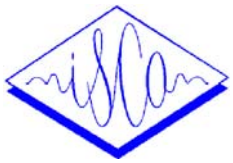

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A SYSTEM FOR LEARNING THE PRONUNCIATION OF JAPANESE PITCH ACCENT

Goh Kawai and Carlos Toshinori Ishi

goh@kawai.com **c_t_ishi@gavo.t.u–tokyo.ac.jp**

University of Tokyo, Department of Information and Communication Engineering,
Tokyo 113–8656 Japan
http://www.gavo.t.u–tokyo.ac.jp/

ABSTRACT

We propose a technique that associates the F0 of Japanese words with native-speaker perceptions of pitch accents. Up till now, computer-aided prosody learning systems that measure fundamental frequency (F0) have not mapped F0 values to perceptual thresholds of native speakers. Existing learning systems often instruct their students to fit F0 contours within a band designated by the system, but the band's boundaries do not to correspond to measures of acceptance by native speakers. Our method estimates intelligibility of pitch accent patterns based on perception experiments. Intelligibility scores motivate learners by guiding them towards meaningful goals.

10.21437/Eurospeech.1999-48

1. INTRODUCTION

We describe a method to teach pronunciation skills related to pitch. Pitch is perceived based partially on physical stimuli (such as intensity, duration and fundamental frequency) that may be intermittently present. Pitch manifests itself differently between languages, dialects, pragmatics, and individuals. Cross-linguistic and inter-speaker variability of the production and perception of pitch are not well understood, and the teaching of pitch for non-native speakers has remained a challenge.

Some existing computer-aided pronunciation learning systems teach prosody by measuring the students' F0 and instructing them to fit their F0 contours within a preassigned band [1]. However the band's edges do not correspond to perceptual thresholds of native speakers. It would be useful to predict the percentage of native speakers who would understand the learner's speech the way the learner intends to be understood.

This paper investigates how F0 measurements of pitch patterns can be mapped to perceptual thresholds of native speakers. The lexical pitch accent of Tokyo-dialect Japanese is taken as an example. Japanese pitch accent is a type of word accent where each word's mora (a sub-syllabic rhythmic unit that is phonemic in Japanese) adheres to specific high (H) - low (L) pitch sequences. Examples of minimal pairs include "umi (H-L)" (sea) vs "umi (L-H)" (pus), and "ikoo (H-L-L)" (after) vs "ikoo (L-H-L)" (go) vs "ikoo (L-H-H)" (shift). Errors in pitch accent clearly signal non-nativeness, and listeners may doubt the speaker's communicative skills. Among the various factors that contribute to pitch accent perception, F0 is the most important for non-native learners.

Our method measures pitch by estimating the F0 of segments in the speech signal. The F0 of Japanese words are matched with native-speaker perceptions of pitch accents that are obtained by perception experiments using synthetic speech stimuli. Learners receive an intelligibility score along with instructions on how to speak better. Intelligibility scores motivate learners and allow them to stop practicing when their pronunciation reaches a certain point even if their pronunciation is not completely native.

2. PITCH ACCENT PERCEPTION

2.1. Perception Experiment Conditions

We ran perception experiments using resynthesized words to quantify the effects of F0 on pitch accent perception. The objectives of the experiment were to:

(a) Verify the feasibility of creating pitch-accent minimal pairs by using artificially modified pitch patterns to resynthesize actual speech signals.

(b) Qualify and quantify the relationship between F0 and moras across pitch accent boundaries (i.e., where H-L movements occur).

Two Japanese speakers (1 male, 1 female) recorded 10 pitch-accent minimal pairs. We used resynthesis techniques to adjust the F0 of the mora immediately following the pitch accent boundary. The F0 value of the mora immediately left to the pitch boundary ($F0_i$) was held constant at approximately 100 Hz or 200 Hz depending on whether a male or female voice was being resynthesized. The F0 value of the mora immediately right to the mora boundary ($F0_{i+1}$) was adjusted in 9 steps, starting from a low value through $F0_i$ to a high value, so that $F0_i >> F0_{i+1}$, $F0_i > F0_{i+1}$, $F0_i = F0_{i+1}$, $F0_i < F0_{i+1}$, $F0_i << F0_{i+1}$.

Both recordings of a single speaker's minimal pair were resynthesized, so that there were 4 resynthesized sets for each pair (2 speakers x 2 words). We defined the pitch difference between two adjacent moras as

$$accent = \log_e F0_{i+1} - \log_e F0_i$$

Each step in the 9-step F0 adjustment mentioned above differed at *accent* value intervals of 0.1. Within each mora, F0 contours were essentially flat, with smoothed

transitions at mora boundaries. As an example, a set of resynthesized speech waveforms of the word “aka” (H-L “red” vs L-H “filth”) is included in the conference proceedings CD-ROM under the filenames “k054001.wav” through “k054009.wav”.

Although an F0 model that decomposes F0 patterns into phrasal and accent components generates better-quality speech, we did not use the model for two reasons: first, automatically calculating the parameters for phrasal and accent components from the acoustic signal is unreliable, and second, even if accurate information were available, learners probably would be unable to comprehend or apply it.

Three native speakers of Japanese listened to a randomized list of the resynthesized words in a quiet room. For each resynthesized word that was played, subjects were asked to choose between the minimal pairs and fill in a response sheet.

2.2. Perception Experiment Results

The subjects’ responses showed there were no significant differences between the results on the stimuli created from recordings of either minimal pair. This means that experiment stimuli can be resynthesized from either H-L or L-H patterns, regardless of changes in pitch accent possibly affecting segmental features.

No significant difference was found between resynthesized stimuli created from male and female utterances. This means that experiment stimuli can be resynthesized from either male or female speech. Together with the results described in the previous paragraph, we found that perception experiments of pitch-accent minimal pair are possible using resynthesized stimuli.

Figure 1 shows the subjects’ responses depending on *accent* values and pitch accent patterns. We found that agreement among native speakers was practically unanimous for most *accent* values, showing that an *accent* value of -0.1 is needed to perceive a H-L fall in 2 or 3-mora words. (An *accent* value of -0.1 corresponds to about a 10-Hz drop when the base F0 is 100 Hz, as was the case for words resynthesized from male speech.) For 4-mora words, an *accent* value of -0.2 was necessary. The fact that the *accent* value increases as a function of the number of moras in the word suggests that even

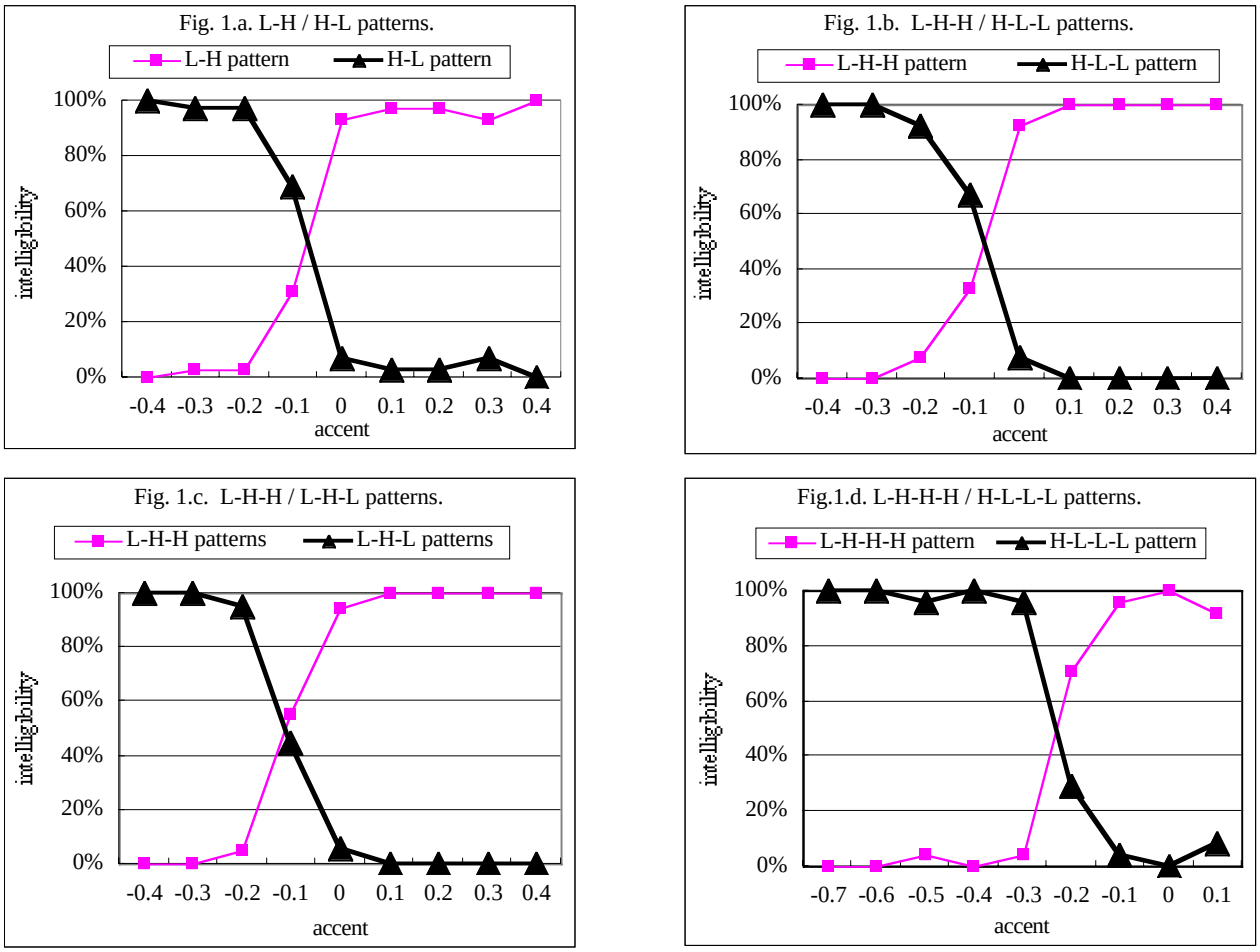

Figure 1. Subjects’ responses in the pitch accent perception experiment, according to *accent* values and word length. Word length is measured by the number of moras in the word. *Accent* values are plotted on the horizontal axis; pitch heights between adjacent moras is H-L towards the left and L-H towards the right of the charts. The subjects’ responses are plotted on the vertical axis. Subjects’ reponses correspond to intelligibility scores for that *accent* value.

though *accent* is the pitch difference between two adjacent mora, the difference must be exaggerated for longer words. The *accent* value may have been larger for longer words because we did not use declination in our F0 modeling.

3. PITCH LEARNING SYSTEM

3.1. System Structure

We implemented a training tool to practice the pronunciation of pitch accent patterns in realtime. The main steps of pronunciation practice are as follows. First, the learner selects reading material from a system-provided list. The learner may choose at any time to listen to a native speaker's recording of the reading material. The learner speaks into either a desktop electret condenser microphone or a close-talking electret microphone. Audio input is encoded in linear format with 16-bit sampling rate at 16-kHz sampling frequency. The learner's speech is sent to two independent processes: forced-alignment using a speech recognizer, and F0 estimation using an autocorrelation algorithm. Both processes use 10-millisecond-wide frames. The process flow of the system is shown in figure 2.

Forced alignment is performed at the phone level using HTK v2.1.1 [2] with Japanese HMMs [3]. The HMMs are gender-dependent, tied-mixture triphones using 12th-order melcepstra, their deltas and delta-deltas, and delta and delta-delta power. The system asks the learner's gender at the beginning of a pronunciation practice session. Based on knowledge of the reading material, phones comprising the same mora are joined together.

The speaker's gender information is used also by the F0 estimator to avoid half-period and double-period errors. Thus F0 estimates have maximum and minimum values depending on the speaker's gender. The F0 estimator gives a confidence measure indicating voicing probability for each 10-ms frame.

When the forced-alignment and F0 estimation processes complete, the mora labels and F0 estimates are aligned with respect to time. Average F0 values of each mora are calculated by summing all non-zero F0 values of frames within the mora, and dividing the sum by the number of the non-zero F0 frames. The F0 values of adjacent moras are used to compute *accent* values. *Accent* values are matched against results from the perception experiment to obtain the percentage of native speakers who will understand the learner's pitch accent.

When a mora consists mainly or totally of voiceless sounds (such as voiceless fricatives, moraic plosive closures, and devoiced or deleted vowels), the mora's average F0 value may be impossible to calculate. In order to reliably estimate F0 values from the speech of non-native learners, we use reading material that is predominantly voiced. Interpolating the average F0 value of a voiceless mora from its neighbors incorrectly assumes the learner is predictably adjusting her pitch across the moras.

After the learner reads the words or phrase, she receives instructions on how to correct her pronunciation. Feedback to the learner consists of (a) the reading material with mispronounced portions highlighted, (b) instructions on where to raise or lower pitch, and optionally, (c) phonetic transcriptions of the reading material and the learner's rendition. For instance, the feedback might be "Your 'ame' (rain) sounds good, but your 'ame' (candy) sounds like 'ame' (rain). Lower the 'a' and raise the 'me'". Depending on the pitch pattern, the system instructs the learner to raise or lower particular portions of the reading material. This kind of feedback is straightforward regardless of the learner's educational background.

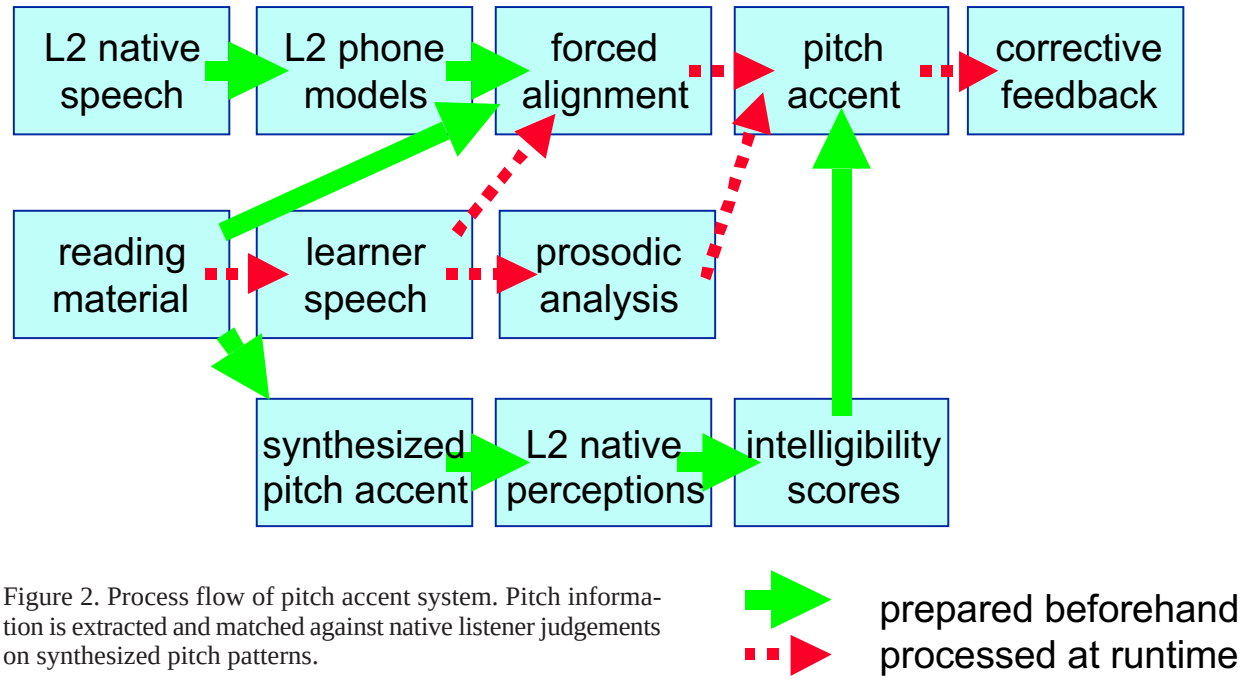

Figure 2. Process flow of pitch accent system. Pitch information is extracted and matched against native listener judgements on synthesized pitch patterns.

Table 1. Recognition accuracy according to speaker type and segmental limitations. Percentages of pitch accent recognition matches between human labelers and the system.

speaker type	number of speakers	all words % (n=157)	words excluding long vowels % (n=112)	2-mora words % (n=56)
natives	19	64.4	70.1	86.8
semi-natives	3	67.7	69.7	91.0
both groups	22	64.9	70.1	87.4

3.2. System Evaluation

Evaluation experiments were done under simulated conditions. 19 native speakers of Japanese (14 males, 5 females) and 3 semi-natives (2 males, 1 female) with at least 10 years experience using Japanese each recorded 157 words containing 79 minimal pairs. (One minimal pair, "ikoo", was a minimal trio; H-L-L vs L-H-H vs L-H-L.) The pitch heights of moras in the recordings were labeled separately by hand and machine. According to phonological rules for lexical pitch accent, the hand-labeling and machine-labeling procedures for determining pitch accent type forced the pitch heights of the first and second mora to be opposite of each other; the only possible combinations between the first and second pitch heights were H-L and L-H. In addition, only one pitch-fall was allowed within the word; once the pitch height dropped from H to L, all subsequent pitch heights were L. Applying these two phonological rules means that the chance probability of detecting pitch accent correctly was 1/n, where n is the number of moras in the word. Table 1 shows results of the comparison between hand-labeled and machine-labeled pitch accent patterns.

We found that pitch accent detection within long vowels (i.e., bimoraic vowels) was approximately 10 percent less accurate than within other segments. Long vowels are spectrally identical throughout except that pitch changes can occur between its two moras. The speech recognizer tended to assign long durations to the long vowel's first mora, meaning that the recognized first mora tended to include much of the second. Thus the average F0 of the first mora became closer to the second, and no pitch change could be detected. Forced alignment between spectrally distinct segmental boundaries is more accurate. Durational modeling may improve recognition performance for long vowels. Results for words without long vowels are included in table 1.

Minimizing the number of moras in the word helped recognition (see results for 2-mora words in table 1). The chance probability rose to 50 percent, because the system was choosing between H-L and L-H patterns. Recognition performance rose significantly for both native and semi-native speakers. These results, especially for 2-mora words, suggest that the system is a useful component technology for pitch accent training.

4. CONCLUSION

We described a method that (a) measures pitch patterns produced by non-native speakers, (b) compares non-native and native speech, obtain a intelligibility score, and (c) instructs the non-native on how to correct his pronunciation.

The proposed system uses speech recognition algorithms and prosody analysis algorithms to accurately measure pitch and align it with the location of each phone in the learner's speech. This technology might allow teaching of pitch accent and intonation contours to non-native learners. Perception experiments showed that while native speakers tolerate a wide range of pitch height within a syllable, subtle changes in pitch across syllables can differentiate high and low pitch accents. The level of agreement among native speakers as particular adjacent syllables being H-L or L-H indicates the appropriateness of the pitch accent pattern for that syllable pair.

Evaluation results suggest that technology has advanced to a point where self-study systems might help learning foreign language pitch production if the appropriateness of a given pitch pattern can be unequivocally agreed upon by native speakers. Our technique might be expanded to teach the intonation of phrases and sentences with fixed intonation patterns. Similar systems might be built for teaching pronunciation skills in any language involving pitch patterns that native speakers deem unanimously correct, such as the pronunciation of fixed expressions (i.e., phrases with fixed lexical composition and pronunciation patterns, such as "How do you do?"). Our method might also help detect prominence in speech recognition applications.

ACKNOWLEDGMENTS

We thank Keikichi Hirose (University of Tokyo) for reseach guidance, and Kazuya Takeda (Nagoya University) and his team for Japanese triphone HMMs [3].

REFERENCES

[1] Rooney, E. et al "Prosodic features for automated pronunciation improvement in the SPELL system" Proc. ICSLP-92 (Banff, Canada), pp. 413-416

[2] Young, S. et al "The HTK Book for version 2.1." Cambridge University, 1997

[3] Takeda, K. et al "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model." Information Processing Society of Japan SIG Notes, paper number 97-SLP-18-3, 1997