ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A Method of Generating English Pronunciation Dictionary for Japanese English Recognition Systems

Tadashi Suzuki, Jun Ishii, and Kunio Nakajima

Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, Japan
tadashis@isl.melco.co.jp

Abstract

In this paper, we propose a method for generating a pronunciation dictionary—extracting typical pronunciations for each word from speech data uttered by Japanese speakers—as one approach to speech recognition targeting English speech uttered by Japanese speakers whose mother tongue is not English. This method includes three processes: a process in which English phoneme HMMs (Hidden Markov Models) are adapted to the speaker using English speech uttered by a Japanese speaker; a process in which English by a Japanese speaker is translated into an English phoneme series using a phoneme typewriter; and a process by which representative phoneme series are selected with a clustering technique from multiple phoneme series derived with respect to each word. We also propose a speaker adaptation method in a recognition phase. In this method, the phoneme HMMs are adapted to the target speaker with a phoneme label series that expresses the typical pronunciation extracted using the above method. Evaluation tests by continuous speech recognition with English speech data uttered by five Japanese speakers using a pronunciation dictionary generated from other five Japanese speakers' data were carried out. The result of the tests indicated that sentence recognition errors were reduced by 72% compared to using a dictionary for native speakers.

1. Introduction

Recently, there has been an increasing demand for speech recognition systems targeting non-native speakers. There is a problem, however, in that recognition systems designed and trained for native speakers do not offer sufficient performance in the case of Japanese speakers for whom English is not the mother tongue, because of factors such as unique accent and instability in pronunciation.

As a method for speech recognition systems targeting non-native speakers of English, there is a proposal for a method in which a recognition network is constructed based on rewriting rules from a text to a phoneme network for Japanese English [1]. In this method, the recognition accuracy is highly dependent on the rewriting rules, and it is difficult to create general rules without a loss of performance. Other methods have been reported as well; for example, using English speech data uttered by Japanese speakers, initial HMMs are adapted to the Japanese speakers based on MAP estimation [2]. There is, however, a problem in which unstable pronunciations of Japanese English data used for speaker adaptation degrades the adaptation accuracy.

In this paper, we will propose two methods for continuous English speech recognition that take into account the characteristics of English as pronounced by Japanese speakers. In the first method, typical Japanese English pronunciations are extracted for each word from speech data uttered by Japanese speakers, and used as a recognition dictionary. This method includes three processes: a process in which English phoneme HMMs are adapted to the speaker using English speech uttered by a Japanese speaker; a process in which English utterance by a Japanese speaker is translated into an English phoneme label series using a phoneme typewriter; and a process by which representative phoneme series are selecting using a clustering technique from multiple series of phoneme label derived with respect to each word. In the second method, for a recognition phase, the phoneme HMMs are adapted to the target speaker, by using the phoneme label series that expresses the typical pronunciation extracted utilizing the first method. The aim is to cope with the adaptation accuracy degradation that occurs as a result of using adaptation training speech data that includes obscure pronunciations by Japanese speakers.

2. Recognition system for Japanese English

2.1 Construction of a Japanese English word pronunciation dictionary

In this section, we discuss the construction of a word pronunciation dictionary that expresses pronunciations representative of English as uttered by Japanese speakers, and methods of applying this dictionary in speech recognition. Fig. 1 shows the process flow for this method.

This method is executed using three elements: a process that extracts partial data corresponding to the each word derived from sentence data uttered by Japanese speakers; an English phoneme typewriter that translates the extracted word data into English phoneme label series; and a clustering process to derive the label series that expresses representative pronunciations in Japanese English from multiple phoneme label series belonging to the same word. In order to minimize the influence of inter-speaker variation on the phoneme typewriter and

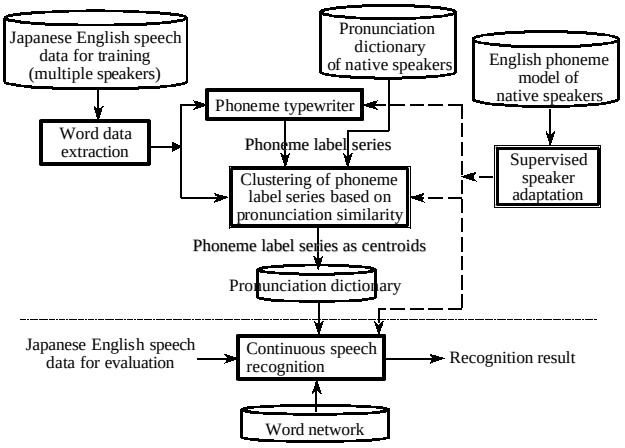

Fig. 1: Procedure of construction of a Japanese English pronunciation dictionary

clustering processes, the English phoneme models are adapted to a speaker using Japanese English speech data uttered by the Japanese speaker. The details of the various processes are described below.

Extraction of word data

Multiple word data belonging to the same word are extracted from English sentence data that has been uttered by Japanese speakers. Segmentation using Viterbi algorithm is carried out with the word model sequences according to the word series of the sentences. In the model sequence, the insertion of pause model between word models is accepted to cope with isolation of word utterances in sentence speech.

Speaker adaptation of English phoneme models

The English phoneme HMMs are adapted to the speaker so as to reduce inter-speaker variation on the phoneme typewriter and phoneme label series clustering processes, which will be discussed later. The Maximum Likelihood Linear Regression (MLLR) method is used for the method of speaker adaptation of English phoneme models. This method transforms the initial phoneme models' spectra into spectra that characterize the target speaker's speech, while maintaining continuity between the phoneme models' spectra. Because adaptation is possible when a rough correspondence is obtained between two sets of the spectra, of the speaker specific training data and of initial (before adaptation) phoneme models, this method has the advantage of being applicable even with a comparatively small amount of speech data for training use. It can thus be assumed that even when using training data with obscure pronunciations, as in the case of English uttered by a Japanese speaker, the system will not easily be influenced by these imperfections.

As shown in the equation below, the MLLR method is carried out for speaker adaptation by transforming the mean vector of Gaussian distribution in each phoneme model m_k into $\hat{m}_k$ via the multi-regression mapping model.

$$\hat{m}_k = A \cdot m_k + b$$

In this equation, A is the DxD transformation matrix, and b is the D dimension vector (D is the number representing the dimension of the mean vector). The transformation matrix A and the vector b is estimated with the adaptation training data based on the maximization criteria for the Baum auxiliary function. To obtain good performance with small amount of training data, the Gaussian distributions of the phoneme models are arranged in a tree structure beforehand using state clustering based on the Bhattacharyya distance, and the Gaussian distribution which shares both the transformation matrix A and vector b is determined based on amount of the training data.

Phoneme typewriter

The output of the English phoneme typewriter (the English phoneme label series) is derived in relation to the extracted speech data for each word. The English phoneme typewriter uses the English phoneme models adapted to the speaker for the word data, and performs matching with the phoneme network that allows all connections between English phoneme models, including pauses (continuation of the same model, however, is unacceptable). For a input word data, the English phoneme series that gives the maximum likelihood for the matching is then output. In this way, the word speech pronunciation in Japanese English is expressed as an English phoneme label series.

Phoneme label series clustering based on similarity in pronunciation

Clustering is carried out using as a criteria of similarity in pronunciation of the M series of English phoneme label obtained from a given word category, and label series expressing representative pronunciations are derived. The pronunciation similarity between two label series is defined by the likelihood of matching the word data corresponding to one label series to the word model obtained from another label series. The word data has been used to make the former label series in the phoneme typewriter process. And the word model is made from the latter label series, connecting speaker adapted HMMs according to the label series.

The average similarity L_{avg} when N label series are taken as the centroids is derived using the equation below. Obtaining the set of N label series maximizes L_{avg}, clustering is conducted for N clusters.

$$L_{avg} = \frac{1}{M} \cdot \sum_{m=1\ldots M} \max_{n=1\ldots N} \left\{ L\left(T_m, \hat{P}_n\right) \right\}$$

In this equation, $L(T_m, \hat{P}_n)$ is the likelihood of matching between the word data T_m and the word model $\hat{P}_n$ created by connecting HMMs adapted to the speaker uttered the word data T_m.

Carrying out this clustering process for each word category, we are able to construct a word pronunciation dictionary that has N units of label series as representative pronunciations of Japanese English for each words. Based on the results of preliminary investigations, a restriction was introduced such that a label series corresponding to the native speaker pronunciation is always taken as one of the centroids. This restriction was introduced to confirm that the recognition rate is improved by combining the dictionary of normal pronunciation by native speakers with the Japanese English word pronunciation dictionary. There are thus always at least two clusters.

2.2 Speaker adaptation using the Japanese English pronunciation dictionary

In supervised speaker adaptation that uses the phoneme label series of native speaker's pronunciation as an adaptation training label sequence, it can be assumed that a sufficient adaptation accuracy cannot be attained when Japanese English speech data that is characterized by unstable pronunciation is used for adaptation data. For example, in the case of the English word "seven (/s/e/v/ə/n/)," a Japanese English word data might be expressed with a phoneme label series such as /s/i/b/u/m/ŋ/, due to the unstable pronunciation of the Japanese speaker's English. When using this data for adaptation, all the phonemes except for the /s/ would be assigned to speech data differing from training phoneme labels as the adaptation supervisor, and the result would be incorrect adaptation.

We have thus proposed a method by which MLLR speaker adaptation is carried out using a phoneme label series—which has the average utterance characteristics found in Japanese English—as the adaptation supervisor in place of the phoneme label series of native speaker pronunciation. That is to say, in

reference to each word of speaker adaptation data, we have generated a phoneme label series using the method described earlier for generating the pronunciation dictionary with different restriction in which without including a native speaker label series one unit of label series are selected as a centroid. MLLR speaker adaptation is then carried out using the English phoneme label series from this pronunciation dictionary. In this way, in the case of the previous example using the word "seven," speaker adaptation is carried out based on assignment of correspondence between the Japanese English speech /s/i/b/u/m/ŋ/ and the adaptation training label series that reflects the average utterance characteristics of Japanese English (e.g., /s/e/b/u/ŋ/). As a result, it can be assumed that the loss of adaptation accuracy due to incorrect assignment of correspondence between the supervisor's label and the speech data can be kept to a minimum.

3. Recognition experiments

3.1 Experimental conditions

The method described above is evaluated using continuous English speech recognition tests. The test conditions are shown in Table 1.

The phoneme model is a 4-state, 3-loop (left-to-right) continuous mixture density HMM (8 mixtures) that has been trained using native speaker speech data. The word pronunciation dictionary was generated using phoneme label series clustering with English business sentence (air traffic control) speech data from five speakers (TH, TI, TS, YA, and YI). The evaluation tests based on continuous English speech recognition were conducted using the same English business sentence data, spoken by a separate set of speakers (5 speakers: AN, JI, KW, MY, and TE). Three conditions were set for the word pronunciation dictionary: native speakers only (when one cluster was used for clustering), 2 clusters, and 4 clusters (in the latter two cases, there were 1 and 3 Japanese English word models, respectively). Two sentences of English speech according to TIMIT database, uttered by each speaker, were used for speaker adaptation. The Japanese English phoneme label series to be used for phoneme HMM adaptation were generated with two sets of speech data—separate from the speech data set used for speaker adaptation in a recognition phase—spoken by ten speakers of Japanese English.

Table 1: Experimental conditions

Speech Data	
Native Speaker	TIMIT 2310 sentences (5 sentences × 462 speakers)
Japanese Speaker	2 sentence × 10 speakers (for speaker adaptation) Business sentence : 274 sentences × 2 sets (by 10 speakers, for evaluation)
AD-trans	16bits * 10kHz
Acoustic Analysis	
Analysis window	Hamming (Length 25.6msec / period 10.0msec)
Pre-emphasis	$1 - 1.0 * z^{-1}$
Order of LPC analysis	15
Feature vector	Mel-Cepstrum (1 - 10) + ΔMel-Cepstrum (0 - 10)
Phonetic label	/b/d/g/p/t/k/jh/ch/s/sh/z/zh/f/th/v/dh/m/n/ng/em/en/l/r/w/y/hh/el/iy/ih/eh/ey/ae/aa/aw/ay/ah/ao/oy/ow/uh/uw/er/ax/ix/axr/h#/
Perplexity	5.0756
Acoustic phonetic model	8-mixed Gaussian HMM × 46 phonemes 4-state 3-loop (Left-to-right/tied-arch)

3.2 Recognition results using Japanese English pronunciation dictionary

First, we conducted evaluation tests for cases using the Japanese English word pronunciation dictionary. Conventional speaker adaptation was carried out with native speaker pronunciation as training label series. The results are shown in Fig. 2. In this figure, the speakers are on the x-axis with the average shown on the far right, and the y-axis is the sentence recognition error rate. In the results for each speaker, the two bars on the right are recognition results based on a word pronunciation dictionary created using the proposed method. The center right bar represents cases using the 2-cluster dictionary (2-cls in the legend), and the bar on the far right represents cases using the 4-cluster dictionary (4-cls). The two bars on the left show the results of recognition using only one cluster for model clustering; that is, using only the native speaker word dictionary. This recognition was carried out for the purposes of comparison. The bar on the far left represents cases using an English phoneme model without speaker adaptation (Native (no-adaptation) in the legend), and the center left bar represents cases using the English phoneme model after speaker adaptation (Native (speaker adaptation)).

By using the Japanese English word pronunciation dictionary, we have reduced the number of recognition errors as compared to cases using only the native speaker pronunciation dictionary. In terms of the average for all five speakers, the error rate of 18.4% in the case of speaker-adapted HMM and native speaker dictionary (Native (speaker adaptation)) has been reduced to 6.2% in the case of 2-cls and 5.1% in the case of 4-cls, thus demonstrating the effects of the dictionary that expresses the typical word pronunciation of Japanese speakers.

3.3 Result of phoneme model speaker adaptation using Japanese English pronunciation dictionary

We conducted recognition tests based on a phoneme model that had been speaker-adapted using a label series which expresses the average pronunciation of Japanese English. The label series are shown below.

SA1: She had your dark suit in greasy wash water all year.
- Native label sequence
  /sh iy /hh ae d /y uh axr /d aa r k /s uw t /ih n /g r iy s iy /w ao sh /w ao t axr /ao l /y ih r
- Japanese label sequence
  /sh iy /hh ah b /y uh /d ah k /s uw t /iy ng /g l r iy z iy /w ax sh /w ax p ah /ow el /iy y eh axr

SA2: Don't ask me to carry an oily rag like that.
- Native label sequence
  /d ow n t /ae s k /m iy /t uw /k ae r iy /ae n /oy l iy /r ae g /l ay k /dh ae t
- Japanese label sequence
  /d ao n t /ah s k /m iy /t uw /k y ae r iy /uh ah n /l oy r iy /r ay g /n ay k /dh ae t

The results are shown in Fig. 3. The two axes in the graph are the same as in Fig. 2. Looking at the bars for each speaker, the three bars on the left represent the results for cases in which the native speaker pronunciation dictionary was used (i.e., the Japanese English pronunciation dictionary was not used). Reading from the left, the bars represent without speaker adaptation, adaptation using conventional methods

(with native pronunciation label series as a training label sequence), and adaptation using the proposed method (with label series of a typical pronunciation in Japanese English as training label sequence).

In the results of the "average" for all five speakers, we see that the error rate for "no adaptation" was 24.6%, traditional adaptation 18.4%, and adaptation with Japanese English pronunciation dictionary as training label sequence 14.9%, which allows us to confirm the effect of the proposed method. In the results for each speaker individually, we find that while there was one speaker (JI) for whom recognition accuracy improved substantially compared to the conventional method, there were others (KW, MY, and TE) for whom no difference appeared in the recognition rate. We can thus see that there is a considerable speaker dependency.

The four bars on the right for each speaker represent the results for cases in which the conventional and proposed speaker adaptation method is applied to recognition using the Japanese English word pronunciation dictionary. From the left, they are: 2-cls with conventional speaker-adapted phoneme model; 2-cls with adapted phoneme model using the proposed method; 4-cls with the conventional adapted phoneme model; and 4-cls with the proposed method. In the cases of speakers AN, JI, and KW, we can confirm the effects of the proposed model, but there was a opposite effect in the case of MY, indicating that speaker dependency was a factor here as well.

We can assume that the reason for the speaker dependency is that adaptation accuracy deteriorated in the case of speakers with utterance characteristics that did not conform to the average utterance characteristics because only one phoneme label series for average utterance characteristics was used as a supervisor. A solution for this problem is to apply a method using multiple series of phoneme label with typical utterance characteristics as a fuzzy supervisor to speaker adaptation.

4. Conclusion

As a method of continuous speech recognition targeting Japanese English, we have proposed a method by which a word pronunciation dictionary expressing pronunciation typical of Japanese English is generated using an English phoneme typewriter and clustering of phoneme label series. We also discussed a method for the process of speaker adaptation of English phoneme models, in which an English phoneme series expressing average Japanese English pronunciation is used as training label sequence. The results of continuous speech recognition evaluation tests using five speakers indicated that a sentence recognition rate of 94.9% had been achieved, thus confirming the effectiveness of the proposed method. In the future, we will continue to investigate speaker adaptation methods that take into consideration the variations in pronunciation of English uttered by Japanese speakers.

References

[1] Hideyuki Suzuki, Seiichi Nakagawa, "Speech Recognition using phoneme models adapted for second language learner," Proc. of the Meeting of the Acoustic Society of Japan, pp. 47-48 (1995-3) (in Japanese)

[2] Meron Yoram, Keikichi Hirose, "Language Training system using speech processing techniques," Technical Report of IEICE, SP96-18, (1996-06)

[3] C.J.Leggetter, P.C.Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," Computer Speech and Language, Vol. 9, pp. 171-185 (1995)

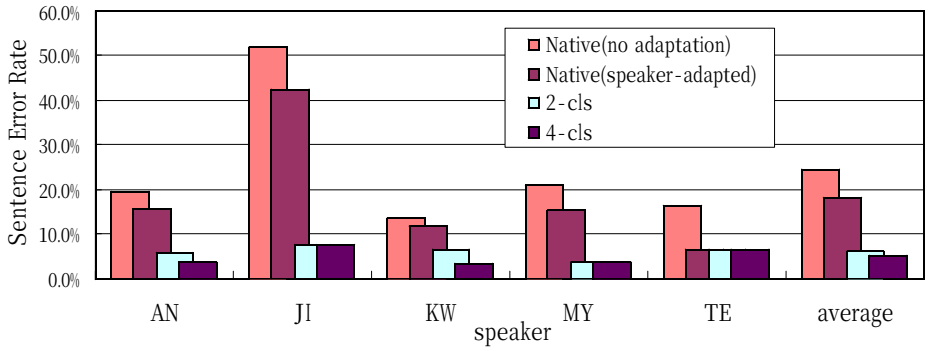

Fig. 2: Effect of using a pronunciation dictionary for Japanese English

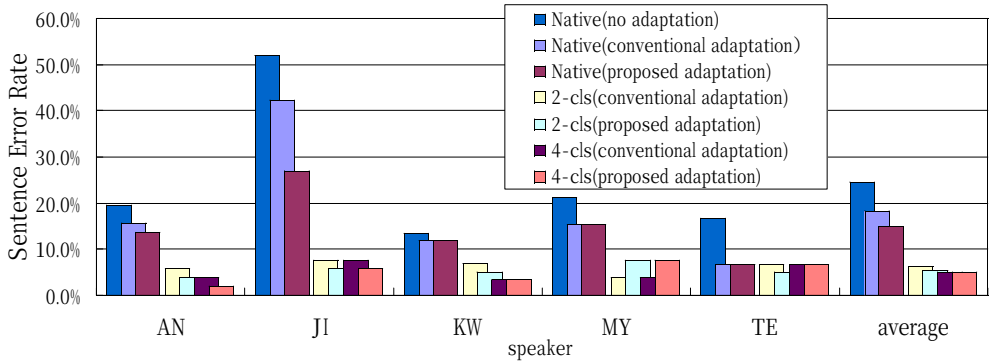

Fig. 3: Effects of speaker adaptation using Japanese English label series