Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models

Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno, *Fellow, IEEE*

***Abstract*—This paper presents an automatic speech recognition (ASR) system that accepts a mixture of various kinds of dialects. The system recognizes dialect utterances on the basis of the statistical simulation of vocabulary transformation and combinations of several dialect models. Previous dialect ASR systems were based on handcrafted dictionaries for several dialects, which involved costly processes. The proposed system statistically trains transformation rules between a common language and dialects, and simulates a dialect corpus for ASR on the basis of a machine translation technique. The rules are trained with small sets of parallel corpora to make up for the lack of linguistic resources on dialects. The proposed system also accepts mixed dialect utterances that contain a variety of vocabularies. In fact, spoken language is not a single dialect but a mixed dialect that is affected by the circumstances of speakers' backgrounds (e.g., native dialects of their parents or where they live). We addressed two methods to combine several dialects appropriately for each speaker. The first was recognition with language models of mixed dialects with automatically estimated weights that maximized the recognition likelihood. This method performed the best, but calculation was very expensive because it conducted grid searches of combinations of dialect mixing proportions that maximized the recognition likelihood. The second was integration of results of recognition from each single dialect language model. The improvements with this model were slightly smaller than those with the first method. Its calculation cost was, however, inexpensive and it worked in real-time on general workstations. Both methods achieved higher recognition accuracies for all speakers than those with the single dialect models and the common language model, and we could choose a suitable model for use in ASR that took into consideration the computational costs and recognition accuracies.**

***Index Terms*—Corpus simulation, mixture of dialects, speech recognition.**

I. Introduction

THE performance of automatic speech recognition (ASR) in real-time has drastically improved recently and has been widely applied to various systems. These systems include automatic transcriptions for the National Congress [1], real-time television subtitling [2], and smartphone applications with speech interfaces [3], [4]. One of the major problems in these systems has been to ignore the dialects of speakers to maintain a practical recognition accuracy for unspecified speakers. For example, Japanese ASR systems recognize utterances of speakers with the common language (Tokyo dialect) but they have difficulty in recognizing utterances by speakers from other areas. Commonly-spoken language is a mixture of dialects because they are not only affected by the speakers themselves but also by their own and their parents' business and residence histories. A more complicated problem is that the way of mixing dialects fluctuates even within a speaker [5]. This makes it much more difficult to recognize dialect utterances accurately. This paper solves the problem by introducing a mixture of dialects. It enables us to estimate how dialects are mixed as well as improve ASR accuracy. This paper focuses on differences in vocabulary in various kinds of characteristics of dialects, which are identified as especially important characteristics [6].

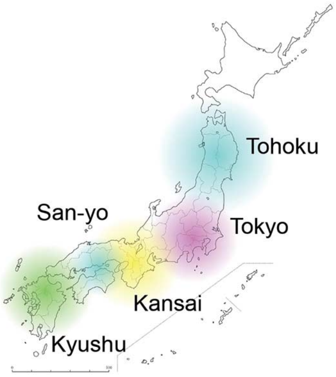


Fig. 1. Dialect mixtures.

This paper considers actually-spoken dialects as a mixture of dialects from each area, and we develop an ASR system that estimates the proportions of mixed dialects in input utterances. Each dialect has different acoustic and linguistic characteristics, which cannot be divided by using specific boundaries [7]. That is why actually-spoken language should be a mixture of each dialect (Fig. 1), where the proportion differs slightly even in small areas.

Dialect characteristics are divided into three types: 1) *pronunciation*, 2) *vocabulary*, and 3) *word order*. One example of *pronunciation* in the English language is in the three words: *"marry"*, *"merry"*, and *"Mary"* [8], [9]. The pronunciations of these three words are the same in one area and different in another. For example, all three are kept distinct in Northeast

Manuscript received April 29, 2014; revised October 12, 2014; accepted December 10, 2014. Date of publication January 05, 2015; date of current version January 16, 2015. This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (S) No. 24220006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

The authors are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: hirayama@kuis.kyoto-u.ac.jp; yoshino@ar.media.kyoto-u.ac.jp; itoyama@kuis.kyoto-u.ac.jp; forest@i.kyoto-u.ac.jp; okuno@i.kyoto-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2387414

America as [mæri], [mɛri], and [mɛəri], while they are merged in most regions in Canada and the Pacific Northwest of America. One example of *vocabulary* is in expressions such as *"watch your step"* and *"mind the gap"* [10]. The former is mainly used in American English while the latter is mainly used in British English within the context of the London Underground. One example of *word order* is *"Tuesday next"* in archaic Canadian English instead of *"next Tuesday"* in contemporary American English [11]. This paper focuses on differences in *pronunciation* and *vocabulary* among dialects. We assume that the set of phonemes and acoustic features of each phoneme would be the same, and pronunciation differences are modeled as differences in vocabulary. For example, if *Mary* is pronounced in the same way as *merry*, we provide the same phoneme sequence of pronunciation as *merry* to *Mary*, instead of training acoustic models for each vowel and consonant.

To develop an ASR system that focuses on vocabulary differences, we have to solve three problems:

Problem 1 Lack of linguistic corpora for dialect language models.

Problem 2 Estimates of the proportions of mixed dialects in utterances.

Problem 3 Trade-off between recognition accuracy and computational costs.

Problem 1: The *lack of linguistic corpora for dialect language models* stems from difficulties in collecting sentences with dialects from newspaper articles or documents on the Web. We develop a method of simulating dialect corpora with a large common language corpus and a parallel corpus between the common language and a dialect [12].

Problem 2: The *estimates of the proportions of mixed dialects in utterances* improves recognition accuracies by constructing optimal language models of mixed dialects suitable for input utterances. Our system recognizes utterances with various mixing proportions in language models and chooses the result that has the maximum recognition likelihood.

Problem 3: The *trade-off between recognition accuracy and computational costs* means that finding optimal language models of mixed dialects incurs too much computational cost to be run in real-time. We investigate a method of integrating recognition results obtained from ASR with each dialect language model to reduce the computational cost. This paper mainly describes solutions to Problems 2 and 3.

Our ASR system is designed to output recognition results as pairs of common language and dialect words so that the layers above including natural language understanding (NLU) can be applied with no modifications. The input and output of our system are below. The experimental evaluation explained in Section IV adopted words in the common language as reference data, because ASR results should accompany corresponding words in the common language for applications using ASR results. Thus, the outputs of our system are capable of being inputs of existing methods of NLU. It compares the words in the common language with reference sentences in the common language to place emphasis on utilizing ASR results for NLU in spoken dialogue systems.

This paper is organized as follows. Section II summarizes studies on previous dialect ASR systems and dialect estimates. Section III describes two methods of dialect ASR: maximized recognition likelihoods and integrated recognition results. Section IV evaluates the effectiveness of these two methods and discusses their pros and cons. Finally, Section V concludes this paper and describes their potential applications.

II. Related Work

Various methods have been developed for constructing models of ASR for dialects and estimating dialects of input utterances. Most of them have not been practical or able to be applied to general purposes. The following sections describe examples of the previous methods that correspond to Problems 1 and 2 in the previous section.

A. Models for Dialect ASR

We should develop models for ASR that correspond to individual dialects to introduce dialects to ASR systems. Research on acoustic modeling for dialects [13] investigated what kinds of characteristics are critical for discriminating spoken dialects. This had an effect on dialects whose acoustic features are dominant, while linguistic features, i.e., differences in vocabulary were not focused on. Vocabulary is especially important in dialect features [6], and methods of introducing linguistic features to dialect ASR systems are also required. Language models for dialect ASR systems that this paper focuses on have also been studied. Lyu *et al.* [14] developed an ASR system for dialects in the Chinese language with a pronunciation dictionary composed of pairs of words in the common language and the corresponding dialect pronunciation. This ASR system output words in the common language and dialect pronunciation for input dialect utterances and was based on the same problem setting as that in this paper. The problem was that the pronunciation dictionary was handcrafted, i.e., additional dictionaries were required to be handcrafted when the number of dialects or the size of vocabulary increased.

We solve this problem by simulating dialect linguistic corpora with parallel corpora between the common language and a dialect [12]. The simulated corpora result in language models of specific dialects, not mixed dialects. The solution to Problem 2 explained in the next subsection enhances the language models to address mixed dialects.

B. Dialect Discrimination for Unknown and Mixed Dialects

One strategy for dealing with unknown dialect utterances is to switch models suitable for the dialect of input utterances, which depend on dialect discrimination results. Dialect discrimination based on acoustic and linguistic features have been studied in some languages. Methods with acoustic features train models with a large amount of speech data for each dialect and estimate the unknown dialects of other speech data. Ching *et al.* [15] developed a discrimination method for two kinds of Chinese dialects (Mandarin and Cantonese) with acoustic features such as power, pitch, and speech rate. Miller *et al.* [16] also studied discrimination between English in the north and south of the U.S.A., and Chitturi *et al.* [17] studied discrimination between English in the U.S.A., U.K., and Australia with acoustic features. Methods with linguistic features estimate the dialect of written sentences on the basis of language models trained with large dialect linguistic corpora. Elfardy *et al.* [18] developed a method of two-class discrimination between common Arabic and its Egyptian dialect. The discrimination features included the rate of out-of-vocabulary words, the average length of words in a sentence, and perplexity of a sentence on the language models that corresponded to the two dialects. This method required more than thousands of dialect sentences.

These methods of discrimination work well for relatively major dialects, but it is costly to prepare new data to adopt them for other minor dialects. It is much more difficult to collect a large amount of speech data and written sentences for dialects that have fewer speakers than those for dialects that have numerous speakers. Furthermore, these methods of discrimination are deterministic; just one dialect is chosen for input utterances even if they consist of multiple dialects. We allow input dialects to be a mixture of dialects and conduct ASR with multiple dialect language models to improve recognition accuracies and estimate the proportions of mixed dialects.

We developed an ASR system for a mixture of the common language and a single dialect of the Japanese language [5]. The model composed of the common language and a single dialect was apparently insufficient to cover the whole language. Therefore, this paper deals with multiple dialects (five in the experimental evaluation). We confirmed that this method also improved recognition accuracies for a mixture of these dialects. The resulting model could automatically estimate the proportions of mixed dialects in input utterances.

III. Dialect ASR with Mixed Dialect Language Models

This section describes a dialect ASR system that accepts mixed dialect utterances. An ASR system for a single dialect is already available with a dialect language model based on simulated dialect corpora [12]. This method requires a method of mixing or integrating multiple dialect language models to recognize multiple dialects. We introduce two kinds of methods to deal with multiple dialect language models: maximized recognition likelihoods and integrated recognition results.

A. *Maximized recognition likelihoods*

This method interpolates dialect language models to construct mixed dialect language models with each combination of mixing proportions. We assume that actually-spoken dialect is a weighted mixture of several dialects since how individual dialects are influenced differs even among people in the same area because of differences in their residential history (Section I). Our main idea is that interpolated dialect language models, called mixed dialect language models after this, are capable of recognizing any dialects because the models cover the vocabulary of all dialects. The main question is how the optimal mixing proportions of each dialect were determined.

This method recognizes an input utterance with a mixed dialect language model of each combination of proportions, and outputs the result that has the maximum recognition likelihood in the recognition results of all combinations. Since the number of continuous mixing proportions is infinite, this method treats a combination of mixing proportions as being discrete, where each proportion is in units of a certain value. Fig. 2 outlines an example of combinations of mixing proportions in units of 20% with three dialects. When mixing K dialects, $D = \{d_1, d_2, \ldots, d_K\}$, with proportions in units of $100/M$ [%] ($0 < M \leq 100$), the set of combinations of mixing proportions, P, is defined as:

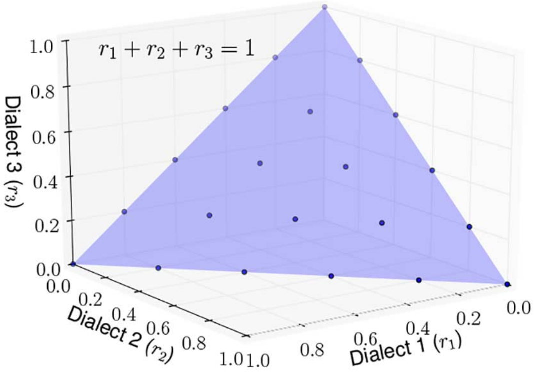


Fig. 2. Combinations of mixing proportions in units of 20% with three dialects.

$$P = \left\{ (r_1, \cdots, r_K) \;\middle|\; \begin{array}{l} \sum_{k=1}^{K} r_k = 1, \\ r_k \in \left\{ \frac{m}{M} \middle| m = 0, 1, \cdots, M \right\} \\ \quad (k = 1, 2, \cdots, K) \end{array} \right\}, \quad (1)$$

where $r_1, \ldots, r_K$ are the mixing proportions of each dialect. We construct mixed dialect language models for each of the combinations, whose number is

$$|P| = \binom{M + K - 1}{M} = \frac{(M + K - 1)!}{M!(K - 1)!}. \quad (2)$$

This result is derived from the fact that the number of combinations of m non-negative integers whose sum equals n is

$$\binom{n + m - 1}{n} = \frac{(n + m - 1)!}{n!(m - 1)!}. \quad (3)$$

In the example in Fig. 2, where $K = 3$ and $M = 5$, the number of mixed dialect language models is

$$|P| = \frac{(5 + 3 - 1)!}{5!(3 - 1)!} = \frac{7!}{5!2!} = 21. \quad (4)$$

The experimental evaluation in Section IV set $K = M = 5$, where $|P| = 126$.

The proposed method chooses the mixing proportions that maximize the recognition likelihood. This poses a question; does the recognition accuracy improve when the recognition likelihood increases? To find out, we calculate the correlation coefficients of pairs of recognition likelihoods and accuracies. If they are positively correlated, maximized recognition likelihoods are proved valid for the method of improving recognition accuracies.

The following describes how we construct the mixed dialect language models with a given set of mixing proportions.

The word n-gram probabilities of the mixed dialect language model, $P_{\mathrm{mix}}(w_i|w_{i-n+1}^{i-1})$, are constructed by linearly interpolating those of each single dialect d, $P_d(w_i|w_{i-n+1}^{i-1})$. The values of $P_{\mathrm{mix}}(w_i|w_{i-n+1}^{i-1})$ are calculated as:

$$P_{\mathrm{mix}}(w_i|w_{i-n+1}^{i-1}) = \sum_d r_d P_d(w_i|w_{i-n+1}^{i-1}), \quad (5)$$
$$\text{s.t.} \quad \sum_d r_d = 1, r_d \geq 0,$$

where r_d is the given mixing proportion of dialect d. Actual interpolation includes back-off probabilities as well as Eq. (5) [19] D We adopt the Kneser-Ney method [20] for back-off smoothing.

The pronunciation probabilities for each word in the common language had to also be interpolated. They are originally output through dialect corpus simulation [12]. The interpolated pronunciation probabilities are determined by interpolating the number of times each pronunciation appears in the simulated dialect corpora of each dialect. When word w in the common language appears $\#(w)$ times in a dialect corpus, and dialect pronunciation y corresponding to w appears $\#(y|w)$ times there, the pronunciation probability of y for w, $P_{\text{class}}(y|w)$, is calculated as

$$P_{\text{class}}(y|w) = \frac{\#(y|w)}{\#(w)} = \frac{\#(y|w)}{\sum\limits_{y} \#(y|w)}. \tag{6}$$

The $\#(w)$ and $\#(y|w)$ in the interpolation of language models are interpolations of $\#_d(w), \#_d(y|w)$, which are those of each single dialect d. Thus, Eq. (6) is transformed as

$$P_{\text{class,mix}}(y|w) = \frac{\#(y|w)}{\#(w)} = \frac{\sum\limits_{d} r_d \#_d(y|w)}{\sum\limits_{d} r_d \#_d(w)}. \tag{7}$$

When all dialect corpora for training language models are simulated from the same common language corpus, $\#_d(w)$ does not depend on d since the number of times common language word w appears remains unchanged before and after the dialect corpora are simulated. Equation (7) is rewritten as

$$\begin{aligned} P_{\text{class,mix}}(y|w) &= \frac{\sum\limits_{d} r_d \#_d(y|w)}{\sum\limits_{d} r_d \#(w)} = \frac{\sum\limits_{d} r_d \#_d(y|w)}{\#(w)} \\ &= \sum_{d} r_d \frac{\#_d(y|w)}{\#(w)} = \sum_{d} r_d P_{\text{class,d}}(y|w) \end{aligned} \tag{8}$$

since $\sum_d r_d = 1$. The $P_{\text{class,d}}(y|w)$ is the probability of dialect pronunciation y for common language word w in dialect d. Equation (8) indicates that the pronunciation probabilities of mixed dialect language models are calculated by interpolating $P_{\text{class,d}}(y|w)$. Since each $P_{\text{class,d}}(y|w)$ satisfies $\sum_y P_{\text{class,d}}(y|w) = 1$, the sum of probabilities of all pronunciations, $\sum_y P_{\text{class,mix}}(y|w)$, equals one.

$$\sum_{y} P_{\text{class,mix}}(y|w) = \sum_{d} r_d \sum_{y} P_{\text{class,d}}(y|w) = 1. \tag{9}$$

Thus, the $P_{\text{class,mix}}(y|w)$ is valid for the probabilities.

Mixed dialect language models constructed in this way have the vocabularies of all dialect language models and they enable ASR systems to recognize utterances in various dialects.

B. Integration of Recognition Results

This section describes how recognition accuracy is improved by integrating recognition results with each dialect language model instead of constructing mixed dialect language models. The method described in Section III-A had to conduct ASR numerous times to cover combinations of mixing proportions. This is unrealistic for methods that work on applications with speech interfaces because they require massively parallel machines to work in real-time; methods that require fewer instances of ASR are necessary. If the proportions in which utterances or speakers are mixed are not required, improved recognition results might be obtained by integrating reliable parts of recognition results with each individual dialect language model. A method like this has to conduct ASR only the same number of times as the number of dialects and dialect ASR is intended to even work on processors for consumers.

This section investigates the integration of a one-best recognition result with each dialect language model. This method first recognizes an input utterance with each single dialect language model, and then integrates reliable parts of each result to decrease the number of errors. We adopt the method of Recognizer Output Voting Error Reduction (ROVER) [21], which is one of the representative methods of integrating recognition results. ROVER first obtains word-wise alignment of multiple recognition results to express them as a sequence of word tuples, and then chooses a word from each tuple by voting. Voting adopts the measure of a weighted average of word probabilities in a tuple and confidence scores given to each recognized word by ASR. Let $\#(m)$ be the number of recognition results for an utterance, i.e., the number of language models to be compared. For each word tuple, the score of word w in the tuple is calculated as:

$$\text{Score}(w) = \alpha \frac{\#(w)}{\#(m)} + (1 - \alpha) C(w), \tag{10}$$

where $\#(w)$ $(0 \leq \#(w) \leq \#(m))$ is the number of w that appears in the tuple, $C(w)$ $(0 \leq C(w) \leq 1)$ is the confidence score based on the confidence measures of all w that appear in the tuple, and α $(0 \leq \alpha \leq 1)$ is a parameter to determine the weights of both terms. The value of $\text{Score}(w)$ ranges from zero to one. Fiscus [21] proposed two ways of determining $C(w)$:

$$C_1(w) = \frac{1}{\#(w)} \sum_{1 \leq i \leq \#(m), w_i = w} \text{Confidence}(w_i) \quad \text{and} \tag{11}$$

$$C_2(w) = \max_{1 \leq i \leq \#(m), w_i = w} \text{Confidence}(w_i), \tag{12}$$

where w_i is the word of the i-th recognition result in the tuple, and Confidence (w_i) is the confidence measure of w_i $(0 \leq \text{Confidence}(w_i) \leq 1)$ output by ASR. We adopted $C_1(w)$ as $C(w)$ in our experimental evaluation, where the parameters to be determined were α and Confidence(ε), which is the confidence measure for empty word ε. We empirically determined these parameters that maximized the ASR accuracy of the test set as $\alpha = 0.6$ and Confidence $(\varepsilon) = 0.4$ in the experimental evaluations. This decision on parameters was not critical for ASR accuracy in the preliminary experiment.

IV. Experimental Evaluation

This section compares two methods of mixed dialect speech recognition explained in Section III in terms of recognition accuracies. Prior to that, Section IV-B explained our calculation of the correlation coefficients between recognition likelihoods and accuracies (Section III-A) to confirm the validity of maximized recognition likelihoods.

TABLE I
AGE AND GENDER OF SPEAKERS (M: MALES AND F: FEMALES). NOS. 1–5 CORRESPOND TO INDICES OF SPEAKERS OF FIVE DIALECTS

	#1	#2	#3	#4	#5
Tokyo	38, F	36, M	32, M	25, M	21, F
Kansai	30, F	27, M	24, F	23, M	20, F
Kyushu	28, M	24, F	22, M	20, F	40, M
Tohoku	26, M	24, F	21, F	20, F	26, M
San-yo	49, F	24, M	22, M	21, F	21, M

A. Experimental Setting

Testing data: The experimental evaluation was conducted with utterances by 25 speakers: five speakers for each of five main dialects of the Japanese language. First we chose five dialects to be evaluated on the basis of the population of their speakers. We then chose five speakers that had lived in the same area from birth to 18 years of age for each dialect so that they could utter their natural ways of speaking. We did not assume their linguistic knowledge of their dialects. The only linguistic requirement for the speakers was that they were conscious of their ability to speak their natural dialects. Table I summarizes the age and gender of each speaker. The five main dialects of Japanese (see Fig. 1) are listed below.

- Tokyo dialect: We adopted an ordinary model of the common language as the language model for the Tokyo dialect.
- Kansai dialect (around Osaka)
- Kyushu dialect (around Fukuoka)
- Tohoku dialect (around Aomori)
- San-yo dialect (around Okayama and Hiroshima, west of Osaka)

The speakers read sentences in their own dialects aloud that they had translated in advance from the same 100 sentences in the common language. The original sentences of the common language were from the daily life category of the Yahoo! *Chiebukuro* (Q&A) corpus (second edition)[1]. Sentences in the corpus had a style similar to the spoken language. The out-of-vocabulary (OOV) rate of the test set was 0.7% in the language model of the common language. The hit-rate of n-grams were 73.38% for trigrams, 20.40% for bigrams, and 6.23% for unigrams. We did not forbid them to change the word order. The speakers of the Tokyo dialect simply read the original sentences.

Here is an example of a set of the original (Tokyo dialect) and four translated sentences by male speakers in their twenties:

Original *Naniga ikenaindarō*. (What causes the trouble?)
Kansai *Naniga akannoyaro*. (speaker: Kansai #4)
Kyushu *Naniga ikancharō*. (speaker: Kyushu #3)
Tohoku *Naniga ikenēndabe*. (speaker: Tohoku #1)
San-yo *Naniga ikennonjarō*. (speaker: San-yo #2)

Training Data: We prepared three kinds of corpora: corpora to train rules to simulate dialect corpora, a large common language corpus to be transformed, and speech data of the common language to train an acoustic model.

The rules to simulate the dialect corpora were trained with parallel corpora edited by National Institute for Japanese Language and Linguistics [22] for each dialect. The parallel corpora had 9,000–13,000 common language words and the corresponding dialect sentences for each dialect.

The common language corpus for simulation was a subset of the Yahoo! Q&A corpus, but they did not overlapped with the spoken sentences at all. The language models in this experiment were trained with three million sentences (71.2 million words) selected from all sentences that belonged to the daily life category; the method of sentence selection was corpus filtering based on perplexity [23], which selected a specific number of sentences in order of perplexity per word on a language model from the smallest, i.e., log-likelihood per word from the largest. The experiment adopted blog sentences[2] from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [24] to train the language model for perplexity calculations with. The value of n for word n-gram models was set to three[3]. Dialect language models had vocabulary composed of words that appeared more than ten times in the common language corpus, whose size was 42,845. No cutoffs of 2-gram or 3-gram were conducted. These conditions resulted in the language model for the common language composed of 42,845 unigrams, 474,161 bigrams and 979,766 trigrams.

The acoustic model consisted of 2,000 states each of which modeled likelihoods of acoustic features as a triphone hidden Markov model (HMM) whose output probabilities were modeled with a Gaussian mixture model (GMM) composed of 64 Gaussian distributions. The acoustic features were based on Mel-frequency cepstrum coefficients (MFCCs); they consisted of 25 dimensions: 12 dimensions of MFCC, 12 dimensions of Δ MFCC, and one dimension of Δ Power. The frame width for feature calculations was set to 25 msec., and the frame interval was the set to 10 msec. This experiment adopted utterances from 500 speakers in the Corpus of Spontaneous Japanese (CSJ) [25] to train the acoustic model with (70.2 hours). The CSJ was composed of speech from lectures, which is more suitable for spoken language recognition than speech that is read aloud. Triphones that did not appear in the CSJ were supplied with speech from phoneme-balanced sentences extracted from Japanese Newspaper Article Sentences (JNAS) [26], which consisted of 23.3 hours of speech by 308 speakers. These data for training acoustic models were independent of a specific dialect or speaker.

ASR Engine: We adopted Julius[4][27] as an ASR engine. The recognition results, which were output as common language sentences, were compared with the original sentences.

B. Correlation Between Recognition Likelihoods and Accuracies

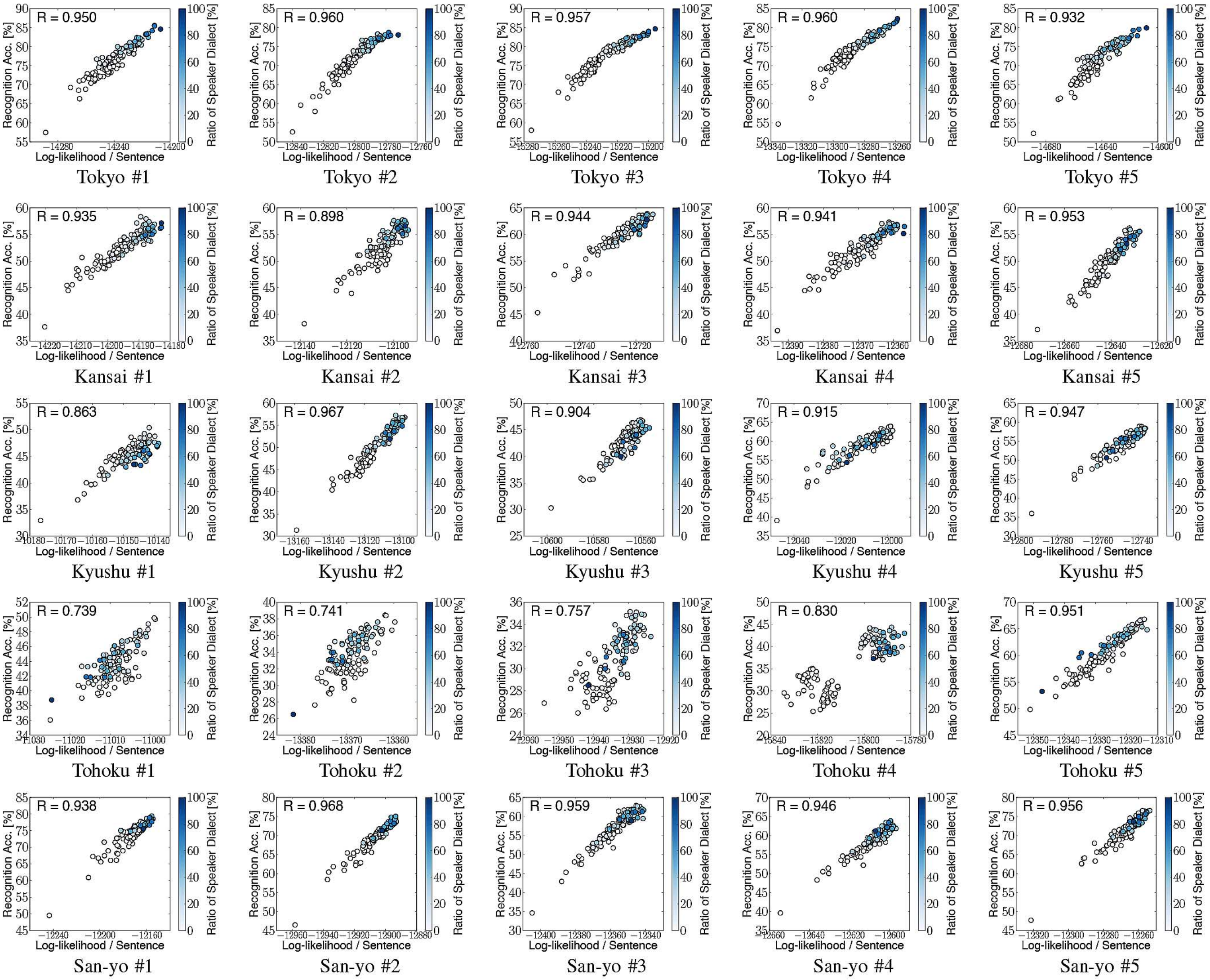
The experiment for this method treated proportions of mixed dialects as being discrete and recognized input utterances with language models for each combination of proportions of mixed

[1] Distributed by Yahoo! JAPAN and National Institute of Informatics.

[2] Only sentences whose word segmentation was manually annotated.

[3] Julius actually required forward 2-gram models and backward 3-gram models.

[4] http://julius.sourceforge.jp/



Tokyo #1 Tokyo #2 Tokyo #3 Tokyo #4 Tokyo #5

Kansai #1 Kansai #2 Kansai #3 Kansai #4 Kansai #5

Kyushu #1 Kyushu #2 Kyushu #3 Kyushu #4 Kyushu #5

Tohoku #1 Tohoku #2 Tohoku #3 Tohoku #4 Tohoku #5

San-yo #1 San-yo #2 San-yo #3 San-yo #4 San-yo #5

Fig. 3. Log-likelihood per utterance (X-axis) and recognition accuracy (Y-axis). Each graph corresponds to one speaker. Each data point corresponds to combination of proportions of mixed dialects. R denotes correlation coefficients.

dialects, as explained in Section III-A. The proportions of mixed dialects were in units of 20%. The number of combinations of proportions of mixed dialects, $|P|$, was calculated with Eq. (2) with $M = 5$ and $K = 5$:

$$|P| = \frac{(5+5-1)!}{5!(5-1)!} = \frac{9!}{5!4!} = 126. \tag{13}$$

Validation of Maximized Likelihoods of Recognition to Improve Recognition Accuracies: Fig. 3 plots the correlations between the logarithms of recognition likelihoods per utterance (X-axis) and the word recognition accuracy (Y-axis) for each speaker. We switched mixed dialect language models and conducted ASR of a whole set of utterances for each fixed mixed dialect language model, which corresponded to each data point. The values of the correlation coefficients were at least 0.739, and more than 0.8 for 22 out of 25 speakers, which indicated strongly positive correlations. These results validated the maximized recognition likelihoods to improve recognition accuracies. The color of most data points that maximized the recognition accuracy indicated 20-60% of the proportion of speaker dialects. This indicates that dialect utterances not only include the dialect in the area where the speaker lives, but also other dialects. That is why mixed dialect language models should improve recognition accuracy.

Possibility of an Updating Model: We evaluated the possibility of estimating the proportions of dialect mixing with an updating model in a preliminary experiment. This experiment investigated a model of updating mixing proportions with a regression model

$$r'_d = \sum_d a_d r_d + C. \tag{14}$$

This model estimated mixing proportions for the next utterance, r'_d, for each dialect d by those for the current utterance, r_d. We determined the mixing proportions of each utterance on the basis of likelihood maximization, and calculated each coefficient a_d and constant C for each speaker with the least squares method and the coefficient of determination, R^2, between the estimated and actual mixing proportions. Since the sum of r'_d is

TABLE II
COMPARISON OF WORD RECOGNITION ACCURACIES [%] FOR COMMON LANGUAGE, SINGLE DIALECT (SAME AS THAT OF SPEAKERS) AND MIXED DIALECT LANGUAGE MODELS

Tokyo speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	84.6	**79.2**	84.2	81.7	79.9	81.9
Mixed dialect (fixed)	**84.7**	78.1	**84.7**	**82.4**	**80.0**	**82.0**
Equally mixed	77.8	74.1	80.0	75.5	72.5	76.0
Common language model	**84.7**	78.1	**84.7**	**82.4**	**80.0**	**82.0**
Kansai speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	**61.4**	**60.1**	**67.3**	**60.3**	**60.0**	**61.8**
Mixed dialect (fixed)	56.3	55.9	63.8	56.5	55.6	57.6
Equally mixed	55.8	57.0	62.1	57.3	52.3	56.9
Single dialect model	57.2	56.6	62.8	55.2	53.3	57.0
Common language model	51.6	49.4	61.2	50.9	50.1	52.6
Kyushu speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	**49.4**	**57.5**	**47.2**	**66.6**	**59.9**	**56.1**
Mixed dialect (fixed)	47.4	56.8	45.3	62.8	58.4	54.1
Equally mixed	47.4	54.3	44.5	61.5	57.3	53.0
Single dialect model	43.5	51.9	39.9	54.4	50.6	48.1
Common language model	44.6	46.0	41.2	57.5	50.4	47.9
Tohoku speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	**49.7**	**42.7**	**37.9**	42.8	**67.9**	**48.2**
Mixed dialect (fixed)	**49.7**	37.6	32.9	**43.0**	64.8	45.6
Equally mixed	45.2	37.3	35.1	39.2	65.2	44.4
Single dialect model	38.8	26.5	28.5	37.3	53.2	36.9
Common language model	44.5	33.0	28.9	33.3	58.8	39.7
San-yo speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	**81.8**	**76.1**	**65.2**	**66.0**	76.1	**73.0**
Mixed dialect (fixed)	78.6	73.6	61.7	61.9	**76.5**	70.5
Equally mixed	76.9	72.0	61.8	60.9	74.1	69.1
Single dialect model	75.6	71.3	58.8	61.2	73.6	68.1
Common language model	66.1	65.5	51.7	54.4	66.3	60.8

TABLE III
MEAN PROPORTIONS OF MIXED DIALECTS ESTIMATED FOR EACH SPEAKER

Tokyo speakers	Estimated proportion [%]				
	Tokyo	Kansai	Kyushu	Tohoku	San-yo
Speaker #1	**70.8**	9.8	6.2	9.0	4.2
Speaker #2	**69.2**	9.2	4.2	11.0	6.4
Speaker #3	**72.4**	9.0	6.0	14.0	15.0
Speaker #4	**70.0**	11.2	4.2	7.6	7.0
Speaker #5	**68.0**	9.6	4.8	9.2	8.4
Kansai speakers	Estimated proportion [%]				
	Tokyo	**Kansai**	Kyushu	Tohoku	San-yo
Speaker #1	20.0	**39.4**	13.8	13.4	13.4
Speaker #2	26.0	**27.8**	16.6	11.2	18.4
Speaker #3	30.2	**34.8**	6.0	14.0	15.0
Speaker #4	24.2	**42.0**	8.6	12.4	12.8
Speaker #5	19.2	**39.4**	11.6	15.2	14.6
Kyushu speakers	Estimated proportion [%]				
	Tokyo	Kansai	**Kyushu**	Tohoku	San-yo
Speaker #1	23.0	19.6	**29.4**	8.4	19.6
Speaker #2	26.6	15.6	**35.2**	9.8	12.8
Speaker #3	25.6	26.0	**22.4**	12.0	14.0
Speaker #4	28.0	32.0	**13.2**	7.2	19.6
Speaker #5	24.6	26.8	**19.6**	9.2	19.8
Tohoku speakers	Estimated proportion [%]				
	Tokyo	Kansai	Kyushu	**Tohoku**	San-yo
Speaker #1	30.6	19.0	13.4	**20.0**	17.0
Speaker #2	21.8	15.6	25.4	**24.0**	13.2
Speaker #3	20.6	17.2	18.0	**31.4**	12.8
Speaker #4	12.4	12.6	14.8	**44.4**	15.8
Speaker #5	38.4	8.4	16.0	**25.2**	12.0
San-yo speakers	Estimated proportion [%]				
	Tokyo	Kansai	Kyushu	Tohoku	**San-yo**
Speaker #1	17.6	21.6	16.6	3.4	**40.8**
Speaker #2	31.2	14.6	17.8	7.8	**28.6**
Speaker #3	17.2	19.0	18.6	6.0	**39.2**
Speaker #4	27.4	20.8	17.8	7.6	**26.4**
Speaker #5	27.0	17.2	14.6	7.8	**33.4**

always one, Tokyo dialect was excluded from the estimation to prevent linear dependence on explanatory variables, r_d.

The calculated values of R^2, composed of 100 numbers (four dialects except Tokyo to be estimated for each of 25 speakers) ranged from 0.0032 to 0.1488, whose mean value was 0.0439. The results did not lead to significant correlations in the estimated and actual mixing proportions. This means that this updating model was statistically insignificant, which made it difficult to estimate the mixing proportions of the next utterance with.

Improved Recognition Accuracies: Recognition accuracies with maximized recognition likelihoods were compared to those with language models of the common dialect and a single dialect. Table II compares recognition accuracies under different conditions. Single dialect language models were those with the same dialect as that of a speaker. Mixed dialect language models (fixed) were determined so that the whole recognition likelihood for all utterance of each speaker was maximized. Mixed dialect language models (variable) were switched for each utterance so that the recognition likelihood for the utterance was maximized. The mean recognition accuracies of all speakers except for the Tokyo dialect with mixed language models were higher than those with the common language and single dialect language models. Recognition with the equally mixed language model, which contained 20% of each language model, resulted in slightly higher recognition accuracies than single dialect models. The mixed language model fixed for each speaker based on likelihood maximization outperformed those with the equally mixed language model. The mixed dialect language models estimated for each utterance outperformed both of them. That was because dialects were mixed in different ways [5] in each utterance as we mentioned in Section I. This showed the importance of mixing language models at precise proportions as well as just mixing them.

The proportions of mixed dialects were averaged for each utterance to estimate the speaker dialects. Table III lists the averaged proportions. The estimated proportions differed for speakers even in the same area. These results support the assumption that multiple dialects are mixed in different ways to be spoken by individual speakers.

We also compared linguistic prediction accuracies according to the value of test perplexity per phoneme on the test set composed of a set of translated sentences by one speaker (#1) for each dialect. Unknown words were based on the phoneme 0-gram model, where all phonemes appeared at the same probability regardless of the context. Table IV summarizes the test perplexity per phoneme of language models of the common language, single dialect models, and the mixed dialect adopted in Table II. The values of perplexity were improved by the mixed dialect model in all dialects except for the Tokyo dialect (= common language), which means the mixed dialect language models predicted utterances by dialect speakers more precisely than the common language model. Furthermore, some single dialect models could improve the perplexity of other dialects, e.g., the Kansai dialect model improved the perplexity of the San-yo dialect more than the common language model (from 6.947 to 6.599). Kansai and San-yo are neighboring

TABLE IV
TEST PERPLEXITY PER PHONEME OF TRANSCRIPTION OF DIALECT UTTERANCES. TESTED DIALECT SENTENCES WERE WRITTEN BY SPEAKER #1 OF EACH DIALECT

	Dialect of the speaker					
LM	Tokyo	Kansai	Kyushu	Tohoku	San-yo	All
Common language	4.530	6.270	6.754	7.280	6.947	6.259
Kansai	5.893	6.068	6.754	8.260	6.599	
Kyushu	6.235	7.106	6.171	8.591	6.878	
Tohoku	9.128	9.655	10.077	9.327	10.859	
San-yo	5.905	6.617	6.736	8.090	6.024	
Mixed dialect	4.530	**5.686**	**5.677**	**6.201**	**5.603**	**5.504**

TABLE V
WORD RECOGNITION ACCURACY [%] WITH THE ROVER METHOD AND MAXIMIZED RECOGNITION LIKELIHOODS

Tokyo speakers	#1	#2	#3	#4	#5	Mean
ROVER	83.4	77.7	84.2	80.8	78.0	80.8
Maximized likelihood	84.6	**79.2**	84.2	81.7	79.9	81.9
Common language model	**84.7**	78.1	**84.7**	**82.4**	**80.0**	**82.0**
Kansai speakers	#1	#2	#3	#4	#5	Mean
ROVER	59.8	59.5	65.8	59.6	56.7	60.3
Maximized likelihood	**61.4**	**60.1**	**67.3**	**60.3**	**60.0**	**61.8**
Single dialect model	57.2	56.6	62.8	55.2	53.3	57.0
Kyushu speakers	#1	#2	#3	#4	#5	Mean
ROVER	**49.7**	57.4	47.1	65.7	59.6	55.9
Maximized likelihood	49.4	**57.5**	**47.2**	**66.6**	**59.9**	**56.1**
Single dialect model	43.5	51.9	39.9	54.4	50.6	48.1
Tohoku speakers	#1	#2	#3	#4	#5	Mean
ROVER	47.8	42.5	**37.9**	**46.1**	66.5	**48.2**
Maximized likelihood	**49.7**	**42.7**	**37.9**	42.8	**67.9**	**48.2**
Single dialect model	38.8	26.5	28.5	37.2	53.2	36.9
San-yo speakers	#1	#2	#3	#4	#5	Mean
ROVER	80.2	75.3	63.7	64.4	75.5	71.8
Maximized likelihood	**81.8**	**76.2**	**65.2**	**66.0**	**76.1**	**73.0**
Single dialect model	75.6	71.3	58.8	61.2	73.6	68.1

dialects and they share some words or pronunciations. As a result, the mixed dialect model additionally improved the perplexity of San-yo dialect speakers more than the single dialect model of the San-yo dialect. This supported our hypothesis that mixing other dialects with the model improved dialect speech recognition.

C. Integration of Multiple Recognition Results

Recognition accuracies were calculated with the ROVER method (Section III-B) to compare them with those with the other methods. Table V summarizes the recognition accuracies from the integrated results. This method also improved recognition accuracies compared with just the common language or a single dialect language model, while it spared us the effort of constructing mixed dialect language models and recognition with language models a large number of times.

We evaluated the real-time factor (RTF) of the integration method (ROVER) for each speaker as:

$$\mathrm{RTF} = \frac{\sum_u (t_{i,u} + \max_m t_{c,u,m})}{\sum_u t_u}. \tag{15}$$

Here, t_u is the time for the utterance, $t_{i,u}$ is the integration time for utterance u, and $t_{c,u,m}$ is the time to calculate the recognition of utterance u with model m. We recognized utterance u with five dialect models in parallel, and $\max_m t_{c,u,m}$ was maximum value for the calculation time with the five models. The results for each speaker ranged from 0.704 to 0.915. The maximum value was less than 1.0 (= real-time); thus, the integration method worked well in real-time. We evaluated the RTF on a workstation that had "Intel Xeon CPU X5560" (clock frequency: 2.80 GHz). We assumed that we could use more than five threads in parallel.

We could also estimate the time to recognize sentences with all of 126 mixed dialect language models. If we could use eight threads in parallel (this is generally assumed if we use a single workstation), its real-time factor would be nearly 14 in the worst case.

We can also estimate the calculation time of mixed dialect language models from this result. If we can use eight threads in parallel (it is general assumption if we recognize with a single workstation), its real-time factor will be nearly 14 in the worst case.

D. Discussion

Performance and other aspects are compared for the methods in Sections IV-B and Section IV-C.

Maximized recognition likelihoods (Section IV-B) outperformed integration with the ROVER method in recognition accuracies. The correlation between likelihood values and recognition accuracies in Fig. 3 implies that this method yielded almost the upper bound for recognition accuracies since this method maximized the likelihood values of each utterance. The main problem is the number of language models to be compared. Although many computers are available due to the popularity of cloud computing, it is still difficult to use a lot of computers for casual ASR. This increases the computational cost of the method and disrupts real-time processing in actual applications. A large number of mixed dialect language models also need to be constructed beforehand.

The integration of recognition results (Section IV-C) only requires single dialect language models and only has to compare the same number of recognition results. This enables ASR systems to work in real-time with multi-core processors available for consumers by conducting ASR with each single dialect language model in parallel. The main problem was that this method did not estimate the proportions of mixed dialects in input utterances. Dialect estimates of input utterances should be conducted with the method of maximized recognition likelihoods as batch processing.

Word recognition accuracies differed for dialects. For example, the recognition accuracies for the Tohoku and Kyushu dialects were lower than those for the other dialects. These dialects have phonemes with very different characteristics from those of the common language. Including acoustic features would improve the recognition accuracy of these dialects.

We conducted ASR with an adapted acoustic model to the Kansai dialect in a preliminary experiment. Although it is the best to introduce different phoneme sets and train acoustic models based on the sets for each dialect, this is difficult to do due to the shortage of data to train the features of the new phonemes. Instead, we adapted the common acoustic model to each dialect, while the phoneme set was still the same. The utterances by each speaker were recognized in our method with the acoustic model adapted to utterances by four speakers

TABLE VI
WORD RECOGNITION ACCURACIES [%] OF KANSAI DIALECT SPEAKERS WITH ACOUSTIC MODELS ADAPTED TO THE KANSAI DIALECT

Kansai speakers	#1	#2	#3	#4	#5	Mean
Mixed dialect (variable)	63.3	62.0	74.8	66.7	62.5	65.9
Mixed dialect (fixed)	59.9	61.8	71.5	62.9	59.6	63.1
ROVER	59.1	59.7	69.9	62.3	58.4	61.9

with the same dialect except for the targeted speaker (4 × 100 utterances altogether).

Table VI summarizes the word recognition accuracies with the adapted acoustic models. Mixed dialect language models with variable and fixed mixing proportions resulted in mean recognition accuracies that corresponded to 4.1 and 5.5 points higher than those with the unadapted acoustic model in Table II. The adapted acoustic models with the ROVER method improved the recognition accuracy by 1.5 points more than that with the unadapted acoustic model in Table V. Only a small adaptation dataset with the unmodified phoneme set achieved these improvements. We intend to adapt acoustic models with more speech data or customization of the phoneme set for each dialect in future work.

V. CONCLUSIONS

We developed a method of recognizing mixed dialect utterances with multiple dialect language models by using small parallel corpora of the common language and a dialect and a large common language linguistic corpus. Our two main methods were *maximization of recognition likelihoods* and *integration of recognition results*. The former constructed mixed dialect language models by using the weighted average of n-gram probabilities and pronunciation probabilities of language models to be mixed. The mixing proportion was chosen for each utterance so that it achieved the largest recognition likelihood of the utterance. One of the main contributions of this paper is that the estimates were conducted for individual utterances and not for individual speakers. The latter integrated recognition results with each single language dialect model to output results that included fewer errors. The former output results with higher recognition accuracies, while the latter improved results with lower calculation costs.

Future work will be twofold. The first is to introduce differences in acoustic features in dialects. We modeled only linguistic differences in dialects, i.e., acoustic features such as the those in a phoneme set and the feature distributions of each phoneme were assumed to be the same as those of the common language. This resulted in lower recognition accuracies for dialects that had very different acoustic features from the common language. If we had also included acoustic modeling in our system, recognition accuracy would be much better even for such dialects. The second is to apply this method to speech data mining. If dialects are treated as differences in vocabulary, various attributes of speakers, including jobs they do and groups they belong to, will be modeled in the same way. Extracting how spoken words are chosen should lead us to discovering knowledge with our method on jobs, majors, communities, favorites, and other kinds of characteristics of speakers.

REFERENCES

[1] Y. Akita, M. Mimura, and T. Kawahara, “Automatic transcription system for meetings of the Japanese National Congress,” in *Proc. Interspeech ‘09*, 2009, pp. 84–87.

[2] A. Lambourne, J. Hewitt, C. Lyon, and S. Warren, “Speech-based real-time subtitling services,” *Int. J. Speech Technol.*, vol. 7, no. 4, pp. 269–279, 2004.

[3] M. Schuster, “Speech recognition for mobile devices at Google,” in *PRICAI 2010: Trends in Artif. Intell.*, 2010, pp. 8–10, Springer.

[4] J. Aron, “How innovative is Apple’s new voice assistant, Siri?,” *New Scientist*, vol. 212, no. 2836, p. 24, 2011.

[5] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, “Automatic estimation of dialect mixing ratio for dialect speech recognition,” in *Proc. Interspeech ‘13*, 2013, pp. 1492–1496.

[6] W. Wolfram, “Dialect awareness, cultural literacy, and the public interest,” in *Ethnolinguistic Diversity and Education: Language, Literacy and Culture*, M. Farr, Ed. *et al.* London, U.K.: Routledge, 2009, ch. 6, pp. 129–149.

[7] D. A. Cruse*, Lexical Semantics*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[8] L. J. Brinton and M. Fee*, English in North America*, ser. ser. The Cambridge history of the English language. Cambridge, U.K.: The Press Syndicate of the Univ. of Cambridge, 2001, vol. 6.

[9] E. R. Thomas, “Rural Southern white accents,” in *The Americas and the Caribbean*, ser. ser. Varieties of English, E. W. Schneider, Ed. Berlin, Germany: Mouton de Gruyter, 2008, vol. 2, pp. 87–114.

[10] D. Ramon, “We are one people separated by a common language,” in *Viagra, Prozac, and Leeches*. Bloomington, IN, USA: iUniverse, 2006, pp. 203–206.

[11] H. B. Woods, “A socio-dialectology survey of the English spoken in Ottawa: A study of sociological and stylistic variation in Canadian English,” Ph.D. dissertation, Univ. of British Columbia, Vancouver, BC, Canada, 1979.

[12] N. Hirayama, S. Mori, and H. G. Okuno, “Statistical method of building dialect language models for ASR systems,” in *Proc. COLING ‘12*, 2012, pp. 1179–1194.

[13] M. Caballero, A. Moreno, and A. Nogueiras, “Multidialectal acoustic modeling: A comparative study,” in *Proc. ITRW Multilingual Speech Lang. Process.*, 2006.

[14] D. Lyu, R. Lyu, Y. Chiang, and C. Hsu, “Speech recognition on code-switching among the Chinese dialects,” in *Proc. ICASSP ‘06*, 2006, vol. 1, pp. 1105–1108.

[15] P. C. Ching, T. Lee, and E. Zee, “From phonology and acoustic properties to automatic recognition of Cantonese,” in *Proc. Speech, Image Process. Neural Netw.*, 1994, pp. 127–132.

[16] D. R. Miller and J. Trischitta, “Statistical dialect classification based on mean phonetic features,” in *Proc. ICSLP ‘96*, 1996, vol. 4, pp. 2025–2027.

[17] R. Chitturi and J. H. L. Hansen, “Dialect classification for online podcasts fusing acoustic and language based structural and semantic information,” in *Proc. ACL HLT 2008 Short Papers*, 2008, pp. 21–24.

[18] H. Elfardy and M. Diab, “Sentence level dialect identification in arabic,” in *Proc. ACL ‘13*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, vol. 2, pp. 456–461.

[19] A. Stolcke, “SRILM—an extensible language modeling toolkit,” in *Proc. Interspeech*, 2002.

[20] R. Kneser and H. Ney, “Improved backing-off for M-gram language modeling,” in *Proc. ICASSP ‘95*, 1995, vol. 1, pp. 181–184.

[21] J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” in *IEEE Workshop Autom. Speech Recogn. Understand.*, 1997, pp. 347–354, IEEE.

[22] Database of Spoken Dialects all over Japan: Collection of Japanese Dialects (In Japanese), Kokushokankokai, vol. 1–20, pp. 2001–2008, Nat. Inst. for Japanese Lang., and Linguist., Ed..

[23] T. Misu and T. Kawahara, “A bootstrapping approach for developing language model of new spoken dialogue systems by selecting Web texts,” in *Proc. ICSLP ‘06*, 2006, pp. 9–12.

[24] K. Maekawa, “Balanced corpus of contemporary written Japanese,” in *Proc. 6th Workshop Asian Lang. Resources*, 2008, pp. 101–102.

[25] K. Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *Proc. ISCA & IEEE Workshop Spontaneous Speech Process. Recogn.*, 2003, pp. 7–12.

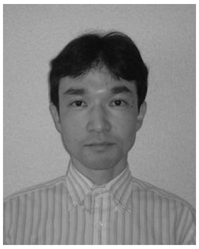
[26] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Acoust. Soc. Jpn. (English Edition)*, vol. 20, pp. 199–206, 1999.

[27] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *Proc. Eurospeech '01*, 2001, pp. 1691–1694.

Naoki Hirayama received his B.E. degree in 2012, and M.S. degree in informatics in 2014 all from Kyoto University, Japan. He currently works for Toshiba Solutions Corporation. His research interest was speech recognition of dialects when he was a student. He is a member of IPSJ.

Koichiro Yoshino received his B.A. degree in 2009 from Keio University, M.S. degree in informatics in 2011, and Ph.D. degree in informatics in 2014 from Kyoto University, respectively. Currently, he is a Research Fellow (PD) of Japan Society for Promotion of Science. His research interests include spoken language processing, especially spoken dialogue system, syntactic and semantic parsing, and language modeling. He received the JSAI SIG-research award in 2013. He is a member of IPSJ, JSAI, and ANLP.

Katsutoshi Itoyama received the B.E. degree in 2006, the M.S. degree in informatics in 2008, and the Ph.D. degree in informatics in 2011 all from Kyoto University. He is currently an Assistant Professor of the Graduate School of Informatics, Kyoto University, Japan. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He received the 24th TAF Telecom Student Technology Award and the IPSJ Digital Courier Funai Young Researcher Encouragement Award. He is a member of IPSJ, ASJ, and IEEE.



Shinsuke Mori received B.S., M.S., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan in 1993, 1995, and 1998, respectively. Then he joined Tokyo Research Laboratory of International Business Machines Co. Ltd. (IBM). Since 2007, he has been an Associate Professor of Academic Center for Computing and Media Studies, Kyoto University. His research interests include computational linguistics and natural language processing. He received the IPSJ Yamashita SIG Research Award in 1997, IPSJ Best Paper Award in 2010 and 2013, and 58th OHM Technology Award from Promotion Foundation for Electrical Science and Engineering in 2010. He is a member of IPSJ, ANLP, and ACL.

Hiroshi G. Okuno (M'03–SM'06–F'12) received the B.A. and Ph.D. from the University of Tokyo in 1972 and 1996, respectively. He worked for NTT, JST, Tokyo University of Science, and Kyoto University. He is currently a Professor of Graduate Program for Embodiment Informatics, Graduate School of Creative Science and Engineering, Waseda University, and a Professor Emeritus, Kyoto University.

He was Visiting Scholar at Stanford University from 1986 to 1988. He is currently engaged in computational auditory scene analysis, music information processing and robot audition. He received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001, 2005, 2010, and 2013, IEEE/RSJ IROS-2001 and 2006 Best Paper Nomination Finalist, and NTF Award for Entertainment Robots and Systems in 2010. He co-edited *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates, 1998), *Advanced Lisp Technology*(Taylor and Francis, 2002), and *New Trends in Applied Artificial Intelligence (IEA/AIE)*(Springer, 2007). He is a fellow of the Japanese Society for Artificial Intelligence, and a member of AAAI, ACM, ASA, RSJ, IPSJ, JSSST, and JCSST.