日本語発音教育への音声認識の利用

河合　剛　広瀬啓吉

東京大学 工学部 電子情報工学科
〒113-8656 東京都 文京区 本郷 7-3-1
email: goh@kawai.com　hirose@gavo.t.u-tokyoac.jp

あらまし　日本語を母語としない者に日本語を教える教育現場では特殊拍(長母音、促音、撥音の総称)の指導が要求されている。特殊拍と非特殊拍とは音素対立するから聞き取りと発音が大切である。しかし特殊拍が日本語固有の発音現象であるため、多くの学習者は母語にかかわらず特殊拍の習得を難しく感じる。しかも日本語教育現場での発音の学習時間が限られているので習得度が低い。教育現場の現実的制約のもとでは学習を自動化して学習効率を高める以外に習得を促す方法がない。しかるに従来の発音自習システムは矯正フィードバックがない点で人間の教師による指導に劣る。たとえば、カセットテープを用いた聴取と発声の教材は発音の適否を自分自身で判断せねばならない。この欠点は音声解析手法を発音教育へ応用した例にも共通している。

幸い、日本語を母語としない学習者の多くは日本語の音そのものをほぼ正しく発音できる場合が多い。たとえば、中国語母語話者の特殊拍の発音誤りのほとんどは特殊拍長(特殊拍の持続時間)の誤りである。単音の発音が正しいのであれば、特殊拍長を誤った場合の矯正フィードバックは「長くのばして発音せよ」「短くちぢめて発音せよ」のように平易明解にできる。そこで特殊拍長を音声認識を用いて測定し、特殊拍長の長短に応じて発音の修正方法を明示する教育システムを作成した。

キーワード　コンピュータ支援言語学習、日本語教育、発音、音声認識、音長、特殊拍

APPLYING SPEECH RECOGNITION TO THE TEACHING OF JAPANESE PRONUNCIATION

Goh Kawai and Keikichi Hirose

University of Tokyo
Department of Information and Communication Engineering
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan
email: goh@kawai.com　hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT　A CALL (computer-aided language learning) system for teaching the pronunciation of Japanese tokushuhaku (long vowels, the mora nasal and mora obstruents) to entry-level learners was developed for TJSL (teaching Japanese as a second language). Conventional self-study methods for pronunciation learning do not tell learners what their mistakes were, whether their speech is intelligible, or what they can do to improve their pronunciation.

The system proposed in this paper uses speech recognition to measure the durations of tokushuhaku phones produced by learners. Tokushuhaku and non-tokushuhaku are spectrally almost identical but their phone durations differ significantly. The learners' durations are compared with distributions of native speakers' perceptions of varying tokushuhaku durations. The CALL system returns learners an intelligibility score that shows the percentage of native speakers who will understand the learner's pronunciation. The learner can terminate training when his communicative performance has met his expectations. For instance, when a learner hits a learning plateau, intelligibility indices can help him decide whether further learning effort is worthwhile. Given that most adult learners can never attain complete nativeness, it is of practical use to be told when nonnative accents cannot be removed further.

Learning experiments 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 curricula.

KEYWORDS　CALL, TJSL, pronunciation, speech recognition, duration, tokushuhaku

1. FOREWORD

The problem addressed is assisting TJSL (teaching Japanese as a second language) teachers in improving their learners' acquisition and retention of pronunciation skills. This is an important issue because spoken language is the most commonly used form of communication. Self-study methods are required because there are not enough teachers to teach students individually. Previous self-study systems forced the learner to evaluate his pronunciation himself — an impossible task given that if the learner could judge pronunciation quality he would have no need to learn it to begin with.

The proposed solution is a CALL (computer-aided language learning) system to train the pronunciation of entry-level nonnative learners. The objective is to build a system that explains to the learner (1) what his mistake was, (2) the percentage of native speakers who will understand his utterance the way he intended it to be understood, and (3) how to correct his mistake.

The proposed system uses speech recognition algorithms to accurately measure the duration of tokushuhaku, a set of phonemically distinct phones most nonnative learners have difficulty with. Focusing on measurable, identifiable pronunciation skills achieves high reliability and validity. By coupling speech recognition technology with quantitative knowledge of how native speakers perceive pronunciation differences, the learner receives training similar to professional instructors. This is the first CALL system for Japanese pronunciation learning that provides feedback similar to human teachers.

This remainder of this paper discusses a statement of the problem and its significance (section 2), an overview of the literature (section 3), the proposed solution (section 4), and conclusion (section 5).

2. THE PROBLEM

The past decade has seen an influx of nonnative speakers entering Japan. Jobs, education, childcare, shopping — the language barrier hinders communication in all aspects of the newcomers' daily lives. Learning to speak Japanese fluently behooves good hearing and listening skills — skills not effectively learned by studying written language. Nonnatives (especially Asians) speaking with an accent are occasionally branded as inferior undesirables [11]. Good pronunciation skills are essential for succeeding in Japan.

Yet only token attention has been paid to pronunciation teaching. The time spent in classrooms practicing pronunciation is minimal. So is the quality of teaching material. According to a survey conducted on 158 TJSL teachers in the nation, most teachers teach pronunciation in some form or another, but limit classroom activities to less than 10 hours total, typically concentrated at the very beginning of entry-level courses [9]. The lack of instruction time is mainly due to the shortage of hours both learners and teachers can afford. Learners are eager to learn useful Japanese as quickly as possible — an understandable desire given they are already in the country. Teachers are under pressure to prepare learners for Japanese language proficiency tests [3]. These tests primarily measure written skills and listening comprehension. Oral production is not a major factor because measuring speaking skills is unreliable without absurdly intense effort [12]. A common misconception that reading and writing Japanese is hard but speaking it is easy does not help. As a consequence, classroom instruction has concentrated on orthography, vocabulary and syntax. It is not an exaggeration to state that students' abilities are measured by how many kanji they can read and write. Students are lucky if their teacher uses a tape recorder or hiragana chart for pronunciation practice. Most teachers do not use textbooks or teaching aids at all [9]. TJSL practitioners both here and abroad are clamoring for a systematic syllabus for Japanese pronunciation training [6].

It is ironic that after months of hard work and a well-deserved Japanese language proficiency certificate in his hand, a nonnative speaker seeking a job would be turned down at his first interview. His poor oral skills suggest he knows less Japanese than he really does. Even if he is hired, his native speaker colleagues may not let him into the communication loop — "He won't understand," they may decide among themselves. Cast away from the social fabric, the nonnative speaker is doomed. The final insult is that by this time it is too late for remedies. His pronunciation mistakes have become solidified, incapable of change.

This perhaps overly melodramatic tragedy can be avoided by using self-study methods for pronunciation skills. The acute shortage of classroom time and the appalling lack of teaching material can be rectified while simultaneously strengthening the acquisition and retention of correct pronunciation.

Implementing such a self-study system is challenging in several ways. First, we need to target pronunciation mistakes that either occur frequently and/or cause communication confusion. Second, we need to measure these mistakes in consistent, meaningful terms. Third, we need to unambiguously instruct the learner how to

correct these mistakes. These three conditions are necessary for the learning system's reliability and validity. As detailed pronunciation curricula for TJSL do not exist, we must build our own.

The next section of this paper critically reviews the literature.

3. THE LITERATURE

3.1. Overview

This section reviews systems for spoken language self-study. We start with a classic example of comparing native and nonnative speech by ear. We then discuss a system that displays the speech waveform, pitch track, and formants of both the native's model and the nonnative's rendition. Next, we review an example that uses speech recognition technology to understand the learner's speech. The system engages the learner in a simulated dialogue, and forces the learner to speak. Last, we review an example that uses speech recognition technology to measure the quality of the learner's pronunciation. This system quantifies pronunciation quality at the phoneme level using metrics derived from acoustic distances.

All systems reviewed except the last are for TJSL, and are representative of the state of the art including systems for the training of non-Japanese languages.

3.2. Comparison by ear

The classical pronunciation self-study method is using a tape recorder [10]. The learner listens to native speech, repeats and records it, and compares his utterance with the native's. Advantages of this method are that the learner can practice listening skills and that the equipment can be portable. Disadvantages are that the learner is not forced to say anything, and that the learner receives no corrective feedback from the system. The learner must be highly motivated and have a good ear for foreign languages.

Learners rarely have these qualifications. Indeed, a common shortcoming of all existing self-study methods including the tape recorder method is that none tell learners whether their speech is intelligible or what they can do to improve their pronunciation. If learners could judge the appropriateness of their renditions by themselves, they would have no need to learn pronunciation in the first place. Learners need to know what their mistakes are, how serious their mistakes are, and how to correct their mistakes. Self-study systems that impose learners to judge the accuracy of their productions are fundamentally flawed.

3.3. Comparison assisted by speech processing

Imaizumi et al proposed a system that displays the speech waveform, pitch track, formant trajectories, and other features derived from the learner's utterance [2]. These features are graphically displayed on a computer screen along with previously processed and stored images of the native's model. The learner changes his articulation so that his features match that the native model's. There is no specific instruction on how this matching might be accomplished.

Imaizumi's system suggests various possibilities to automated pronunciation training, but the effectiveness of the system is unclear. This is because the system was not intended primarily as a pronunciation teaching aid. It was proposed as an extension to Imaizumi's existing speech analysis program for F0, F1 and F2 feature extraction; as such, Imaizumi did not run educational experiments. The lack of instruction on how to alter the learner's speech with the native's leads one to imagine that the system may not be effective.

Another concern is that the native model is a single utterance of a particular individual. If the learner's speech waveform, formant trajectories and so forth were to overlap completely with the native model's, then the learner must sound exactly like that particular native speaker, mirroring the native's idiosyncratic speech mannerisms and voice quality. Imaizumi's system risks turning into a method for voice actor training unless a native speaker is chosen carefully, perhaps by using multiple native speakers over the course of study, or by pairing natives with nonnatives having similar physical characteristics.

Imaizumi's system can be expanded to include more speech processing information. But the learning system should not demand learners to acquire new skills unrelated to pronunciation practice. For instance, displaying spectrograms should be avoided, because learning to read spectrograms is unessential for learning pronunciation. (Many TJSL learners lack technical backgrounds.) The learning system should concentrate on analyzing pronunciation errors and suggesting remedies. The learner should follow the system's suggestions. In this regard, even relatively simple processing

of speech (such as time waveforms) may be unsuitable.

3.4. Interactive dialogue

The system designed by Ehsani et al allows user-system interactions to be completely oral, thereby forcing the learner to speak [1]. The system simulates situation-based, system-learner dialogues. Learners reply to the system's verbal prompts by saying whole sentences. Ehsani's dialogues has exactly one correct path the student should follow. Errors elicit corrective feedback for content; for instance, when the student says "Pleased to meet you," the system might suggest "This isn't the first time you meet."

The learner needs to understand the conversation's context, find the correct answer, and say it more or less correctly. The problem is, as far as pronunciation learning is concerned, that there is no incentive to speak more clearly or intelligibly. The system does not grade pronunciation. Insofar as the learner reads one of the offered sentences aloud, the system will accept his pronunciation. The acceptance threshold seems to have little if anything to do with the intelligibility of the utterance as perceived by native speakers.

Ehsani's system is essentially a sentence recognizer. The instructor scripts a dialogue consisting of system prompts and student responses. The instructor anticipates correct and incorrect student utterances based on the instructor's knowledge of the learner's linguistic skills. The student does not choose from a set of sentences shown on the computer screen; he is free to respond to the system's verbal prompts in any way. The strong advantage of this approach is that the learner must actively generate a correct response instead of passively choosing one from a system-provided list. The equally strong disadvantage is that if the dialogue is not scripted correctly, the student can say something unforeseen by the system, causing the interaction to fail. The system's lack of responsiveness was cause for concern during field trials. In general, giving feedback in realtime is important in pronunciation training because the learner must receive corrections immediately after speaking or he will forget how he articulated.

3.5. Pronunciation evaluation

Witt's system was designed for British English but studying it is useful because the basic algorithm is language-independent [13]. Witt defines the goodness for each phone in the utterance as the posterior probabilities that the learner uttered phone p given the acoustics $\boldsymbol{O}$ and the set of all phones $\boldsymbol{Q}$. He assumes that (1) all phones are equally likely, and (2) the total likelihood of all phones in $\boldsymbol{Q}$ yielding the acoustics $\boldsymbol{O}$ can be approximated by the maximum likelihood of any single phone yielding $\boldsymbol{O}$. (Strictly speaking, neither of these assumptions is true in any language, but is chosen as a first approximation.) Witt's goodness metric can quantify the pronunciation quality of each phone individually or as a group; for instance, a particular token of q can be analyzed on its own, or all phones of type q occurring in the utterance can be treated at once.

Witt's method is one of the most systematic approaches to automated pronunciation grading to date. Its weakness is relying on a single probability likelihood measure, which in turn is based on multidimensional acoustic distance measures. Analyzing the acoustic signal of a phone or an entire utterance, either by calculating the distance in acoustic feature space or by evaluating the probability likelihood of nonnative speech being produced given native HMMs, is incorrect because the learner's speech is a mixture of characteristics specific to the individual (such as voice quality) and the learner's native language (namely its phonology).

To improve the speech recognizer's performance, Witt uses speaker adaptation [5]. This method is risky because it does not necessarily guarantee separation of speaker-specific characteristics from that of the training-data population. In the case of well-trained native speaker HMMs, a particular native speaker (to whom we wish to adapt HMMs) differs only on speaker-specific characteristics; the rest are language-common features shared with the HMMs. This is not true for nonnative speakers, again because their speech is heavily influenced by their native language's phonology as well as their particular personal features. It is probably better to use both native and nonnative phone models to quantify the proximity of the learners's speech to his native and target languages.

3.6. Summary

Although various proposals have been made for TJSL pronunciation teaching, none provide the learner with useful feedback such as "Your pronunciation is intelligible," "Native speakers will not understand you," "You should do such and such to improve your pronunciation," and so forth. The system ought to determine when nonnative accents cannot be removed further — given that most adults can never attain complete nativeness, it would be useful if the system were to say "Your pronunciation is at a level where native speakers will un-

derstand you. However, your skills are not likely to improve beyond this point. I suggest you stop practicing at this time." A system that explains how to improve and indicates when to stop training is a system that teaches similarly to human instructors. The next section of this paper describes a proposed CALL system with this goal in mind.

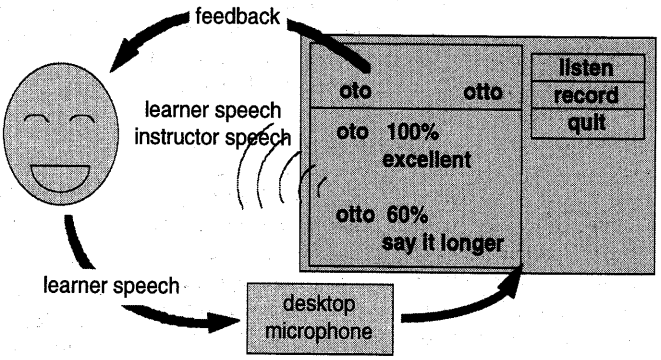

Fig. 1 System-user interaction

4. THE SOLUTION

4.1. Overview

The long-term goal of my research is to discover how adults can efficiently learn to pronounce nonnative languages. By choosing pronunciation problems that occur frequently and/or disrupt communication seriously, and by implementing a CALL system that teaches how to avoid such mistakes, automated pronunciation learning can be shown to be feasible, reliable and valid.

A sensible research strategy is to implement pronunciation CALL systems that have features extendable to pronunciation problems found in many languages. It is advantageous to concentrate on pronunciation errors at the phone or word level rather than at the sentence level, because the CALL system needs high reliability and validity. This can be achieved by targeting subskills, such as particular phoneme sets, rather than judging an entire sentence as an amalgam. Once a method for phone-level and word-level pronunciation learning has been established, we can tackle sentence-level issues.

This paper proposes a method to deal with a particular pronunciation skill in Japanese. We developed a CALL system for teaching the pronunciation of Japanese tokushuhaku (long vowels, the mora nasal and mora obstruents) to entry-level learners. Tokushuhaku and non-tokushuhaku are spectrally almost identical but their phone durations differ significantly. Our system uses speech recognition to measure the durations of tokushuhaku phones produced by learners. The learners' durations are compared with distributions of native speakers' perceptions of varying tokushuhaku durations. The CALL system returns learners an intelligibility score that shows the percentage of native speakers who will understand the learner's pronunciation. The learner can terminate training when his communicative performance has met his expectations. For

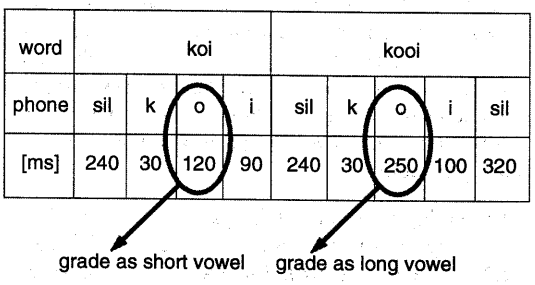

Fig. 2 Phone duration measurements

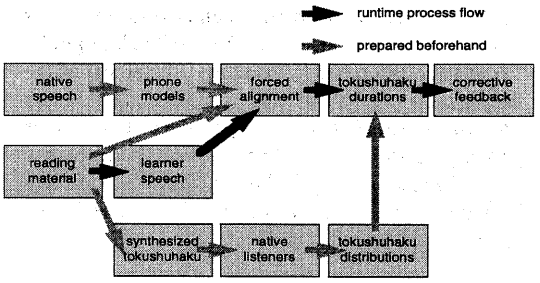

Fig. 3 System's process flow

instance, when a learner hits a learning plateau, intelligibility indices can help him decide whether further learning effort is worthwhile.

4.2. System overview

The overall system-user interaction is shown in figure 1. First, known reading material is presented 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 tokushuhaku phone durations are measured (figure 2). Each duration is compared with results from perception experiments ran on native speakers (this procedure is explained in section 4.3). Feedback to the learner consists of (a) an intelligibility score showing the percentage of native speakers who will understand the learner's pronunciation, (b) instructions on whether to lengthen or shorten the tokushuhaku, and optionally, (c) the tokushuhaku duration in milliseconds. An example of a feedback display is included in figure 1. The process flow of the system is shown in figure 3.

For speech recognition, HTK v2.1 [14] was used with gender-dependent phone models based on [8]. 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).

The reading material of this system is comprised of minimal pairs of actual words, for example "kado" (corner) and "kaado" (card). The learner may choose at any time to listen to a native speaker's recording of the reading material. Doing so tends to sway the learner's speech rate towards the native model's. After the learner reads the word pairs, he immediately receives an intelligibility score and instructions on how to correct his pronunciation. For instance, the feedback might be "Your kado can be understood by 100 percent of native speakers, but your kaado can be understood by only 10 percent. Say kaado longer." Depending on the phone's duration, the system instructs the learner to "say it longer" or "say it shorter." This kind of feedback is straightforward regardless of the learner's educational background.

4.3. Tokushuhaku intelligibility

My previous research showed that the teaching of tokushuhaku can be automated using tokushuhaku duration information [4]. Although durations of tokushuhaku produced by native speakers are available, the exact durations of auditory stimuli perceived by native speakers as tokushuhaku remained unclear. The latter information is useful in language pedagogy because it can determine phone duration ranges that are unanimously perceived by native speakers as tokushuhaku. Language learners can target these ranges during pronunciation practice. Conversely, ambiguous durations indicate lower intelligibility, and the level of confusion can be used to grade the learners' productions.

To clarify what durations are unambiguously perceived by native speakers as tokushuhaku for particular cases, I ran perception experiments of native speakers judging artificially altered tokushuhaku durations.

Minimal pairs of actual Japanese words differing solely on the presence or absence of tokushuhaku were chosen. Next, each word was synthesized in isolation with 13 varying tokushuhaku durations. Vowel durations were adjusted at roughly constant ratios. Preplosive closures were varied at 20 ms steps. For nasals, a moraic nasal of varying length was combined with a non-moraic nasal of fixed length (45 ms). A shallow falling pitch based on the Fujisaki model was added to all words, but no lexical pitch accent was used. Table 1 shows the minimal pairs along with the minimum and maximum phone durations created by the terminal analog speech synthesizer [7]. Double phones such as [aa] denote tokushuhaku; single phones such as [a] are non-tokushuhaku. Durations for each synthesized word were measured by hand.

Table 1 Minimal pairs and their synthesized tokushuhaku duration ranges

word pairs		duration [ms]	
		min	max
kado	kaado	40	328
nasu	naasu	37	360
biru	biiru	60	288
chizu	chiizu	64	197
kuro	kuuro	44	250
kutsu	kutsuu	12	191
kaite	kaitee	45	335
seki	seeki	44	345
koi	kooi	45	288
toru	tooru	50	301
supai	suppai	20	260
hata	hatta	20	260
ita	itta	20	260
haka	hakka	20	260
kokee	kokkee	20	260
ichi	icchi	20	260
sachi	sacchi	20	260
kona	konn-na	30	200
hone	honn-ne	30	200

All 13 different varieties of each word were played twice in random order. Twelve native speakers of Japanese were asked to categorize the words as a word containing tokushuhaku, not containing tokushuhaku, or neither of the above.

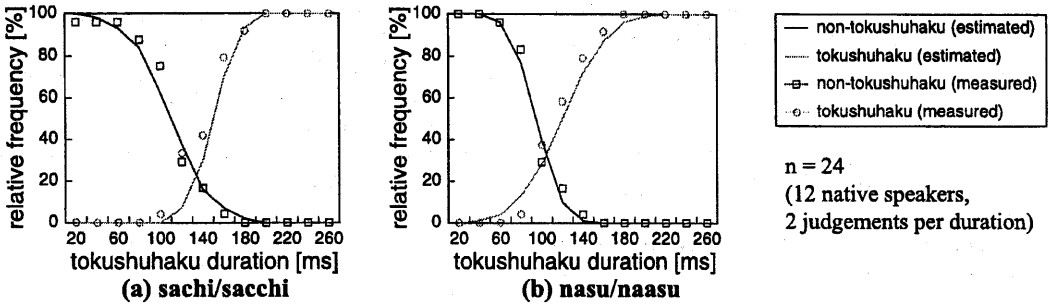

Fig. 4 Relative frequency of subjects' responses to various tokushuhaku durations

Some frequency plots of the subjects responses are shown in figures 4(a) and (b). There was almost perfect agreement among subjects with regard to short and long durations. As expected, mid-range durations were judged as ambiguous, as were overly short or long durations. Discrimination curves closely matched normal distributions (table 2). Tokushuhaku and non-tokushuhaku are clearly distinguishable.

The percentages of curves shown in figures 4(a) and (b) can be interpreted as intelligibility indices based on tokushuhaku duration. For instance, the tokushuhaku curve in figure 4(b) can be interpreted as the percentage of native speakers understanding a hypothetical learner's rendition of "naasu" produced with varying tokushuhaku lengths. The level of agreement among native speakers as a particular phone being tokushuhaku or not indicates the appropriateness of that phone. By knowing the learner's intention beforehand, we can provide corrective feedback that quantifies the likelihood of the learner being understood correctly.

Table 2 Means and standard deviations of tokushuhaku durations

kado	N(56,19)	kaado	N(192,78)
nasu	N(95,20)	naasu	N(130,35)
biru	N(85,22)	biiru	N(200,55)
chizu	N(80,17)	chiizu	N(147,28)
kuro	N(63,20)	kuuro	N(168,52)
kutsu	N(41,26)	kutsuu	N(130,43)
seki	N(57,19)	seeki	N(177,103)
kaite	N(68,29)	kaitee	N(218,71)
koi	N(64,94)	kooi	N(190,60)
toru	N(72,27)	tooru	N(194, 69)
supai	N(85,12)	suppai	N(185,51)
hata	N(52,38)	hatta	N(192,48)
ita	N(50,37)	itta	N(192,45)
haka	N(45,5)	hakka	N(108,53)
kokee	N(45,32)	kokkee	N(185,48)
ichi	N(42,41)	icchi	N(185,57)
sachi	N(110,35)	sacchi	N(150,20)
kona	N(83,13)	konn-na	N(149,152)
hone	N(40,17)	honn-ne	N(110,45)

4.4. Reliability

In order to determine how accurately the system measures phone duration, we compared hand-labeled and system-measured segment durations. Due to improved HMMs, significantly higher accuracy was obtained compared to my previous work. Comparing approximately 1200 phones obtained from 3 nonnatives showed that practically all durations of phones occurring within a word were measured at differences at or below 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.

5. CONCLUSION

Presenting learners with an intelligibility score that shows the percentage of native speakers who will understand the learner's pronunciation is significant improvement over conventional techniques, which at most merely return a good/bad categorical result. The new method informs the learner how far he has progressed in easy-to-grasp terms.

Mistakes in phone quality are detected using a speech recognizer incorporating bilingual monophone models of both the learner's native and target languages. HMMs for the two languages are trained separately on language-dependent speech data, but are bundled together during recognition so that the closest phone recognized indicates the nonnativeness of the utterance should the recognized phone not be Japanese. Knowing the phonetics and phonology of the learner's native language can identify nonnative articulatory gestures that result in Japanese pronunciation errors, thus allowing precise corrective feedback to the learner. The

paper includes an overview of the system, its reliability, validity, and effectiveness in the foreign language classroom.

ACKNOWLEDGMENT

We thank Kazuya Takeda for providing us with monophone HMMs [8].

REFERENCES

[1] Ehsani, F. et al "Subarashii: Japanese interactive spoken language education" Proc. Eurospeech-1997 (Rhodes, Greece), 681-684, 1997

[2] Imaizumi, 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

[3] Ishii, E. "Japanese language education in Japan: issues and possible improvement" J. of Japanese Language Teaching, vol. 94, 2-12, 1997

[4] Kawai, G. and Hirose, K. "A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruents" Proc. Eurospeech-1997 (Rhodes, Greece), 657-660, 1997-9

[5] Leggetter, C. et al "Speaker adaptations of HMMs using linear regression" Report CUED/F-INFENG/TR.181, Cambridge University, 1994

[6] Ramalakshmi, V. et al "Panel discussion: TJSL pronunciation education abroad" in "The Japanese language in international society", Ministry of Education, 146-163, 1997

[7] Sakata, M. "Analysis and synthesis of prosodic features in spoken dialogue of Japanese" Unpublished master's thesis: University of Tokyo, 1995

[8] Takeda, K. et al "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model" IPSJ SIG Notes, 97-SLP-18-3, 1997

[9] Taniguchi, H. "Results of a survey on Japanese pronunciation teaching" in "Prosody and its role in TJSL", Ministry of Education, 17-21, 1991

[10] Toki, S. et al "Japanese exercises for nonnative speakers vol. 12: pronunciation and listening (with cassette tape)" Tokyo: Aratake Shuppan, 1988

[11] Toki, S. "Pronunciation teaching" Teramura, H. ed "Japanese and Japanese language teaching" vol. 13, Meijishoin: Tokyo, 111-138, 1989

[12] Woodford, P. "Language testing at ETS: its development and evaluation" J. of Japanese Language Teaching, vol. 94, 160-170, 1997

[13] Witt, S. et al "Language learning based on non-native speech recognition" Proc. Eurospeech-1997 (Rhodes, Greece), 633-636, 1997

[14] Young, S. et al "The HTK Book" Cambridge University, 1996