English and Japanese CALL Systems Developed at Kyoto University

Tatsuya Kawahara*, Hongcui Wang*, Yasushi Tsubota*, and Masatake Dantsuji*
* Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—This paper gives an overview of the English and Japanese CALL systems which have been developed at Kyoto University. Both systems incorporate automatic speech recognition (ASR) technologies to detect pronunciation errors. In order to cope with non-native speech, error prediction mechanisms are prepared based on linguistic knowledge and corpus-based decision tree learning. Several choices of acoustic modeling for non-native speech including erroneous pronunciations are also investigated. The English CALL system is designed for Japanese college students so that they can introduce Japanese cultures to foreign people, thus the acoustic model and error prediction are tuned to the specific native language (L1=Japanese). On the other hand, the Japanese CALL system is for foreign visitors of any L1, and focuses on basic-level sentence production and adopts GUI for easy practice.**

I. Introduction

Second language learning has become very important in the modern globalized society, in which tremendous amount of information is exchanged globally and in almost real-time. Computer-assisted language learning (CALL) provides an effective learning environment so that students can practice in an interactive manner using multi-media content, either with the supervision of teachers or on their own pace in self-learning. The advancement of speech and language technologies has opened new perspectives on CALL systems, such as automatic pronunciation assessment and simulated conversational-style lessons.

With incorporation of automatic speech recognition (ASR), CALL systems have been used for pronunciation learning, specifically evaluating pronunciation and correcting errors, such as the system in [1], FLUENCY [2], WebGrader [3], and EduSpeakTM [4]. One of the most significant problems in this scheme is accurate recognition and error detection of non-native speech. While the ASR system needs to adapt to non-native speech, it must detect critical errors in terms of intelligibility, by predicting possible errors effectively. We have approached this problem both in acoustic modeling and language modeling. These techniques are reviewed in this paper. We also present an overview of the CALL systems we have developed and deployed at our university.

The Academic Center for Computing and Media Studies (ACCMS) of Kyoto University introduced CALL systems in 1998, the first in major universities in Japan. Since then, we have been working on the advanced CALL using ASR technologies. Our major targets have been English CALL and Japanese CALL although we have been engaged in other languages such as Chinese, French and German languages.

The English CALL system is designed for Japanese college students. The content of the system is Japanese cultures such as temples in Kyoto, so that students can explain them by themselves to foreigners. Although Japanese students have been studying English for more than six years before admitted to universities, their English communication skill, for example measured by TOEFL and other standard tests, is very low compared with students in other countries, partly because Japanese and English languages are much different in terms of the phonetic and grammatical structures. Therefore, the English CALL system is focused on Japanese students. By limiting the native language (L1), we can prepare a dedicated acoustic model and error prediction/feedback mechanisms. Specifically, we exploited a database of Japanese speakers for acoustic modeling, but there are a number of erroneous pronunciations that are not faithfully labeled. Thus, several choices of training and adaptation schemes were investigated and compared. Error prediction rules were devised based on linguistic knowledge to realize robust error detection in Japanese-accented English. Moreover, we incorporated automatic error detection of stresses, in which Japanese students have much difficulty. The system has been used in CALL classes in Kyoto University, and we have found and fixed a number of technical problems.

The Japanese CALL system is designed for foreign students coming to Japan, focusing on elementary levels for their survival in Japan. Although the lesson content is relatively easy, the system does not assume any particular native language (L1). Since it is difficult to devise universal error prediction rules, we turned to a data-driven method; decision tree learning was introduced to find critical error patterns, which optimize the balance of coverage and perplexity of the grammar network. This system has been tested and will be released to public when the full content is complete.

II. English CALL System: Hugo

A. System Overview

The English CALL system covers English learning in two phases: (1) role-play conversation and (2) practice of individual pronunciation skills. In the first phase, students play the role of a guide who provides information on famous events and landmarks in Kyoto, as shown in Fig. 1. As a guide, the student (B) answers questions asked by a native English speaker (A). Each question is presented to the student in audio/video format

Proceedings of the Second APSIPA Annual Summit and Conference, pages 804–810,
Biopolis, Singapore, 14-17 December 2010.

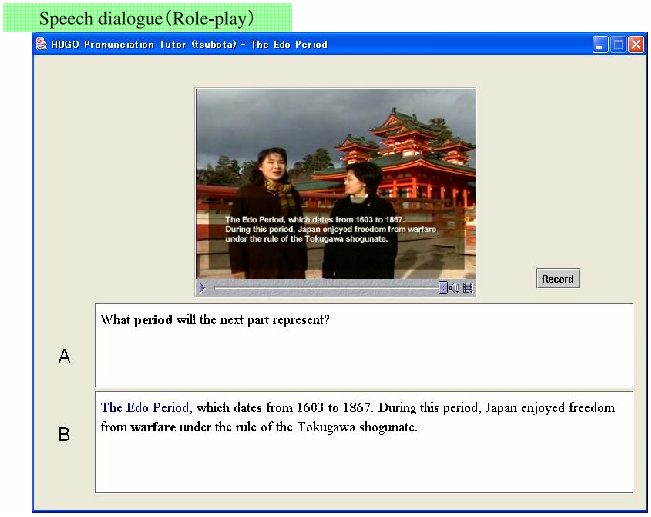

Fig. 1. Screen shot of role-play practice

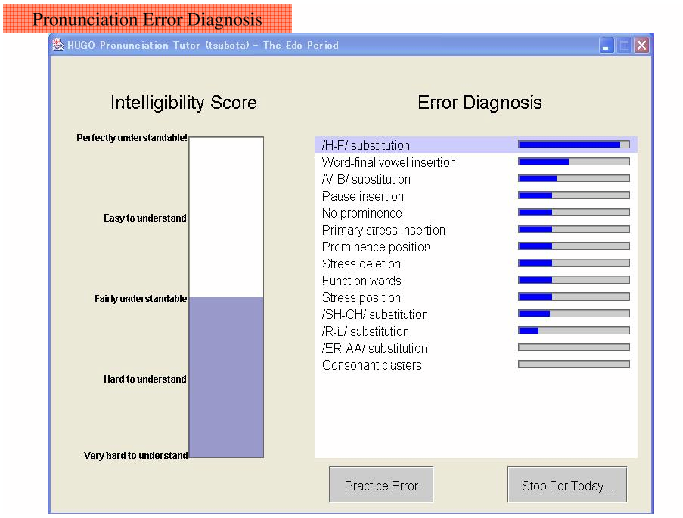

Fig. 2. Example of pronunciation profile

at the beginning of the practice session. The student records his/her spoken answers by following the script which appears on the screen. During the recording, the system works in the background to detect the student's pronunciation errors and stores a profile of his/her pronunciation skills. However, at this stage, the system does not inform the student of his/her errors so that the student can focus on the flow of the conversation.

At the end of the role-play session, the system provides a pronunciation profile for the student. It consists of two parts: (1) an intelligibility score and (2) priority scores for various pronunciation aspects. An example of the profile is shown in Fig. 2. The intelligibility score indicates how well the student's pronunciation would be understood by native speakers of English. It is computed from the error rates for the categorized pronunciation aspects, such as word-final vowel insertion and /r-l/ substitutions (right-hand side of Fig. 2), which cover typical errors made by Japanese students. To determine the priority of the error category for the student in the following practice, the system identifies the critical error categories for improving the intelligibility, based on the difference between the student's error rate and the average error rate of those in the same intelligibility level [5].

In the second phase, the student practices correcting the individual pronunciation errors identified by the above procedure. The training samples are chosen, based on the error category, from the sentences used during the role-play phase. In this phase, the student focuses on correctly pronouncing these words or phrases. During this stage, results of error detection and further instructions for correcting errors are presented.

B. Phoneme Error Prediction

To predict pronunciation errors, we modeled error patterns of Japanese students according to the linguistic literature [6]. The hand-crafted rule set includes 79 kinds of error patterns. There are 37 patterns concerning vowel insertions, such as which vowels are inserted between a certain pair of consonants or after the final consonant of words. In addition, there are 35 patterns for substitution errors. For deletion errors, we have 7 patterns: /w/, /y/, /hh/ deletion at the word beginning and /r/ deletion in some contexts [7].

C. Speaker Adaptation of Acoustic Model

Speech recognition and error detection in CALL is not easy since the speech of students using the system is different from that of native speakers. To compensate for acoustic variation, we can introduce speaker adaptation using Maximum Likelihood Linear Regression (MLLR) [8]. There is a problem in applying supervised adaptation in the case of CALL applications in which the students' pronunciation is not necessarily correct. Thus, we compared two phonemic labels for adaptation: lexicon labels (base form) and hand-labels counting pronunciation errors.

We prepared a native English model using the TIMIT database. The database was collected from eight major dialect regions of the United States. It contains a total of 6300 sentences (10 sentences spoken by 630 speakers). We trained monophone HMMs for 41 English phonemes. Each HMM has three states and 16 mixture components per state. The acoustic features consist of 12-dimensional MFCCs (Mel-Frequency Cepstral Coefficients), their Δs and Δpower.

For evaluation, we conducted phoneme recognition experiments with a corpus of English words spoken by Japanese students. The test corpus consists of 5950 speech samples. Seven Japanese speakers (2 male, 5 female) uttered 850 basic English words respectively. The database contains phonemic hand-labels, which were transcribed faithfully including erroneous pronunciations. For each speaker, 100 word samples were used for adaptation and the remaining 750 samples for evaluation.

Phoneme recognition rates are listed in Table I. Speaker adaptation with the lexicon labels was found to improve accuracy by about 5%, which is comparable to the result obtained using the hand-labels. Thus, we concluded it is

TABLE I
EFFECT OF SPEAKER ADAPTATION (PHONEME RECOGNITION RATE)

Model	No adaptation	Lexicon label	Hand-label
Native English	75.4%	80.6%	81.0%

TABLE II
COMPARISON OF ACOUSTIC MODELS (PHONEME RECOGNITION RATE)

model	baseline	speaker adapted
Native English	75.4%	80.6%
Japanese students' English		
base form	78.0%	81.8%
automatic label	77.1%	81.5%

acceptable to use the lexicon base form for speaker adaptation in the following experiments.

D. Comparison of Native and Non-native Acoustic Models

We have also explored the use of speech data spoken by Japanese students. We used the English corpus compiled from Japanese students' speech and funded by MEXT. [1] The corpus contains a total of 13129 sentences spoken by 178 Japanese speakers (85 male, 93 female). Although the corpus includes a large amount of pronunciation errors, it does not have faithful phonemic labels. Thus, we investigated two kinds of phonemic labels for acoustic model training: labels from base form and automatic labeling using the ASR with error prediction. The specification of the phoneme HMM is same as the previous Sub-section.

Table II lists the phoneme recognition results with the various acoustic models. In this evaluation, we also applied speaker adaptation. Thus, two kinds of results for each model were computed: baseline and speaker adapted. The effect of speaker adaptation is confirmed in this result, too. As we expected, the best performance is achieved by the acoustic model trained with the Japanese students' database. Here, the labels based on the base form are sufficient for training the acoustic model. This model yielded 2.6% better accuracy than the native English model without speaker adaptation (baseline). However, the superiority is decreased to 1.2% when speaker adaptation was applied. The results demonstrate that with speaker adaptation, the native English model can compete with the Japanese student's model .

E. Automatic Detection of Sentence Stress

In English, stressed syllables are characterized by not only power level, but also pitch, duration and vowel quality [9]. Based on the observations of typical error patterns made by Japanese students, we prepared the following classes for modeling stressed syllables.

- Stress level
  We classify the stress level into three categories. Primary-stressed syllables (PS) carry the major pitch change in a tonal group (phrase). There is only one PS in each phrase, usually placed on the word containing the most important piece of information. Secondary-stressed syllables (SS) are all other stressed syllables. Non-stressed syllables (NS) do not bear any mark of stress. Usually, all syllables but one in a word tend to be non-stressed in continuously spoken sentences.
- Syllable structure
  As the syllables of complex structures have a tendency of being stressed [10], we introduce classification of syllable structures based on four categories: V, CV, VC, CVC. We also classify vowels into four categories: schwa (Vx), short vowel (Vs), long vowel (Vl), and diphthong (Vd). Thus, combinations of these two factors give rise to 16 possible categories of syllables.
- Position in phrase
  Pitch in natural speech rises rapidly at the beginning of each phrase unit and falls gradually, resulting in strong influences on the sentence stress. Thus, we also classify syllables into three types according to their position in a phrase: head (H), middle (M) and tail (T).

We used the following acoustic features for detection of sentence stress: pitch (log(F0)), power (log(power)) and spectral (MFCC) parameters. These features can be regarded as independent, and are thus processed by three different streams in the model. The TIMIT database was used for training. Preliminary experiments showed that modeling the distribution with a mixture of eight Gaussians brought about the best result.

In order to reliably align the syllable sequence which includes the phoneme insertions and substitutions by non-native speakers, we apply the ASR with error prediction for a given sentence. Based on this alignment, the syllable units together with their structures and positions within a phrase are determined. According to the classification results, the corresponding PS, SS and NS models are applied to estimate the stress level. Syllables whose detected stress level differs from the correct level are marked as pronunciation errors. If the syllable structure and/or position in the phrase are incorrect, such information is presented to the student as possible causes of the stress error.

Since PS, SS and NS have different acoustic characteristics, the effective features for discrimination will differ. For example, PS is characterized by a tonal change, thus F0 should be the most important feature. We propose a two-stage recognition method. During the first stage, the presence of stress is detected. Here, a stress model (ST) that merges PS and SS is compared against NS using weights optimized for the two-class discrimination. For syllables detected as stressed, the stress level (PS or SS) is identified in the second stage using different weights. By tuning the weights with linear discriminant analysis, we achieved stress detection accuracy of 95.1% for native English speakers and 84.1% for Japanese students, which is a significant improvement from a naive combination using the same weight for all three features (93.7% for native English and 79.3% for Japanese students) [11].

[1]Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research on Priority Areas, No.12040106.

III. System Trials in Classrooms

The system was implemented with Java for Windows OS, and installed in a CALL classroom in ACCMS. In this CALL classroom, there are 48 PCs, each equipped with a headset microphone. We have been using this system in an English class for second-year students of Kyoto University.

A. Analysis of logged data

When we first used this system in the classroom, we encountered a number of unexpected problems. These problems are classified into the following categories.

- Errors in recording
  A number of errors in recording or voice activity detection were observed during the first trial of the system. We identified they were caused by improper configuration of recording levels. Thus, during the second trial of the system, we instructed students to set their recording levels prior to the practice, and the number of errors was reduced by 75%.
- Unpredicted pronunciation errors
  The system is designed to predict possible pronunciation errors for a given sentence based on the linguistic knowledge. However, students make a number of unexpected pronunciation errors. Most of them involve repetition of words and incorrect reading of phrases. For example, "sixteen-o-seven (1607)" and "sixteen three" for a phrase "1603 (sixteen-o-three)." These errors occurred because the students were not familiar with these words. There is essentially a limitation in predicting possible errors, and adding too many candidates would degrade the ASR performance. An alternative solution would be to simply add an explanation for the reading of the phrase in question and a function for re-recording.
- Speech recognition errors
  The system delivers a message indicating a recognition error when the utterance differs greatly from the correct model. While errors of this type were frequently observed during the first trial, there was no errors in the second trial after fixing the recording level.

B. Evaluation by the Students

We have received numerous positive opinions on this system. For example, a student remarked, "It is very interesting as I haven't experienced this kind of English practice. I want to practice more with this system." Another student commented, "Other classes don't offer the opportunity to use interesting systems like this one." On the other hand, some students complained improper configuration of the microphone settings.

We also counted the number of utterances and the number of errors students made using the logged data, and compared the results for the two trials. Table III lists the number of utterances per session and ratio of errors in recording and ASR. It is observed that in the second trial, the number of utterances was more than doubled and the number of errors was drastically reduced, which suggests that meaningful practices were conducted.

TABLE III
Analysis of logged data

	#Utterances	Error Rate (Recording)	Error Rate (Recognition)
1st trial	52.1	20.4	1.2
2nd trial	111	4.9	0

IV. Japanese CALL System: CALLJ

A. System Overview

The Japanese CALL system is organized to cover elementary grammar points and vocabulary from levels 4 and 3 of the Japanese Language Proficiency Test (JLPT[2]). These levels cover approximately 1500 words (of which around 200 are verbs), 300 kanji characters, and 95 grammar points. The grammar points are distributed across a set of 30 lessons. Each lesson consists of exercises and self-learning material, which help students master key grammar points and key sentence patterns. The exercises are a collection of related questions (=sentences) connected to some key sentence patterns, such as "like to do something". Before practicing, students look through the overview of the lesson points, notes of the grammar points, and examples of questions. Specifically, the overview briefly shows key sentence patterns and grammar forms. The notes give more information on the grammar structures that are used in the lesson. With these documents, students get an idea on sentence patterns in the current lesson before they exercise using the system.

A process flow of the exercises is depicted in Fig. 3. Each question involves the students being shown a "Concept Diagram", which is a picture representing a certain situation. The students are then asked to describe this situation with an appropriate Japanese sentence using text input or speech input. Thus, the system allows students the freedom to create their own sentences. If the answer is given via a microphone, ASR is conducted using a language model in the form of a grammar network for the target sentence. Errors are detected and feedback information is generated for the students. This process of question, answer and feedback is repeated.

Unlike the conventional textbooks or prepared materials, the system generates questions on the fly, by selecting subjects, objects and optional phrases with regard to time and place and so on. Accordingly, the diagram and the grammar network is generated by dynamically combining the relevant parts. Thus, students can try as many questions as they want.

Fig. 4 shows the user practice interface.

B. ASR Grammar Network Generation with Error Prediction

As the system has an idea of the desired target sentences, the system easily generates a language model to cover them in the form of a network. The major problem is to predict errors (possible answers different from target sentences) that non-native students tend to make, and to integrate them into the language model.

[2]http://en.wikipedia.org/wiki/JLPT

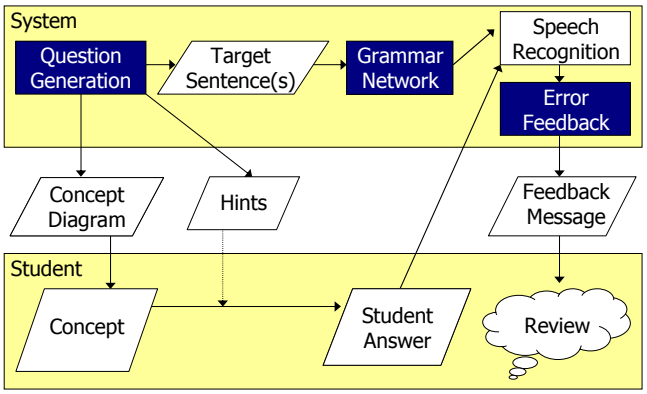

Fig. 3. Overview of CALLJ

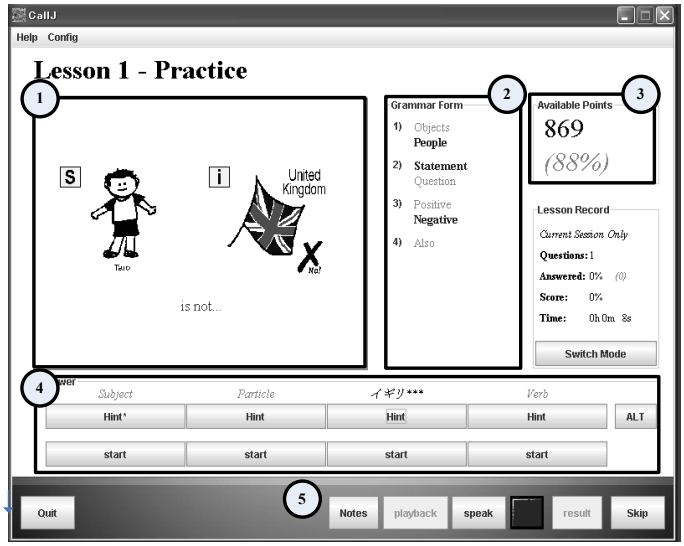

Fig. 4. Screen shot of CALLJ; 1: Concept diagram, 2: Desired form guide, 3: Score, 4: Answer area and hint display, 5: Control button panel

In the conventional CALL systems using ASR, the linguistic knowledge is widely used to achieve error prediction. In our English CALL system Hugo, too, pronunciation error patterns were hand-crafted to recognize Japanese students' English. However, the learner of the system was limited to Japanese students. Obviously, a larger number of error patterns will exist if the system allows any non-native speakers. Moreover, we need to handle more variations in the input, if we allow more freedom in the sentence generation, like CALLJ. These factors would drastically increase the perplexity of the grammar network, causing adverse effects on ASR. In order to find critical errors and avoid redundant patterns, a decision tree is introduced for error classification [12].

The error classification is conducted by comparing the features of the observed word to those of the target word. The features include same POS (Part-Of-Speech; verb, noun etc), same base form, similar concept, wrong inflection form, and so on. To select effective features and find critical error patterns, an "impact" criterion is introduced to find an optimal decision tree that balances the tradeoff of the error coverage

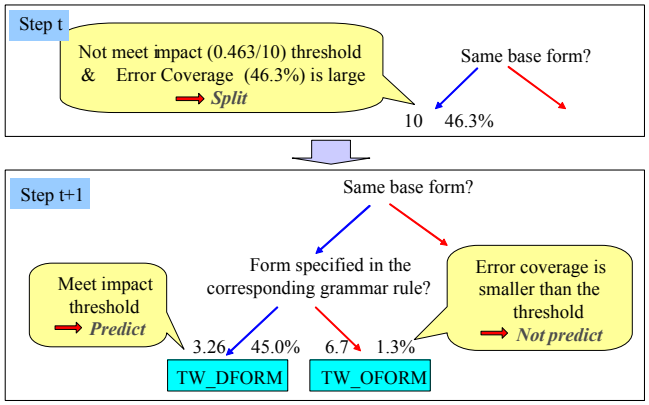

Fig. 5. Example of decision tree training process

and perplexity. It is used to expand a certain tree node from the root node (containing everything), and partition the data contained in the node according to some feature. For a given error pattern, it is defined as below:

$$impact = \frac{\text{error coverage}}{\text{perplexity}}$$

Error coverage is defined as the proportion of errors being predicted among all errors. It is measured by the frequency in the training data set, so that more frequent errors are given a higher priority. Perplexity is approximated by the average number of predicted competing candidates for every word in the training data set. The larger value of this impact, the better recognition performance can be achieved with this error prediction. Our goal is reduced to finding a set of error patterns that have large impacts. If a current node in the tree does not meet this criteria (threshold), we expand the node and partition the data iteratively until we find the effective subsets and mark "to predict", or the subset's coverage becomes too small and marked "not to predict". Fig. 5 shows an example of one step of the tree training for verbs. In each node, perplexity and error coverage of the node is labeled from left to right.

The training data for the decision tree learning were collected through the trials of the prototype CALLJ system with text input. They consist of 880 sentences, containing 653 errors. Since some errors can never happen or be tolerant in the speech input, we performed a pre-processing. Specifically, we corrected the input errors which are caused by typing or spelling mistakes and result in same pronunciation, such as "*o*" for "*wo*" (a particle) and "*tanaka san*" for "*tanakasan*".

After the training process, a decision tree is derived for each POS. As for verbs, eleven leaves are extended with a maximum depth of six in a binary tree. Among them, four leaf nodes are chosen for prediction as listed in Table IV.

Each error pattern falls within one of four error types: *Lexical*, *Grammatical*, *Concept*, and *Input*. Lexical errors are out-of-vocabulary words and inappropriate choice of words which are similar in concept. Features to identify similar-concept word pairs depend on the word component type. For verbs, they are: substitution between words that are grammar

TABLE IV
ERROR PATTERNS BEING PREDICTED FOR VERBS

Pattern	Type	Description
TW_DForm	grammatical	Target Word (base form) in Different Form
DW_SForm	lexical	Different Word in Same Form
DW_DForm	lexical	Different Word in Different Form
TW_WIF	grammatical	Target Word in Wrong Inflection Form

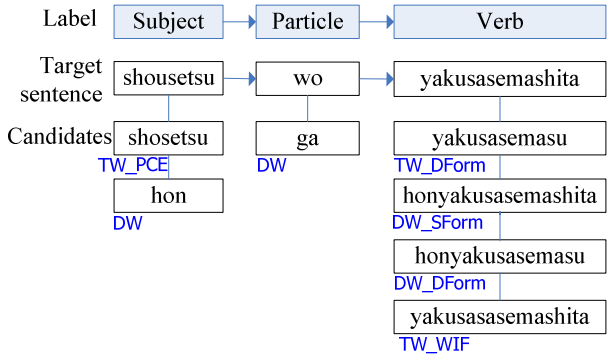

Fig. 6. Prediction result for a given sentence

points (such as "*ageru*", "*kureru*", and "*morau*"), between words having same meaning (such as "*honnyakusuru*" and "*yakusu*"), between the transitive and intransitive verb pair (such as "*okosu*" and "*okiru*"). Grammatical errors include wrong forms or wrong inflections of the correct word and inappropriate particles. Concept errors are mistakes not in the language itself, but in the interpretation of the situation that the students need to describe. Input errors are mistakes in the input format, such as *hiragana* being used instead of *katakana*.

As we identified the errors to predict, we can exploit this information to generate a finite-state grammar network. Given a target sentence, for each word in the surface form, we extract its features needed such as POS and the base form, and compare the features with error patterns to predict using the decision tree. Then, we generate potential error patterns with the prediction rules and add them to the grammar node. Fig. 6 shows an example of a recognition grammar based on the proposed method for a sentence "*shousetsu wo yakusasemashitaka*".

C. Experimental Evaluation

Twenty one foreign students of Kyoto University took part in the first trials using the text-input prototype system. The data collected in this trial were used for training of the decision tree. In the second trial, ten foreign students tested the system which incorporates speech-input capability. The data collected in this trial were used for evaluation of ASR. Ten students are from seven different countries including China, France, Germany and Korea. All students were studying Japanese in the Kyoto University Japanese language course, and thus their approximate language proficiency was known based on the course level in which they were enrolled in (Elementary, Intermediate 1 or Intermediate 2).

All students had no experience with the CALL system before the trial, but were briefly introduced before undertaking the task. Each student ran through a set of lessons, answering a set of generated questions before seeing the correct answers and feedback for errors they made. In the second trial, ASR based on a grammar network was executed at run time. After the trial, all utterances (140) were transcribed including errors by a Japanese teacher.

To evaluate the performance of ASR, we use the conventional WER (word error rate), error detection rate and false alarm rate. We define the error detection rate as the number of detected errors divided by the total number of errors the students made. The false alarm rate is the number of words erroneously flagged as a student error, divided by the total number of words students spoke correctly.

Comparing the system's output to the faithful transcript of utterances including errors made by the students, the WER of ASR is 11.2%. It is quite lower compared with the case (28.5%) using the baseline grammar, which is hand-crafted and does not consider errors made by foreign students. The baseline method simply includes all words in the same concept such as foods and drinks in the grammar network, and can be applied to any sentences in the same lesson. The error detection rate is 75.7% with the false alarm rate of 8.6%, though 85.7% of errors were covered by the grammar network and could be recognized in theory. The error coverage (85.7%) and perplexity measure (4.1) for the test data are comparable to those (77.9% and 5.1) for the training data. The result confirms the generality of the decision tree training.

V. CONCLUSIONS

This paper has given an overview of the English and Japanese CALL systems which have been developed at Kyoto University. Both systems incorporate ASR technologies to detect pronunciation errors. The English CALL system focuses on Japanese students, thus the acoustic modeling and language modeling are designed to reflect Japanese students' characteristics. While the acoustic model trained with the Japanese speakers provided better performance, we also showed that the native speakers' model can work comparably if speaker adaptation is allowed. The language model for error prediction was based on a set of rules which includes typical errors made by Japanese students. In the Japanese CALL system designed for any foreign speakers, we introduced an empirical error prediction method based on decision tree learning. The method successfully found critical error patterns without increasing the perplexity.

The English CALL system was installed in our classroom, to be used in the English courses or self-learning. Although it is not easy to measure the effect of the system on proficiency, the system provides a new learning environment in which the students enjoy practicing. The content of the Japanese CALL system is still under construction, and when it is complete, the system will be released to public.

ACKNOWLEDGMENT

The authors are grateful for former staffs and students engaged in this project, including Antoine Raux, Kazunori Imoto and Christopher Waple.

REFERENCES

[1] G. Kawai and K. Hirose. Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology. *Speech Communication*, 30:131–143, 2000.

[2] M. Eskenazi and S. Hansma. The Fluency Pronunciation Trainer. In *STiLL*, 1998.

[3] L. Neumeyer, H. Franco, V. Abrash, L. Julia, O. Ronen, H. Bratt, J. Bing, V. Digalakis, and M. Rypa. WebGrader: A Multilingual Pronunciation Practice Tool. In *STiLL*, 1998.

[4] H. Franco, V. Abrash, k. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari. The SRI EduSpeakTM System: Recognition and Pronunciation Scoring for Language Learning. In *STiLL*, 2000.

[5] A.Raux and T.Kawahara. Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *Proc. ICSLP*, pages 737–740, 2002.

[6] S. Kohmoto. Applied English Phonology: Teaching of English Pronunciation to the Native Japanese Speaker. *Tanaka Press*, 1965.

[7] Y.Tsubota, T.Kawahara, and M.Dantsuji. Recognition and verification of English by Japanese students for computer-assisted language learning system. In *Proc. ICSLP*, pages 1205–1208, 2002.

[8] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

[9] M. Sugito. English Spoken by Japanese. *Izumi Shoin*, 1996.

[10] R.M. Dauer. Stress-Timing and Syllable-Timing Reanalyzed. *Journal of Phonetics*, 11:51-62. 1983.

[11] K.Imoto, Y.Tsubota, A.Raux, T.Kawahara, and M.Dantsuji. Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system. In *Proc. ICSLP*, pages 749–752, 2002.

[12] H.Wang and T.Kawahara. Effective error prediction using decision tree for ASR grammar network in CALL system. In *Proc. IEEE-ICASSP*, pages 5069–5072, 2008.