definitions. Among the four candidates of *F0mora*, *avg-VC* and *tgt-CV* produced better recognition, showing a good match with the results in Table 1. Although, in the current experiments, *w* version of *F0mora* was not tested, it may yield a better result.

Table 2: *Rates of accent type recognition for F0ratio-based and F0-based models.*

		Accent type recognition (%)						
	Accenttype	0	1	2	3	4	5	all
	Samples	1039	800	486	309	169	64	2867
F0ratio	*avg-CV*	82.4	76.3	43.2	54.4	63.9	62.5	69.5
	avg-VC	84.8	83.1	56.4	62.8	61.5	54.7	75.1
	tgt-CV	82.1	85.8	61.9	60.5	59.8	56.3	75.5
	tgt-VC	82.1	71.9	49.0	55.7	49.7	43.8	68.0
$F_0/\Delta F_0$	HMM_{ref1}	75.2	76.2	46.8	45.2	39.3	33.3	65.3
	HMM_{ref2}	71.0	56.4	50.8	38.5	31.1	21.4	57.4

5. 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 HMM's. Then the accent types of the homonyms (*bunsetsu*'s for sentence utterances) are recognized 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.

In the current paper, 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.

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. The four digit figures (such as "49.96") corresponds mora F_0 values in MIDI musical scale.*

6. 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) [2]. 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 [7]. 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.

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.

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 the basis during modification. The detail is already explained elsewhere [5].

Adding to the F_0 and duration modifications, power was also modified and short pauses, not present in the teacher's signal, were deleted.

7. Visual feedback

We combined our audio feedback with visual information to enhance the changes that occurred during the speech modification process. 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.

8. 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.

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.)

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.

9. Conclusion

After defining *mora* F_0 value, which has a good correspondence with perceived pitch, through a perceptual experiment, it was used to develop an accurate accet type recognition method for continuous speech. Then, a CALL system to train Japanese lexical 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.

The author's appreciations are due to Prof. Nobuaki Minematsu, Dr. Carlos Toshinori Ishi and Mr. Frédéric Gendrin for their contribution to the work.

10. References

[1] Kawai, G.; Ishi, C. T., 1999. A system for learning the pronunciation of Japanese pitch accent. *Proc. EUROSPEECH*, Vol.1, 177-180.

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

[3] Hirose, K.; Gendrin, F.; Minematsu, N., 2003. A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice. *Proc. EUROSPEECH*, Vol.4, 3149-3152.

[4] Sato, H., 1987. Rule-based speech synthesis. *PhD thesis*, 55-92.

[5] Ishi, C.; Hirose, K.; Minematsu, N., 2003. *Mora* F_0 representation for accent type identification in continuous speech and considerations on its relation with perceived pitch values. *Speech Communication*, Vol.41, Nos.2-3, 441-453.

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

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