A CALL SYSTEM USING SPEECH RECOGNITION TO TRAIN THE PRONUNCIATION OF JAPANESE LONG VOWELS, THE MORA NASAL AND MORA OBSTRUENTS

Goh Kawai and Keikichi Hirose
Department of Information and Communication Engineering
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113 Japan
E-mail: goh@kawai.com hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

We developed a CALL (computer-aided language learning) system for teaching the pronunciation of Japanese long vowels, the mora nasal and mora obstruents to non-native speakers of Japanese. Long vowels and short vowels are spectrally almost identical but their phone durations differ significantly. Similar conditions exist between mora nasals and non-mora nasals, and between mora and non-mora obstruents. Our system uses speech recognition to measure the durations of each phone and compares them with distributions of native speakers while correcting for different speech rates. Results show that learners quickly capture the relevant duration cues. The amount of learning time spent on acquiring these durational skills is well within the time constraints of TJSL (teaching Japanese as a second language) curricula.

1. INTRODUCTION

TJSL has received significant attention over the last decade owing to the influx of non-native speakers entering Japan. Classroom instruction has concentrated on orthography, vocabulary and syntax. Although poor verbal skills brand non-natives (especially Asians) as inferior undesirables, training of pronunciation has received significantly less attention, partly due to lack of classroom time.

Self-study methods that exist today do not tell learners whether their speech is intelligible or what they can do to improve their pronunciation [1][2][3]. If learners could judge the appropriateness of their renditions by themselves, they would have no need to learn pronunciation skills in the first place. Clearly, some form of authoritative, corrective feedback is necessary to acquire skills efficiently.

One of the most acutely needed pronunciation skills is distinguishing between double-mora and single-mora vowels (also referred to as long and short vowels), the mora and non-mora nasal, and mora and non-mora obstruents [4]. The TJSL term "tokushuhaku" collectively refers to these three phonemic pair sets.

This study aims to train tokushuhaku without requiring an instructor's presence. The remainder of this paper analyzes tokushuhaku duration (section 2), describes a CALL system for training tokushuhaku (section 3), and reports evaluation results (section 4).

2. TOKUSHUHAKU DURATIONS

The predominant phonemic contrast between tokushuhaku is phone duration. While long and short vowels are often contrasted by both duration and pitch accent, and the mora nasal has several allophones, mispronouncing these differences does not compromise intelligibility [5]. Vowels, nasals and plosives can each be treated as a group, because the phones comprising each group behave similarly with respect to duration and speech rate.

The distribution of tokushuhaku durations of 20 speakers from the ATR database was analyzed [6]. The database consists of wideband, clean, native speech collected from 10 male and 10 female professional broadcast announcers. Speakers were instructed to read a set of context-dependent sentences at a speech rate comfortable and consistent for each speaker. The sentences were identical across speakers. For each speaker there were 115 sentences and 3933 mora. Hand-labeled segment durations were used for this study. In this paper, double letters denote tokushuhaku phones (for instance, [a] is single-mora and [aa] is double-mora). The phones [p][t][k] and [pp][tt][kk] refer to the preplosive closures of non-mora and mora plosives (the plosive bursts for non-mora and mora plosives are identical).

Phonological context plays a role in delimiting the durations of moraic and non-moraic stops. Although stops lengthen when speech rate falls, their elasticity is limited because over-lengthening leads to confusion with other phones. The mora nasal cannot lengthen beyond a certain threshold because it becomes ambiguous with a mora nasal followed by a non-mora nasal. (The mora nasal occurs syllable-finally; the non-mora nasal syllable-initially.) Likewise, over-lengthened mora plosives are confused with pauses; hence their maximum duration. Figures 1 and 2 show the distribution of nasals and plosives regardless of speech rate or phonological context.

For the purposes of corrective feedback to non-native learners, phone durations are divided into five zones: for nasals, for instance, (1) too short to be a non-moraic nasal, (2) acceptable non-moraic nasal, (3) ambiguous between non-moraic and moraic nasals, (4) acceptable moraic nasal, and (5) too long to be a moraic nasal. These ranges, which are derived from perception experiments, are superimposed on figures 1 and 2.

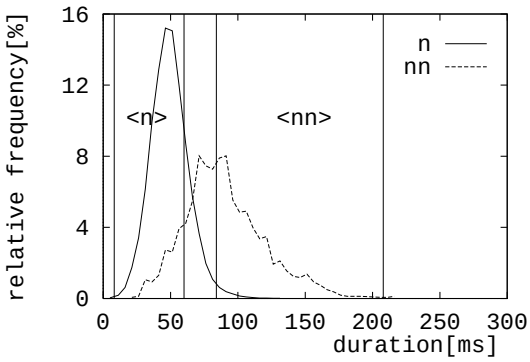


Fig. 1 Distributions and appropriate ranges of nasal durations

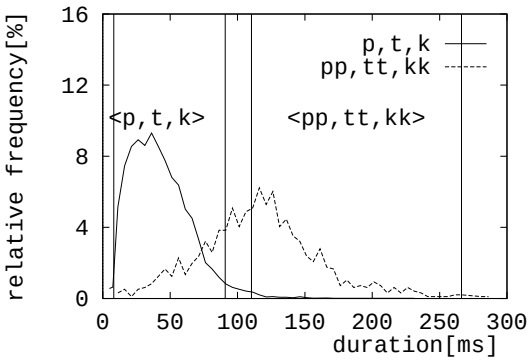
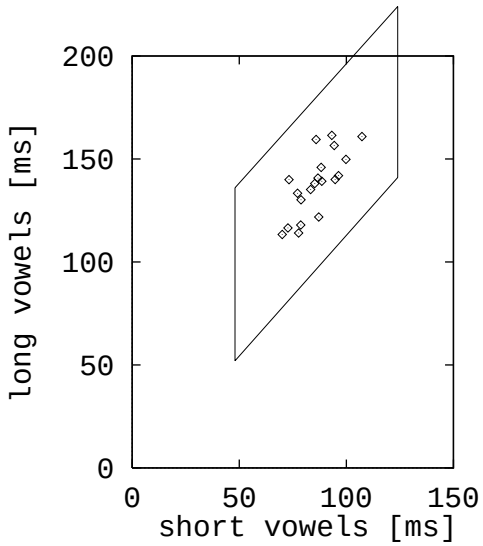


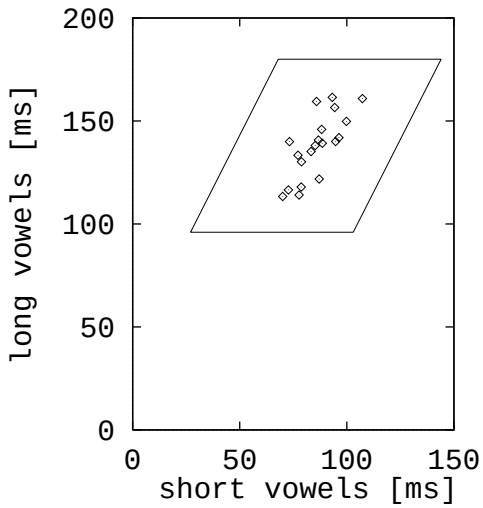
Fig. 2 Distributions and appropriate ranges of plosive durations

Compared with stops, both double-mora and single-mora vowels are spread over a significantly wider range, and are affected strongly by speech rate. Long and short vowels are indistinguishable by absolute ranges in duration; in fact, in conversational speech, short vowels said slowly can be longer than long vowels said quickly by the same speaker. Non-natives must discriminate long and short vowels based on local, short-term speech rates. Speech rate is thus a variable which should be controlled for the structured instruction of phonetic features.

When providing corrective feedback to a non-native learner's production of long and short vowels, it is important to know how the learner's native language affects his or her Japanese pronunciation, because this can determine whether the learner can produce short or long vowels correctly. For instance, native speakers of Chinese pronounce Japanese short vowels with correct durations because Chinese does not shorten vowels, but native speakers of Chinese cannot distinguish Japanese long and short vowels. An opposite example is the native speakers of English, who, on the one hand, produce Japanese long vowels with correct duration because English has diphthongs (however, as a separate problem, they diphthongize Japanese long vowels), but on the other hand reduce or delete short vowels because English does so with unstressed syllables [7]. Figure 3 shows appropriate tokushuhaku durations depending on whether the learner's short or long vowel durations are assumed to be correct.



(1) appropriate range when short vowel durations are assumed correct



(2) appropriate range when long vowel durations are assumed correct

Fig. 3 Distributions and appropriate ranges of vowel distributions

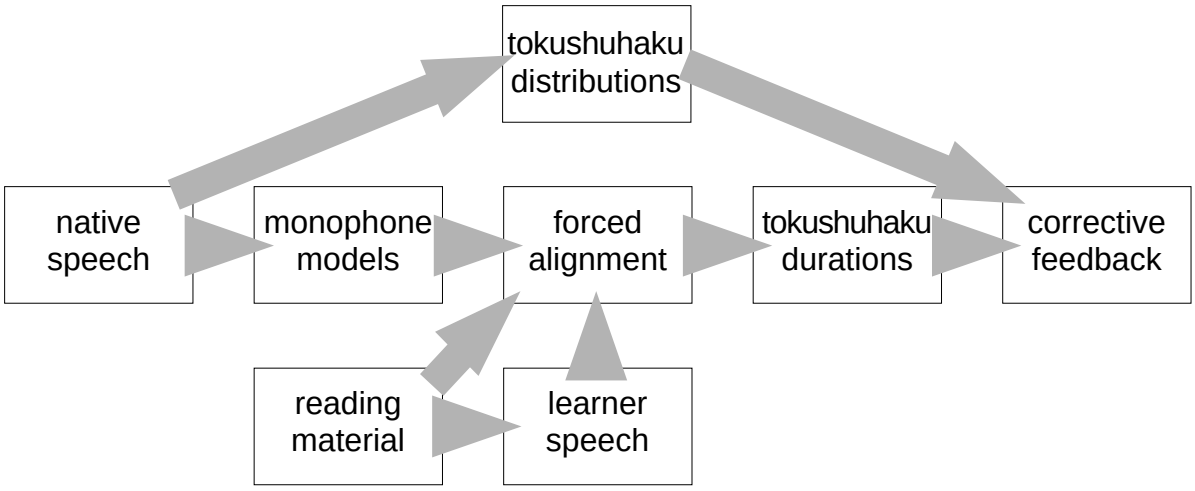


Fig. 4 Process flow of the system

3. SYSTEM DESCRIPTION

Figure 4 shows the overall structure of our system. Process flow is as follows. First, known reading material is shown to the learner. Next, the learner's speech is forced-aligned by the speech recognizer (i.e., phone boundary locations are obtained with respect to the beginning of the utterance given a correct transcription or similarly tightly constrained language model of the utterance), and phone durations are measured. Finally, each phone duration is compared with the correct duration ranges described above in section 2 (see figures 1, 2 and 3), and feedback is sent to the learner.

For speech recognition, HTK was used to generate monophone models trained on the same speech data used for tokushuhaku analysis in section 2 [8]. Spectrally identical tokushuhaku and non-tokushuhaku were trained and recognized as one phone (for example, [a] and [aa] were trained using a single phone model [a]). Prior knowledge of the reading material is used to determine whether a phone was a tokushuhaku or not. Audio input is 8 bit mulaw sampled at 16 kHz, using a desktop electret condenser microphone. The entire system runs on a Sun workstation in realtime. Each pronunciation practice turn takes 6 seconds (3 seconds record, 3 seconds playback).

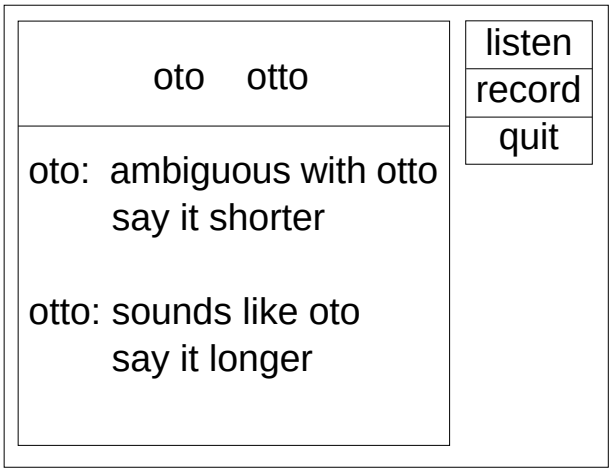


Fig. 5 System's user interface (translated)

The reading material of this system is comprised of minimal pairs of actual words, for example "oto" (sound) and "otto" (husband). The learner may choose at any time to listen to a native speaker's recording of the reading material. After the learner reads the word pairs, she receives corrective feedback immediately. For instance, if the learner cannot distinguish the pair, the feedback might be "Your oto is excellent, but otto sounds the same as oto. Say otto longer." (All examples of feedback in this paper are given in English for explanation purposes.) If otto is not long enough, the feedback might be "Your otto is ambiguous with oto -- say otto longer". If the learner hypercorrects, the feedback might be "Your otto is way too long -- make it shorter". Some learners read the word pair in the wrong order, which elicits "Your oto sounds like otto. Your otto sounds like oto". In all cases, the feedback is one of "say it longer", "say it shorter", or "excellent". This kind of feedback is straightforward to understand regardless of the learner's educational background. Feedback can optionally include phone durations in milliseconds, which is helpful for technology-oriented learners to control their phone durations. Figure 5 shows the graphic user interface along with feedback to the student.

Some students complained that the system is unforgiving, because a mere 10 ms difference in duration can result in negative feedback. Instead of using rigid boundaries to divide acceptable and unacceptable duration zones, incorporating fuzzy logic might help steer students towards correct pronunciation. Graphical representations of target zones may be useful as well.

4. SYSTEM EVALUATION

Two experiments were run to determine (1) how accurately the system measures phone duration, and (2) how useful the system is for teaching tokushuhaku.

By comparing hand-labeled and system-measured segment durations of 1176 phones collected from 3 non-native speakers, we found that the median difference was 30 ms (i.e., half of the durations were measured at differences of 30 ms or less). In particular, at least two-thirds of voiceless plosives were measured at differences within 10 ms (10 ms being one frame width for the speech recognizer). Given that the durations of tokushuhaku and non-tokushuhaku differ at magnitudes significantly larger than 10 ms, measurement differences of 10 ms seem well-within the acceptable threshold. On the other hand, at least two-thirds of the voiceless fricatives [h][s] were confused with noise. Measurement differences often exceeded 100 ms. It is thus necessary to design reading material using phones whose durations can be measured with sufficient accuracy.

The system's usefulness is determined by whether the system helps teach pronunciation skills that are not being taught due to lack of time in the TJSL classroom. The system is not intended to replace teachers; rather it is designed to supplement instruction which is impractical by teachers. Comparing the performance of the system with human teachers is unrealistic. Instead, the system's usefulness can be suitably measured by the length of time learners spend acquiring tokushuhaku production skills. The system can be judged successful if students acquire skills within the timeframe allowed in the TJSL curriculum, i.e., if pronunciation training keeps pace with subject matter being taught in the classroom, such as orthography, vocabulary and syntax.

To determine the amount of time necessary for tokushuhaku acquisition, we observed changes of tokushuhaku duration over time. Figure 6 shows an example of durations of [t] and [tt] from the word pair "oto otto" spoken by a learner from Anhui, China, who had been in Japan for 6 months. The subject's renditions began to fall within acceptable ranges (shown as shaded areas on the chart) after about 4 minutes of practice.

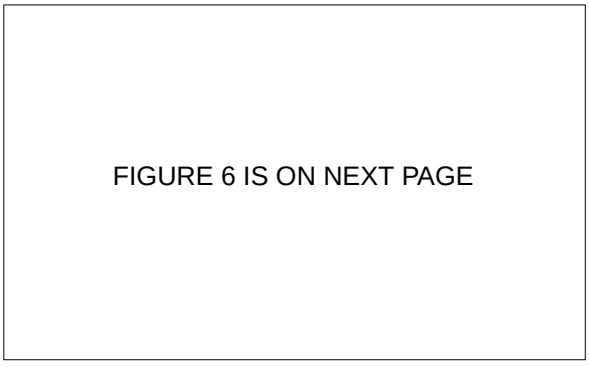


Fig 6. Phone durations of a single subject over time

5. CONCLUSION

A CALL system using speech recognition to measure segment duration was found to be useful for teaching tokushuhaku phones. The system is significant because it is the first CALL system to provide corrective feedback similar to what human teachers would provide when teaching Japanese pronunciation. The feedback is easy to understand and follow. The system provides an intensive environment for pronouncing tokushuhaku, which helps the learner acquire skills in a short period of time.

Improvements in speech recognition robustness may yield higher accuracy when measuring duration; in particular, some phones were easily confused with noise. Intelligent language models may become necessary when the system is used with novice learners who exhibit significant hesitation phenomena and other disfluencies. Comprehensive evaluation in the TJSL classroom is required to determine the system's usefulness with respect to the learners' native languages. Another area of experiments includes the retention, attrition and recovery of tokushuhaku skills.

With respect to applying speech recognition technology to TJSL, the next step of this research is towards the teaching of pitch accent. When the pitch patterns of Japanese words are incorrectly pronounced, usually no semantic confusion results, but the speaker is negatively categorized as of either rural or foreign origin. Precise pitch extraction and segment alignment will be needed to develop a CALL system for training pitch accent production.

6. REFERENCES

[1] Imagawa, H. et al “Realtime extraction of pitch and formants using DSPs and its applications to pronunciation training” Technical Report of the IEICE, SP89-36. 1989

[2] Saida, I. et al “Speech CAI for non-native speakers of Japanese” Annual report of the study group on Japanese prosody and its instruction, 5-15. Tokyo: Ministry of Education. 1993

[3] “Triple Play Plus! Japanese, version J1.2” Syracuse Language Systems, 1996

[4] Taniguchi, H. “Results of survey on teaching of Japanese pronunciation” Proceedings of the symposium on Japanese prosody and its instruction. Tokyo: Ministry of Education. 1991

[5] Muraki, M. et al “The mora nasal and mora obstruents pronounced by native speakers of English or Chinese” In Sugitou, M. ed “Japanese phonetics and phonology” Tokyo: Meiji Shoin. 1990

[6] Takeda, K. et al “Japanese speech database for research purposes” Kyoto: ATR. 1988

[7] Hibiya, J. “Pronunciation teaching outside of Japan” J. of the Phonetic Society of Japan, 211:43-48. 1996

[8] Young, S. et al "The HTK Book" Cambridge University. 1995

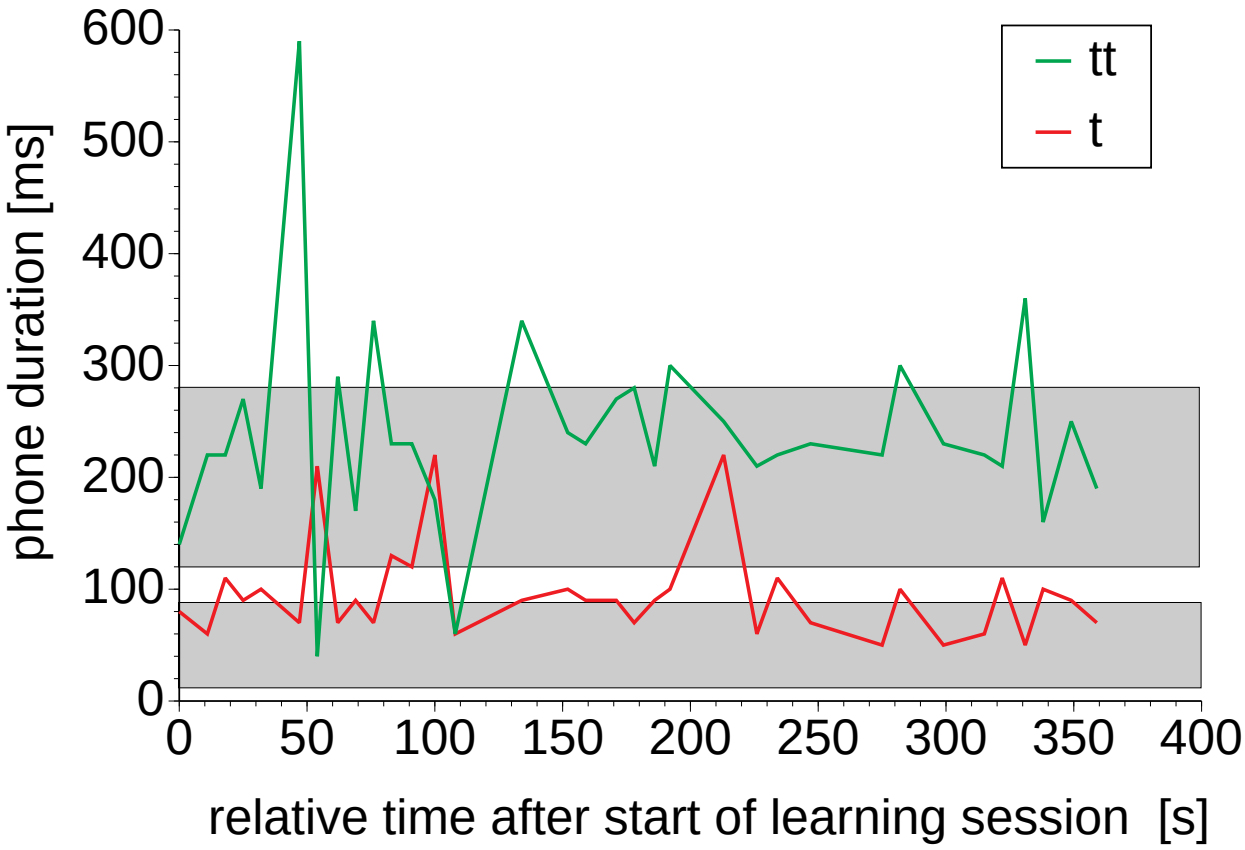


Fig 6. Phone durations of a single subject over time