Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition

Yuya Akita, *Member, IEEE*, and Tatsuya Kawahara, *Senior Member, IEEE*

***Abstract*—We propose a novel approach based on a statistical transformation framework for language and pronunciation modeling of spontaneous speech. Since it is not practical to train a spoken-style model using numerous spoken transcripts, the proposed approach generates a spoken-style model by transforming an orthographic model trained with document archives such as the minutes of meetings and the proceedings of lectures. The transformation is based on a statistical model estimated using a small amount of a parallel corpus, which consists of faithful transcripts aligned with their orthographic documents. Patterns of transformation, such as substitution, deletion, and insertion of words, are extracted with their word and part-of-speech (POS) contexts, and transformation probabilities are estimated based on occurrence statistics in a parallel aligned corpus. For pronunciation modeling, subword-based mapping between baseforms and surface forms is extracted with their occurrence counts, then a set of rewrite rules with their probabilities are derived as a transformation model. Spoken-style language and pronunciation (surface forms) models can be predicted by applying these transformation patterns to a document-style language model and baseforms in a lexicon, respectively. The transformed models significantly reduced perplexity and word error rates (WERs) in a task of transcribing congressional meetings, even though the domains and topics were different from the parallel corpus. This result demonstrates the generality and portability of the proposed framework.**

***Index Terms*—Automatic speech recognition (ASR), language model (LM), pronunciation model, spontaneous speech, statistical transformation.**

I. Introduction

THE targets of large-vocabulary continuous speech recognition (LVCSR) research have been extended in recent years to spontaneous speech such as telephone conversations, lectures, and meetings. Large corpora of conversational telephone speech (CTS), such as Switchboard and Fisher corpora, were collected and a number of LVCSR techniques have been developed with these corpora [1], [2]. The NIST Rich Transcription (RT) project has dealt with conversational speech recognition of meetings [3]. Such multiparty meetings have also been investigated by the AMI/AMIDA project conducted by European research institutes [4]. A number of speech and linguistic studies for lectures have been conducted using the Corpus of Spontaneous Japanese (CSJ) [5], which is a collection of academic lectures and public speeches. Oral presentations and seminars have also been recorded by European projects such as the TED corpus [6] and the CHIL project [7]. The LVCSR of classroom lectures has also been tackled mainly by universities [8], [9]. Speeches made at public gatherings such as in parliaments and courts are yet another target of LVCSR [10]–[13]. LVCSR and machine translation of European Parliament have been investigated in the TC-STAR project by several research institutes. We have also been studying LVCSR for the National Congress (Diet) of Japan.

These kinds of spontaneous speech have different acoustic and linguistic characteristics from those of read or broadcast news speech, since speakers are neither professional announcers nor narrators, and their behavior is spontaneous. Specifically, fast speaking rates, unclear articulation, and pronunciation variants are frequently observed. Redundant expressions, ungrammatical sentences, and disfluencies such as fillers and repairs are also observed. These characteristics are distinct especially in spoken Japanese, which is very different from written or formal Japanese. These acoustic and linguistic phenomena should be modeled in an LVCSR system to accurately transcribe spontaneous speech.

Hundreds of hours of well-matched training data have been collected to build LVCSR systems that cover these phenomena in some of these spontaneous speech recognition tasks including CTS and CSJ. However, collecting such large-scale corpora is usually impractical, because of the cost of manual transcriptions. Therefore, a practical solution is to combine corpora representing domain-specific characteristics (such as the proceedings of lectures and newspapers) with those representative of the characteristics of spontaneous speech (such as the Switchboard/Fisher corpora and the CSJ). Although this simple combinational approach is often adopted by many LVCSR systems [3], [8], [14], it still suffers from an intrinsic problem where the two kinds of characteristics cannot truly be harmonized. For example, it is impossible to estimate N-gram entries that have both topic words and fillers with this mixture-based training. Also, actual pronunciations of topic words are not necessarily obtained. Moreover, resulting models inevitably contain redundant, irrelevant, or inconsistent N-gram entries, which may cause confusion in LVCSR and thus degrade performance.

Manuscript received March 31, 2009; revised October 23, 2009. First published November 24, 2009; current version published July 14, 2010. This work was supported in part by the Strategic Information and Communications R&D Promotion Program (SCOPE), Ministry of Internal Affairs and Communications, Japan. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

The authors are with the Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan (e-mail: yuya@media.kyoto-u.ac.jp; kawahara@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TASL.2009.2037400

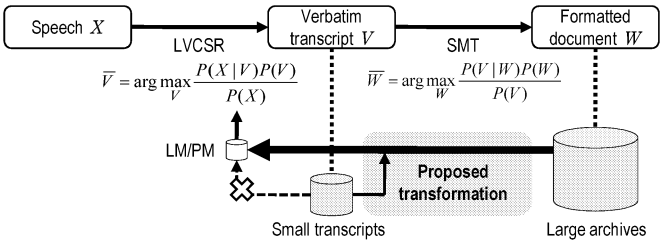

Fig. 1. Concept underlying proposed statistical transformation.

In this paper, we propose a novel approach to transform the language and pronunciation models to better model spontaneous speech, in which the statistical characteristics of spontaneous speech are modeled separately from task-dependent characteristics. Large document archives such as those of meeting minutes and lecture proceedings are generally easy to access. Although these archives are expected to match the target domain of LVCSR, they do not have spoken-style characteristics. Spoken-style models can be obtained from such large archives by modeling general transformations between orthographic-document and spoken styles. This transformation can be applied to various spontaneous speech recognition tasks, because of the task-independent framework. Another advantage of the proposed approach is that since the transformation models are much smaller than conventional language and pronunciation models, they can be trained with a smaller amount of training data.

This paper is organized as follows. The basic concept underlying the proposed transformation framework is described in Section II. Then, the actual methods of transforming the language and pronunciation models are fully explained in Sections III and IV. Section V describes typical characteristics of spontaneous Japanese, which are modeled with the proposed approach. The approach is evaluated for LVCSR with real data from congressional meetings. Our experiments and the results are discussed in Section VI. Section VII concludes the paper.

II. Basic Concept Underlying Statistical Transformation

The basic idea behind the proposed transformation framework is illustrated in Fig. 1. Spoken utterances in the conventional documentation process are transcribed once into a verbatim transcript V and then formatted as a document W. When considering the formatting process as a "translation" from V to W, an LVCSR system is followed by a statistical machine translation (SMT) [15] system, to automatically produce these documents. Here, a large number of spoken transcripts V are necessary to train the LVCSR system; however, faithful transcripts are expensive and the size of transcripts is actually limited. Therefore, direct estimates of the language and pronunciation models are virtually impossible with such small transcripts. In contrast, formatted documents W such as the minutes of meetings, which are manually edited, are archived on a large scale. These archives are often recorded in electronic form and are easily available. Consequently, the key idea behind the proposed framework is to transform W into language and pronunciation models for spontaneous speech recognition, by inverting the formatting process.

This inversion is modeled as a transformation model using transcripts V together with corresponding formatted texts W (called a parallel corpus hereafter). The transformation model for language model predicts N-gram entries and estimates their occurrence statistics from document archives. The transformation model for pronunciation model generates real pronunciation (surface form) entries from the orthodox pronunciation (baseform) of words found in the documents. Pronunciation probabilities are also predicted and assigned to all pronunciation entries by the model. The correspondences between spontaneous speech phenomena and orthographic expressions are solely targeted in both cases, and the correspondences are extracted in a more generalized form than word-level mappings. Consequently, the framework is expected to model transformation effectively and efficiently using a small amount of training data. Both transformations are described in detail in the following sections.

III. Statistical Transformation of Language Model

We address the statistical transformation of the language model based on the framework described above. The simulation and generation of spoken text from written or formal text have previously been proposed, as an alternative to the baseline mixture-based method mentioned in Section I. For example, Schramm *et al.* [16] proposed generating a simulated spoken-style text by randomly inserting fillers into written-style text. Petrik and Kubin [17] proposed restoring a literal transcript from non-literal documents by using speech recognition and phonetic matching. However, predicting a verbatim transcript V from a formatted text W is not usually deterministic, for example, insertion of fillers is arbitrary, though not random. The reliability of N-gram statistics was not always guaranteed in these approaches. As the purpose of transformation for LVCSR is to build a spoken-style language model, and not to obtain a text itself, it is more straightforward to estimate the language model statistics directly, rather than producing a text. Hori *et al.* [18] proposed using a weighted finite-state transducer (WFST) to transform a language model. However, linguistic expressions transformed into the spoken style were mostly handcrafted, and variations and statistics were not well represented. We extract the characteristics of the spoken style automatically from a corpus with a sufficient number of reliable statistics.

A. Basic Formulation

The concept underlying the proposed statistical transformation of the language model is illustrated in Fig. 2. It is based on the framework of statistical machine translation [15], where a sentence V of the target language is generated from a sentence W of the source language, which maximizes the posterior probability $P(V|W)$ using Bayes' rule

$$P(V|W) = \frac{P(W|V)P(V)}{P(W)}. \tag{1}$$

$P(W|V)$ is usually computed with a translation model.

We consider document-style and spoken-style languages as different languages, and denote the former as W and the latter as

TABLE I
MAJOR DIFFERENCES BETWEEN SPONTANEOUS SPEECH AND DOCUMENT-STYLE TEXT

Type	Typical cases in Japanese	Document-style text W		Spoken language V
Insertion	Fillers, end-of-sentence expressions, conjunctive	I think ...	→	(pause) **ah** I think ...
Deletion	Particles, end-of-sentence expressions	*watashi* ***wa*** *omoimasu* (I / my / me) (think) "*wa*" indicates the nominative case (I)	→	*watashi omoimasu*
Substitution	End-of-sentence expressions, colloquial expressions	***iroiro na*** (various)	→	***ironna***

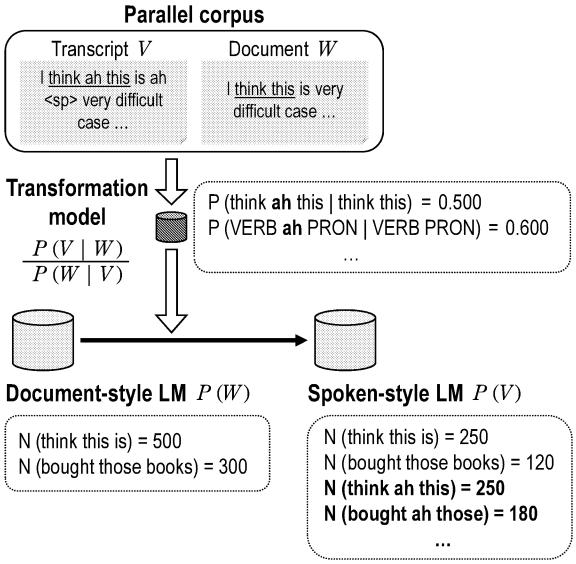

Fig. 2. Flow of language model transformation.

V. We can estimate spoken language model $P(V)$ by rewriting (1)

$$P(V) = P(W)\frac{P(V|W)}{P(W|V)}. \tag{2}$$

The conditional probabilities $P(V|W)$ and $P(W|V)$, i.e., the transformation model, can be estimated using a parallel corpus of faithful transcripts and corresponding document-style texts. When we assume N-gram language model for both V and W, actual estimation is carried out over N-gram entries and their statistics, because N-gram probabilities are basically proportional to corresponding N-gram statistics. Thus, (3) is derived from (2) to express an occurrence count of a spoken-style N-gram entry v using that of the corresponding document-style N-gram entry w found in an input corpus

$$N_{\mathrm{LM}}(v) = N_{\mathrm{LM}}(w)\frac{P(V|W)}{P(W|V)} \tag{3}$$

where N_{LM} denotes the occurrence count of N-gram entries.

When we further limit contextual information for $P(V|W)$ and $P(W|V)$, these probabilities are actually calculated as $P(v'|w')$ and $P(w'|v')$ for every pair of a spoken-style word sequence v' and a document-style word sequence w' that are found in the parallel aligned corpus. The possibility of transformation depends on the context of the target spoken-style expression; therefore, word contexts (neighboring words) are included in v' and w' to generate transformation patterns and to estimate probabilities.

In this paper, we consider three types of transformation: insertion, deletion, and substitution. Suppose a filler "ah" is inserted in the middle of "I think this is," as shown in Fig. 2, the words "think" and "this" are regarded as contexts, and transformation is modeled as $w' =$ *"think this"* and $v' =$ ' *"think ah this."* Similarly, deletion of a particle "for" from "to wait for him now" is expressed with $w' =$ *"wait for him"* and $v' =$ *"wait him,"* and substitution of words "am not" to the colloquial expression "ain't" in "today I am not fine because" is modeled as $w' =$ *"I am not fine"* and $v' =$ *"I ain't fine."* Then, N-gram entries such as "$*$ think ah," "think ah this" and "ah this $*$" are generated with input "$*$ think this" or "think this $*$," where "$*$" means any word. If context words of the input N-gram entries are not sufficient to provide transformed N-gram entries, for example, generating a trigram $v =$ "ah this $*$" after insertion of "ah" in "think this," necessary context words are added.

Thus, (3) is revised as follows:

$$N_{\mathrm{LM}}(v) = N_{\mathrm{LM}}(w)\frac{P(v'|w')}{P(w'|v')}. \tag{4}$$

In case that v can be generated from multiple w entries, the resulting occurrence count is a sum of counts estimated from respective w by (4)

$$N_{\mathrm{LM}}(v) = \sum_{w} N_{\mathrm{LM}}(w)\frac{P(v'|w')}{P(w'|v')}. \tag{5}$$

This context-dependent model improves precision, but encounters the problem of data sparseness, because the parallel corpus is usually small. To mitigate this problem, we present three models, based on back-off, linear interpolation, and maximum entropy (ME) schemes, which take into consideration part-of-speech (POS) information.

B. Probabilities to be Estimated

The language model of spontaneous speech is estimated within the above framework. Types of transformation can be classified into three categories, *viz.*, insertion, deletion, and substitution, as listed in Table I. We estimated conditional probabilities $P(v'|w')$ and $P(w'|v')$ for these cases. Note that a detailed analysis of differences between spoken and orthographic Japanese is described in Section V-A.

One of these three types is the insertion of words, especially fillers. Fillers are often observed at the beginning or the end of utterances and accompanied by a pause. However, their occurrence is not limited to these points, and the frequency depends on actual filler words and contexts. Hence, insertion probability

$P(v'|w')$ must be estimated. In contrast, deletion probability $P(w'|v')$ is always one for fillers, since fillers must be removed from transcripts for documentation.

The second type is the deletion of words such as particles. Not all particles are deleted in utterances, and particles cannot be inserted at all possible points when transforming a transcript into a document-style text. Thus, both $P(v'|w')$ and $P(w'|v')$ must be estimated.

The third type is substitution of colloquial words and phrases. Similar to the first case, not all possible words are actually substituted; however, observed colloquial expressions must always be corrected in document-style texts. Therefore, $P(v'|w')$ should be estimated, while $P(w'|v')$ is set to one.

C. Training of Transformation Model

The basic word-based transformation model is directly trained using the occurrence counts of word sequences. In a training corpus, spoken- and document-style word sequences (v' and w') are annotated and aligned beforehand. Then, the statistics (occurrence counts) for document-style word sequence $N(w')$ and those for corresponding spoken-style word sequence $N(v' \leftarrow w')$ are calculated using this parallel aligned corpus. Then, word-based transformation probabilities $P_{\mathrm{word}}(v'|w')$ are estimated as follows:

$$P_{\mathrm{word}}(v'|w') = \frac{N(v' \leftarrow w')}{N(w')}. \tag{6}$$

For reliability of probabilities and computational saving, we cut-off transformation patterns (v', w') whose count $N(v' \leftarrow w')$ is smaller than or equal to θ_c, or probability $P_{\mathrm{word}}(v'|w')$ is smaller than θ_p.

However, the problem of data sparseness arises with this basic model. We introduced a transformation model based on POS tags to improve coverage and estimation accuracy. A POS tag is assigned to every contextual word by using a morphological analyzer ChaSen [19]. We use top-level categories of POS tags, such as verb, adjective and adverb, except that nouns and particles are classified into subcategories such as general noun, proper noun, and pronoun. The transformation probabilities $P_{\mathrm{POS}}(v'|w')$ for POS-based patterns, e.g., $w' =$ "*VERB PRONOUN*" and $v' =$ "*VERB ah PRONOUN*" are estimated in the same way as defined in (6). The same thresholds θ_c and θ_p are applied to counts and probabilities of POS-based patterns. Conditional probabilities $P_{\mathrm{word}}(w'|v')$ and $P_{\mathrm{POS}}(w'|v')$ are also estimated in the same manner.

D. Back-Off Scheme

The word-based and POS-based transformation models can be applied to the N-gram entries of a document-style language model using (4). A back-off scheme is used as the simplest way of combining word-based and POS-based models:

$$P(v'|w') = \begin{cases} P_{\mathrm{word}}(v'|w'), & \text{if } w' \rightarrow v' \text{ exists} \\ P_{\mathrm{POS}}(v'|w'), & \text{else if POS}(w') \rightarrow v' \text{ exists} \end{cases} \tag{7}$$

where $POS(w')$ is a pattern made from w' by changing its context words to corresponding POS tags. Each word-based pattern in this scheme such as "*think this→think ah this*" is first applied to N-gram entries in turn, and transformation is carried out with $P_{\mathrm{word}}(v'|w')$ when the pattern matches. If it does not match, a POS-based pattern such as "*VERB PRONOUN→VERB ah PRONOUN*" is applied. When the POS-based pattern matches, the POS contexts of a transformed entry are replaced with original context words such as "*think*" and "*this*" to produce a new N-gram entry. Unlike standard back-off N-gram modeling, we do not use back-off weights to $P_{\mathrm{POS}}(v'|w')$ in this scheme, because $P_{\mathrm{word}}(v'|w')$ is small in most cases and therefore back-off weights tend to be almost one.

E. Linear Interpolation Scheme

We introduce a scheme of linear interpolation of the two models as an alternative to the back-off method. The weighted sum of the word-based and POS-based probabilities is used as the transformation probability in this scheme, and transformation is conducted if either the word-based or POS-based pattern matches an original N-gram entry:

$$P(v'|w') = \lambda P_{\mathrm{word}}(v'|w') + (1-\lambda) P_{POS}(v'|w'). \tag{8}$$

F. Maximum Entropy Scheme

The word-based and POS-based models in the above schemes are first separately estimated and then combined. We introduce the maximum entropy (ME) scheme [20] to better count the lexical and POS information in a more integrated manner. A conditional probability $P(v'|w')$ is determined by

$$P(v'|w') = \frac{1}{Z} \exp\left[\sum_i \lambda_i f_i(w', v')\right] \tag{9}$$

where f_i is a feature function, λ_i is a feature weight, and Z is a normalization factor. We used the preceding and following words and their POS tags as features $\{f_i\}$ for ME. The ME model is applied to every N-gram entry of the document-style model, and a spoken-style N-gram is generated if the transformation probability is larger than a threshold.

G. Generation of N-Gram Entries and Language Model

For every input word sequence w, matching with every transformation pattern w' is performed. For all matched patterns, spoken-style word sequence v is generated from v' which corresponds to the matched pattern w'. The occurrence count of v is then calculated in real numbers, based on the transformation probabilities which are defined in either of back-off, interpolation and ME schemes as described above. Then, statistics of N-gram entries are calculated over the generated word sequences. Fractions of N-gram occurrence counts are finally rounded off, and resulting integer counts are used to train a language model in a standard manner. There are a large number of N-gram entries whose count is less than one, and these entries are discarded in this process. Although rounding errors are inevitable here, their effect is empirically insignificant, and this process leads to memory efficiency and compatibility with existing toolkits.

IV. Statistical Transformation of Pronunciation Model

Next, we will describe the transformation of the pronunciation model. The design of a pronunciation lexicon was conventionally an empirical issue. Manual editing of lexicons [21] is, however, extremely costly and not practical for LVCSR. Therefore, various frameworks of pronunciation modeling have been proposed to automatically generate a lexicon. Previous studies include the knowledge-based approach such as the application of phonological rules. However, this approach does not provide the probabilities of the rules, which are necessary to suppress false matching caused by increased numbers of entries. The data-driven approach has also been studied, e.g., pattern extraction using automatic phone recognition [22]. Most of the previous work, however, has assumed that the domain and lexicon for the training data are the same as those of the test set.

Pronunciation modeling for spoken Japanese was studied using the CSJ [5]. Faithful pronunciations of all speech materials in the CSJ have been transcribed as well as orthographic transcriptions. Thus, pronunciation variations observed in spontaneous speech can be extracted by matching these two kinds of transcriptions. Nanjo and Kawahara have already addressed pronunciation modeling using the CSJ [23], where matching was done word by word and the pronunciation probability was estimated for all possible pronunciation variants of a word. They also investigated language modeling that separately handled pronunciation variants [23]. However, these word-based approaches are obviously limited to the vocabulary observed in the CSJ, and cannot be applied to different tasks.

We investigate subword-based modeling to achieve portability to other domains. Pronunciation variation can be described as the transformation of one subword to another. Surface forms are obtained by applying such a model to subword sequences of baseforms. The decision tree [24], neural network [25] and confusion matrix [26] have been proposed as the modeling frameworks. Phones are often used as the modeling units. Although pronunciation variations depend on preceding and following contexts, most methods have not considered the context or have only counted neighboring phones. Moreover, the three methods mentioned above [24]–[26] do not necessarily yield appropriate pronunciation probabilities, because they do not permit statistical estimates with reliable and sufficient amounts of data.

We therefore propose a statistical transformation model for pronunciation modeling. We adopt probabilistic rewrite rules like Yang *et al.* [27] with a variable-length phone context as a modeling framework, which is a kind of statistical transformation model. Pronunciation variations are detected by aligning phonetic and orthographic transcriptions, and variation patterns including neighboring phone contexts are extracted. Furthermore, variation probabilities are derived from the occurrence counts of baseforms and surface forms. The appropriate length for phone contexts is determined based on the statistics, and the context back-off mechanism is introduced to enable the model to be robustly estimated and matched.

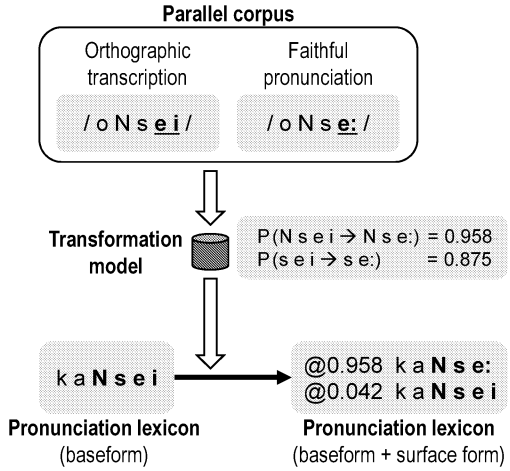

Fig. 3. Flow of pronunciation model transformation.

The proposed modeling method involves three steps. First, patterns in pronunciation variations are detected, and the necessary statistics for variation patterns and their phone contexts are estimated. Next, a set of rewrite rules is derived with appropriate contexts and probabilities. Finally, these rules are applied to baseforms to generate new pronunciation entries (surface forms).

A. Extraction of Pronunciation Variations

First, phonetic transcriptions of spoken words in the training data are compared with baseforms to detect pronunciation variations. The input text must be segmented into words to make comparisons, and each word must have a pronunciation baseform. In Japanese, morphological analysis is applied to the input text to insert word boundaries and generate baseforms. Then, baseforms and phonetic transcriptions (surface forms) are automatically aligned for each word using a dynamic programming technique, given utterance boundaries. As Japanese words often have distinct multiple baseforms, the most likely baseform is also determined by the alignment process.

As we can see from Fig. 3, if a mismatched pair of a baseform and a surface form (i.e., variant) is found, their phone sequences are identified. Each variation is extracted together with its preceding and following phone context, and the number of times it occurs is counted. We considered up to two phones in both directions as the phone context. The length of the context seemed reasonable, since at most five phones (quinphones) are used as a modeling unit in context-dependent acoustic modeling. Note that the word boundary is also considered as a context, because it provides useful information for pronunciation variations.

B. Generation of Probabilistic Rewrite Rules

Next, probabilistic rewrite rules are generated based on the statistics of variations obtained in the previous step. Let q be a certain phone or phone sequence with phone context c, and q' be a variant of q. $N(q|c)$ and $N(q \rightarrow q'|c)$ correspond to the occurrence counts of baseform q and surface form q' with context c. A threshold, θ_1, is introduced for $N(q|c)$ to determine the adequate length of context c so that the model has reliable statistics. Namely, patterns that are more frequent than θ_1 (i.e.,

TABLE II
EXAMPLES OF SPOKEN-STYLE EXPRESSIONS AND THEIR OCCURRENCE COUNTS IN CONGRESSIONAL SPEECH

(a)

Fillers:	
eh	5,623
oh	2,434
ano	2,097
End-of-sentence expressions:	
desu ne	2,728
to	1,063
Conjunctive:	
de (and)	1,078

(b)

Particles:	
wo (nominative case)	586
wa (nominative case)	535
ga (nominative case)	233
ni (objective case)	150
End-of-sentence expressions:	
iru	52
desu	34

(c)

End-of-sentence expressions:			
te iru	→	*teru*	2,209
toiu	→	*teiu*	406
keredomo	→	*kedomo*	235
Colloquial expressions:			
yahari	→	*yappari*	246
iroiro na	→	*ironna*	181

$N(q|c) \geq \theta_1$) are adopted as rules, and their probabilities are computed as

$$P(q \rightarrow q'|c) = \frac{N(q \rightarrow q'|c)}{N(q|c)}. \tag{10}$$

The contextual patterns eliminated by the threshold θ_1 are backed off to shorter-context rules.

We used at most two phones as preceding and following contexts. Let i and j be the respective lengths of the preceding and following contexts, and R_{ij} be a set of rules whose context length is i and j. Rules are defined in a descending order, from the longest context set, R_{22}, to a context-independent rule, R_{00}. These, once adopted, should be excluded from the back-off computation. For example, the adjusted frequency of variation pattern "$a\ b - q + d$," which has preceding context "$a\ b$" and following context "d," is computed as

$$N'(q|ab:d) = N(q|ab:d) - \sum_{\substack{(ab:dx) \\ \in R_{22}}} N(q|ab:dx). \tag{11}$$

The rewrite rules for variation $q \rightarrow q'$ consist of context sets $R_{ij}(0 \leq i,j \leq 2)$, and individual rule entries have their own probabilities $P(q \rightarrow q'|c)$. Finally, we also introduce a threshold, θ_2, for the probabilities, and rules that have probabilities larger than θ_2 (i.e., $P(q \rightarrow q'|c) \geq \theta_2$) are adopted. This threshold is intended to save computation at the time of application, by discarding trivial rewrite rules.

C. Application of Variation Rules

Then, new surface forms are generated by applying the set of rules to baseforms in a lexicon.[1] Rules with longer contexts are applied with higher priority, and then backed-off to shorter contexts if necessary. The probabilities of a resulting new pronunciation entry p' and the original p for a lexical entry w are updated as

$$P(p'|w) \leftarrow P(p|w) \prod_{i:p \rightarrow p'} P(q_i \rightarrow q_i'|c) \tag{12}$$

and

$$P(p|w) \leftarrow P(p|w) \prod_{i} \{1 - P(q_i \rightarrow q_i'|c)\} \tag{13}$$

[1]Using a finite-state transducer (FST) would realize a more solid implementation, but it was not adopted in this work.

where $q_i \rightarrow q_i'$ is all variation patterns applicable to the input entry p, and the initial probabilities of $P(p|w)$ are equal, i.e., one divided by the number of baseforms. The rules are applied to every possible position in the baseform, and the probability of a new entry p' is calculated by multiplying probabilities of the actually applied rules to derive p' from p [(12)]. On the other hand, the total amount of probabilities assigned to these new entries is deducted from the initial probability of the baseform entry p [(13)]. A resulting entry is discarded if its probability is smaller than a threshold θ_3, to avoid false matching with rare entries during a decoding process.

V. ANALYSIS ON CHARACTERISTICS OF SPONTANEOUS JAPANESE

This section investigates spoken and formal Japanese using parallel aligned corpora to clarify the differences between them. The comparison is made in lexical and phonetic sequences. Note that we did not analyze the inversion and disorder of phrases and sentences as these were beyond the scope of N-gram-based LVCSR.

A. Linguistic Characteristics in Spontaneous Japanese

We used transcriptions and the minutes of congressional meetings as a parallel corpus for the lexical comparison. We prepared faithful transcriptions of speech from meetings in the House of Representatives of the National Diet of Japan. Most of the speech was extracted from the budget committee in 2003, where a variety of national issues were discussed. Committee meetings in the Diet are more spontaneous than plenary sessions in parliament, which were mainly dealt in the TC-STAR project. The House provides minutes of its meetings, which were edited to meet the strict orthographic standards by professional stenographers. They are not faithful transcripts since redundant expressions such as fillers and end-of-sentence expressions are deleted and several spoken expressions are modified for purposes of documentation. They were manually compared and aligned to faithful transcriptions of the original speech, and each edit was annotated. The transcripts in this analysis contained 737 K words.

Table II lists typical differences and their occurrences in the parallel aligned corpus. The most significant phenomenon in spontaneous speech is the insertion of fillers and end-of-sentence expressions, as listed in Table II(a). A similar tendency has been reported in lectures in the CSJ [28]. As these kinds of end-of-sentence expressions and conjunctives are often treated as fillers, many such expressions are removed from minutes.

TABLE III
EXAMPLES OF PRONUNCIATION VARIATIONS EXTRACTED FROM THE CSJ

Variations	Types	Examples			
e-i → e:	Long vowel	*o N s e i*	→	*o N s e:*	(speech)
o: → o	Short vowel	*h o N t o: n i*	→	*h o N t o n i*	(really)
n-i → N	Nasal	*m a i n i ch i*	→	*m a i N ch i*	(everyday)
k-u → q	Glottal stop	*hy a k u*	→	*hy a q*	(hundred)
k → g	Voiced consonant	*k a i sh a*	→	*g a i sh a*	(company)
r →	Dropped	*s o r e*	→	*s o e*	(that)
e-r-e → e:	Dropped, long vowel	*k e r e d o m o*	→	*k e: d o m o*	(though)
i → u	Foreign word	*e k i s u p o*	→	*e k u s u p o*	(expo)
u → i	Other	*sh u ts u j o*	→	*sh i ts u j o*	(join)

The deletions and substitutions in Table II(b) and (c) occur fewer times than insertions, but have typical patterns. Most deletions are postpositional particles, which indicate the relationship between words or phrases, such as the linguistic case structures. Note that not all postpositional particles are deleted, e.g., those indicating the nominative case (such as *wa* and *ga*) are often omitted while those indicating the possessive case are rarely dropped. Lexical substitutions with colloquial expressions are often used for short smooth utterances, just as the contracted forms of "can't" and "ain't" are often used in English conversation. A large number of such substitutions appear at the ends of sentences, which corresponds to verbs and auxiliary verbs in Japanese.

TABLE IV
EXAMPLES OF REWRITE RULES

Variation pattern	Context		Probability
	preceding	following	
e i → e:	#-t	r-i	0.9647
	#-t	t	0.8077
	—	r-i	0.6531
k-u → q	g-a	k-a	0.5385
	a	k	0.1818
	—	k	0.1549
a-w-a → a:	#-m	r-i	0.2770
	#-g	#	0.1408
	a-z	—	0.4286

"#" denotes word boundary.

B. Pronunciation Variations in Spontaneous Japanese

We used the CSJ for phonetic comparison, which has faithful transcriptions of speech from lectures. We generated an orthographic pronunciation of speech using a morphological dictionary, and then compared it against the faithfully transcribed pronunciation. A total of 630 K words was used for this analysis.

Typical variations in the CSJ are listed in Table III. One of the major differences occurs in vowels, i.e., a short vowel or a diphthong becomes a long vowel, and a long vowel becomes a short vowel. Japanese syllables typically consist of a consonant followed by a vowel; therefore, vowels surrounded by consonants, typically /k/, sometimes vanish (glottal stop). Most of these variations cause us to speak faster. Unvoiced consonants are often voiced in compound words of nouns. However, some consonants are also dropped to simplify pronunciation in limited phone contexts.

Some examples of derived rewrite rules are listed in Table IV. The derived rule set includes typical cases of pronunciation variations that are phonologically predictable, e.g., "/e i/ → /e:/" (diphthong to long vowel) and "/k u/ → /q/" (vanishing vowel). However, our results attach appropriate probabilities to these. Moreover, a number of variants that are characteristic to spontaneous speech and cannot be predicted by using phonology were also found.

VI. EXPERIMENTAL EVALUATIONS

We carried out experiments on real congressional speech to evaluate the proposed transformation schemes of language and pronunciation models. First, we preliminarily examined three modeling schemes for transforming the language model, and they were then compared with conventional methods. Finally, we transformed the pronunciation model and evaluated it.

A. Comparison of Methods of Transforming Language Model

We preliminarily evaluated how effective the back-off, linear-interpolation, and ME-based methods were for transforming the language model. We collected archives of the minutes of the National Diet over four years (1999–2002) for this preliminary experiment and used these to train the baseline trigram language model. They contained 71 M words in total. We used a parallel corpus of the transcripts and minutes of meetings described in Section V-A to train the transformation model. Part of the corpus (63 K words) was used as a test set. As there was a total of 23 distinct speakers in the test set, various spontaneous expressions were observed. The rest of the corpus of 666 K words was used for training. The corpus was too small to directly train a language model. Throughout experiments, cutoff thresholds for the training of transformation model were set as $\theta_c = 1$ and $\theta_p = 0.001$.

We prepared a language model ("+CSJ") for comparison by linearly interpolating the baseline model with extemporaneous public speeches in the CSJ (2.9 M words) that were not matched to the task domain. We also prepared another mixture model ("+Transcript") by interpolating the baseline model and a model trained with the faithful transcripts of the parallel corpus. Compared to the proposed modeling expressed by (4), the conventional linear interpolation of language models is formulated as

$$N_{\mathrm{LM}}(v) = \lambda' N_{\mathrm{LM}_W}(v) + (1-\lambda') N_{\mathrm{LM}_V}(v) \qquad (14)$$

where N_{LM_W} and N_{LM_V} are occurrence counts in W and V, respectively, and λ' is an interpolation weight, which must be

TABLE V
PERPLEXITY FOR BUDGET-COMMITTEE-MEETING TEST SET

Model		PP	OOV rate	# of Bigrams	# of Trigrams
Baseline		80.0	3.74%	0.99M	3.63M
Mixture-based	+CSJ	75.9	0.48%	1.05M	3.76M
	+Transcript	54.2	0.44%	1.00M	3.65M
Transformed (Proposed)	Back-off	50.0	0.48%	1.06M	3.99M
	Linear	52.2		1.03M	3.88M
	ME	51.8		1.08M	3.96M

tuned depending on characteristics of W, V and the target domain. For the reference models in this experiment, the best interpolation weight was separately chosen based on the perplexity over the test set.[2]

The perplexity (PP), out-of-vocabulary (OOV) rates and numbers of bigram and trigram entries for the two mixture models and the three transformed models are listed in Table V. The combination with CSJ ("+CSJ") improved the perplexity and the OOV rate, however, the improvement against the baseline was smaller, because it could not provide sufficient N-gram entries for this task. In contrast, the combination with transcripts of the parallel corpus ("+Transcript") considerably improved perplexity. These transcripts matched the task, and N-gram entries in the transcripts well covered spoken expressions in the test set, although the increase of N-gram entries was small.

All transformed models further reduced perplexity. Among these models, the back-off model achieved smallest perplexity. The reduction by this model over the baseline model was 37.5% and that over the conventional "+CSJ" model was 34.1%. The reduction over the "+Transcript" model was 7.7%. It was confirmed that the proposed transformation generated more N-gram entries than the mixture-based models, and it led to these improvements. We also confirmed that most of these N-gram entries were generated by the POS-based transformation. When POS-based transformation was omitted and word-based transformation was solely applied, only 3.70 M trigram entries were obtained, and perplexity by this model was 75.2.

When we compare the three transformed models, there were fewer trigram entries in the linear interpolation model than in the other two transformed models, because the effect of POS-based transformation was reduced by the interpolation. The ME model generated more N-gram entries than the back-off model; however, the perplexity by the former was larger. Because the set of features for the ME model seemed too large, the parameters could not be estimated appropriately. Based on the results, we adopted the back-off scheme for transforming the language model.

B. Transformation of Language Model in Different Topic Domains

We applied the proposed transformation model to different topic domains in the National Diet to evaluate the generality. We prepared new test sets for various committee meetings held in 2006, including those of the budget committee, as listed in Table VI. There were a total of five meetings and the test-set transcripts contained 144 K words in total.

[2]We did not prepare held-out data for tuning, but this would give a positive bias to the reference models.

TABLE VI
SPECIFICATIONS OF 2006 CONGRESSIONAL-SPEECH TEST SET

Meetings	# of Words	OOV rate
Budget	48,966	0.13%
Question time	7,884	0.05%
National security	46,486	0.17%
Foreign affairs	40,416	0.18%
Sum/Average	143,752	0.15%

TABLE VII
SPECIFICATIONS OF LANGUAGE MODELS

Model	Baseline	Interpolation with CSJ ("+CSJ")	Interpolation with transcripts ("+Transcripts")	Transformed ("Proposed")
Training corpus	Minutes	Minutes +CSJ	Minutes +transcripts	Minutes +parallel
# of Words	119M	119M + 7.3M	119M + 0.7M	119M +0.7M
Vocab. size	54,301	54,321		
# of Bigrams	1.37M	1.78M	1.40M	1.46M
# of Trigrams	5.32M	7.29M	5.45M	5.86M

The transformation model was retrained for this experiment using the entire parallel corpus (737 K) explained in Section V-A. We obtained 6339 transformation patterns, and 2310 (36%) of these were POS-based. We also enhanced the baseline language model used in the previous experiment by adding the minutes for year 2003 to 2005 to the training texts. Then, conventional mixture models were created by combining it with the transcripts in the CSJ or the parallel corpus. Best mixture weights were preliminarily chosen for these models using perplexity on the test set. The transformed model was generated based on the back-off scheme. The specifications for all models are listed in Table VII. These models except the baseline model have the same vocabulary of 54 321 items. The difference between this vocabulary and the baseline vocabulary was mostly fillers that were found only in the transcripts of the parallel aligned corpus. The transformation added these words to the baseline vocabulary, and resulting vocabulary was also applied to the mixture models. The average OOV rate by this vocabulary over the test set is 0.15%, as shown in Table VI. Note that the baseline model was only used to mix with the other models and not to directly evaluate, as it was obviously inappropriate for speech recognition. The acoustic model in this experiment was newly trained using 140 hours of speech from meetings in the Diet, with applying vocal tract length normalization (VTLN) [29], [30] and minimum phone error (MPE) training [31].

As we can see from Fig. 4, perplexity was drastically reduced by the "Proposed" model over the "+CSJ" and "+Transcripts" models, in all meetings. The average perplexity by "+CSJ," "+Transcripts," and "Proposed" models corresponded to 52.0, 49.1, and 41.2. Reduction in perplexity by the transformed model was 20.8% against the "+CSJ" model, and 16.1% against the "+Transcripts" model. Although the "+CSJ" model had many more trigram entries (7.29 M) than the other two models, its perplexity was larger. This suggests that a simple

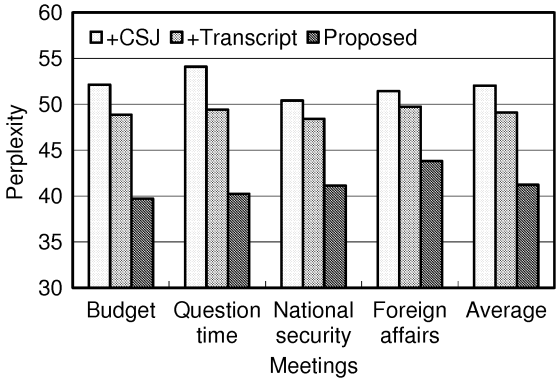

Fig. 4. Perplexity by language models.

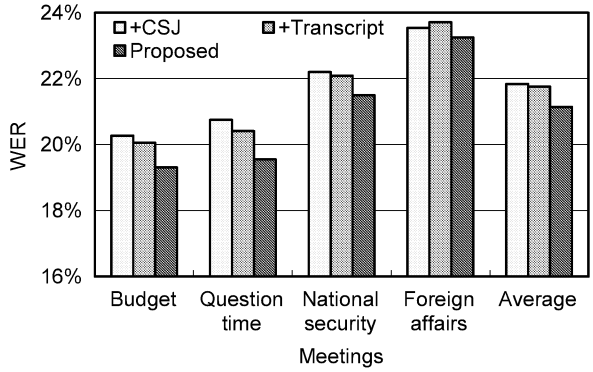

Fig. 5. Word error rates by language models.

mixture-based language model does not sufficiently cover effective N-gram entries. The "+Transcripts" model reduced perplexity over the "+CSJ" model; however, the reduction was relatively smaller. This is because mixture with the transcript yielded a limited number of trigram entries. This result demonstrated efficient and effective prediction could be accomplished with the proposed transformation.

The reduction in perplexity with the proposed model against the "+Transcripts" model was larger in this experiment than that obtained in the previous experiment (6.9%) where the training and test-set data were taken from the same budget committee, and the date of the meetings were close. Since the topics in both data were similar, the mixture-based "+Transcripts" model could accomplish high coverage and small perplexity. In contrast, the test data in this experiment were chosen from a different year, and thus the topics were also different, even in the same budget committee. While the "+Transcripts" model could not greatly reduce perplexity, the proposed model accomplished significant perplexity reduction. This means that the proposed approach is general and even more effective in different topic domains.

Fig. 5 is a bar graph of the word error rates (WER) for the three language models. Here, we used a pronunciation lexicon without the surface forms provided by our method that were described in Section IV. The WERs obtained by the "+CSJ" and "+Transcripts" models were comparable, whereas the proposed "Transformed" model significantly reduced the WER. The reduced WER by using the "Transformed" model over the "+CSJ" and "+Transcript" models corresponded to 3.2% and 2.8% relative, and these were statistically significant at $p < 0.01$. The proposed method particularly reduced errors around the fillers, which might be inserted at many points. The mixture-based approach achieved limited error reduction around fillers, since it could only provide N-gram entries observed in the training corpus. Actually, we conducted an evaluation in which fillers were removed from automatic speech recognition (ASR) results and reference texts before comparison. Here, we obtained WERs of 21.6%, 21.5% and 20.9% by "+CSJ," "+Transcripts" and "Transformed" models, respectively. Improvements by the proposed transformation over "+CSJ" and "+Transcripts" models were 3.1% and 2.6% relative, which were still statistically significant at $p < 0.01$. This result demonstrates the advantage of the proposed method, i.e., it could predict unseen N-gram entries in the training corpus.

C. Effect of Training Data Size

In the proposed transformation, the size of training data might affect the performance of the transformed language model. To investigate this, we made another experiment using transformation models that were trained with a half, a quarter and one-eighth portions of the parallel aligned corpus. Average perplexity over the test set by language models transformed with these models were 42.0, 43.1, and 44.4, respectively, while the full-size model achieved 41.2 as mentioned above. This result suggests that the proposed method properly works even if the size of training data is one eighth, i.e., smaller than 100 K words, and the performance was almost saturated with the 737 K corpus.

D. Transformation of Pronunciation Model Using the CSJ

Finally, we evaluated the proposed transformation scheme of the pronunciation model. We used the CSJ in Section V-B to train the transformation model. We used all lectures in the CSJ (6.3 M words). Thresholds θ_1, θ_2, and θ_3 were determined as $\theta_1 = 20$ and $\theta_2 = \theta_3 = 0.1$, based on our previous experiments on discussion tasks [32]. As a result, 265 kinds of variation patterns and 1381 rules were obtained. By applying these rules, the lexicon of 57 462 baseform entries was expanded to 64 857 entries by adding 7395 surface forms.

Fig. 6 is a bar chart of the WER on the same test set by using the baseline pronunciation lexicon and the transformed lexicon, which contains surface forms generated by the rewrite rules. The "Transformed" language model in Table VII was used in this experiment. Reduced WER was achieved in all meetings, and the average reduction was relatively 4.6%. This reduction is statistically significant at $p < 0.01$. The training data used for the transformation model was the CSJ, which had a completely different domain from the Diet task. Therefore, the results demonstrated that the proposed framework could model spontaneous characteristics separately from the domain characteristics of the training data, and achieved portability to other task domains.

The average WER by transforming the language and pronunciation models was 20.2%. The improvement in both transformations over the "+CSJ" language model and the baseline pronunciation lexicon was relatively 7.6%, which is almost the

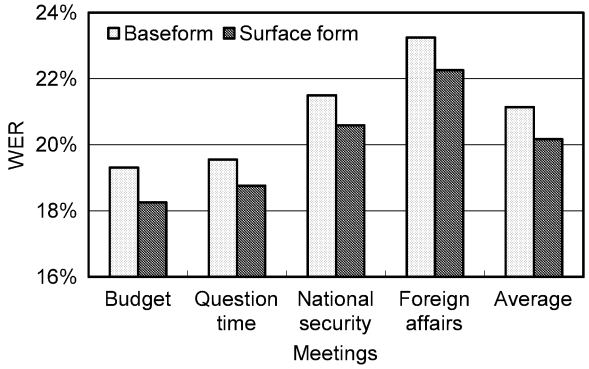

Fig. 6. Word error rates by pronunciation models.

same as the sum of individual improvement (3.2% and 4.6%). This means that two transformations could work synergistically, and effectively improved LVCSR.

VII. Conclusion

We proposed a novel approach of statistical transformation that efficiently generated a language and a pronunciation model for spontaneous speech recognition. The transformation model for the language model contained context-dependent probabilistic patterns of transformation from document-style to spontaneous speech. These patterns and their probabilities were determined based on occurrence statistics in a parallel corpus of faithful transcripts and their corresponding document-style texts. Since these training data were small, contexts were backed-off to the POS-level, which provided more robust prediction. The transformation model was applied to the N-gram counts of a document-style language model, and N-gram entries for spontaneous speech were generated with the estimated number of occurrences.

The transformation model for a pronunciation model consisted of probabilistic rewrite rules with phone contexts of variable lengths. The pronunciation variations between baseform and surface forms were extracted from a large-scale spontaneous speech corpus (CSJ), then phone context-dependent variation patterns and their occurrence probabilities were trained. Since the probabilistic model was generalized, it can be applied to any lexicon of new domains to generate appropriate surface forms with their probabilities.

In experimental evaluations on real congressional speech, the proposed method efficiently and effectively generated a language model and a pronunciation model, and reduced both perplexity and WER. The transformed language model was tested on meetings whose topics and times were different from those in the training data, and it accomplished reduced WER without side effects of increasing vocabulary. The result demonstrated the generality of the proposed transformation of language model for new domains. As for pronunciation model, the transformation model was trained with the CSJ, and tested on the Diet corpus. Although the two corpora were completely different, the generated pronunciation model reduced WER. This clearly demonstrated the portability of our approach, and thus transformation is expected to be applied to any spontaneous speech recognition tasks.

References

[1] G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, and H. Gish, "The BBN Byblos 1997 large vocabulary conversational speech recognition system," in *Proc. ICASSP*, 1998, pp. 905–908.

[2] T. Hain, P. Woodland, T. Niesler, and E. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," in *Proc. ICASSP*, 1999, pp. 57–60.

[3] J. Garofolo, C. Laprun, and J. Fiscus, "The rich transcription 2004 spring meeting recognition evaluation," in *Proc. ICASSP Meeting Recognition Workshop*, 2004.

[4] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. ASRU*, 2007, pp. 238–247.

[5] S. Furui, K. Maekawa, and H. Isahara, "Toward the realization of spontaneous speech recognition—Introduction of a Japanese priority program and preliminary results," in *Proc. ICSLP*, 2000, pp. 518–521.

[6] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED corpus lectures," in *Proc. ICASSP*, 2003, pp. 232–235.

[7] L. Lamel, G. Adda, E. Bilinski, and J. Gauvain, "Transcribing lectures and seminars," in *Proc. Eurospeech*, 2005, pp. 1657–1660.

[8] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. Eurospeech*, 2007, pp. 2553–2556.

[9] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in *Proc. ICASSP*, 2008, pp. 4929–4932.

[10] R. Prasad, L. Nguyen, R. Schwartz, and J. Makhoul, "Automatic transcription of courtroom speech," in *Proc. ICSLP*, 2002, pp. 1745–1748.

[11] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, "The LIMSI 2006 TC-STAR EPPS transcription systems," in *Proc. ICASSP*, 2007, vol. 4, pp. 997–1000.

[12] J. Loof, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter, and H. Ney, "The 2006 RWTH parliamentary speeches transcription system," in *Proc. ICSLP*, 2006, pp. 105–108.

[13] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro, "The IBM 2006 speech transcription system for European parliamentary speeches," in *Proc. ICSLP*, 2006, pp. 1225–1228.

[14] M. Cettolo, F. Brugnara, and M. Federico, "Advances in the automatic transcription of lectures," in *Proc. ICASSP*, 2004, pp. 769–772.

[15] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.

[16] H. Schramm, X. Aubert, C. Meyer, and J. Peters, "Filled-pause modeling for medical transcriptions," in *Proc. Workshop Spontaneous Speech Process. Recognition*, 2003, pp. 143–146.

[17] S. Petrik and G. Kubin, "Reconstructing medical dictations from automatically recognized and non-literal transcripts with phonetic similarity matching," in *Proc. ICASSP*, 2007, vol. 4, pp. 1125–1128.

[18] T. Hori, D. Willett, and Y. Minami, "Language model adaptation using WFST-based speaking-style translation," in *Proc. ICASSP*, 2003, vol. 1, pp. 228–231.

[19] M. Asahara and Y. Matsumoto, "Extended models and tools for high-performance part-of-speech tagger," in *Proc. COLING*, 2000, pp. 21–27.

[20] A. Berger, V. Della Pietra, and S. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.

[21] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary, continuous speech recognition," in *Proc. ICSLP*, 1996, pp. 6–9.

[22] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP*, 1996, pp. 2328–2331.

[23] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 391–400, Jul. 2004.

[24] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Commun.*, vol. 29, pp. 209–224, 1999.

[25] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks," *Speech Communication*, vol. 27, pp. 63–73, 1999.

[26] D. Torre, L. Villarrubia, J. Elvira, and L. Hernandez-Gomez, "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," in *Proc. ICASSP*, 1997, pp. 1463–1466.
[27] Q. Yang, J.-P. Martens, P.-J. Ghesquiere, and D. Compernolle, "Pronunciation variation modeling for ASR: Large improvements are possible but small ones are likely to achieve," in *Proc. ICSLP Workshop Pronunciation Modeling Lexicon Adaptation for Spoken Lang. Technol.*, 2002, pp. 123–128.
[28] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. Workshop Spontaneous Speech Process. Recognition*, 2003, pp. 7–12.
[29] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, 1996, vol. 1, pp. 346–349.
[30] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, 1996, vol. 1, pp. 353–356.
[31] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, vol. 1, pp. 105–108.
[32] Y. Akita and T. Kawahara, "Generalized statistical modeling of pronunciation variations using variable-length phone context," in *Proc. ICASSP*, 2005, vol. 1, pp. 689–692.

Yuya Akita (M'06) received the B.E., M.Sc., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 2000, 2002 and 2005, respectively.

Since 2005, he has been an Assistant Professor in the Academic Center for Computing and Media Studies, Kyoto University. His research interests include spontaneous speech recognition and spoken language processing.

Dr. Akita received the 2007 Awaya Memorial Award from the Acoustical Society of Japan.

Tatsuya Kawahara (M'91–SM'08) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively.

In 1990, he became a Research Associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor in the Academic Center for Computing and Media Studies and an Adjunct Professor in the School of Informatics, Kyoto University. He was also an Invited Researcher formerly at ATR and is currently at National Institute of Information and Communications Technology. He has published more than 150 technical papers covering speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of the IEEE Signal Processing Society Speech Technical Committee. He was a general chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU-2007).