Estimating Speech Recognition Accuracy Based on Error Type Classification

Atsunori Ogawa, *Member, IEEE*, Takaaki Hori, *Senior Member, IEEE*, and Atsushi Nakamura, *Senior Member, IEEE*

***Abstract*—Methods for estimating the speech recognition accuracy without using manually transcribed references are beneficial to the research and development of automatic speech recognition technology. This paper proposes recognition accuracy estimation methods based on error type classification (ETC). ETC is an extension of confidence estimation. In ETC, each word in the recognition results (recognized word sequences) for the target speech data is probabilistically classified into three categories: the correct recognition (C), substitution error (S), and insertion error (I). Deletion errors (D) that can occur at interword positions in the recognition results are also probabilistically detected. By summing these CSID probabilities individually, the numbers of CSIDs and, as a result, the two standard recognition accuracy measures, i.e., the percent correct and word accuracy (WAcc), for the speech data can be estimated without using the reference transcriptions. Two recognition accuracy estimation methods based on ETC are proposed. In the first easy-to-use method, ETC is performed by converting the recognition results represented as word confusion networks into word alignment networks (WANs). In the second and more accurate method, the WAN-based ETC results are refined with conditional random fields (CRFs) using various types of additional features extracted for each of the recognized words. Experiments using English and Japanese lecture speech corpora show that the recognition accuracy can be accurately estimated with the CRF-based method. The correlation coefficient and root mean square error between the lecture-level true WAccs calculated using the reference transcriptions and those estimated with the CRF-based method are 0.97 and lower than 2%, respectively. A series of additional experiments and analyses are also conducted to better understand the effectiveness of the CRF-based method.**

***Index Terms*—Automatic speech recognition, conditional random fields, error type classification, recognition accuracy estimation, word alignment network.**

I. Introduction

Recently, great progress has been made on automatic speech recognition (ASR) technology and various types of ASR-based applications, e.g. voice search services and speech-to-speech translation systems, have been actively developed [1]. Despite this great progress, ASR performance, e.g. accuracy and speed, varies greatly when there are changes in factors such as speakers and noisy environments [2]. Therefore, if we are to develop an ASR-based application, it would be useful to know in advance the level of ASR performance that can be obtained under the target conditions [3]–[7].

One of the most important ASR performance measures is the *recognition accuracy*. In continuous speech recognition, each word in the recognition results (recognized word sequences) for the target speech data is classified into one of three categories: the *correct recognition* (*C*), *substitution error* (*S*) or *insertion error* (*I*). Hereafter, this classification is referred to as *CSI classification*. *Deletion errors* (*D*) that occur at inter-word positions in the recognition results are also detected. Hereafter, this detection is referred to as *D detection*. CSI classification and D detection are performed by making *word alignments* between the recognition results and their corresponding *reference transcriptions* with the dynamic programming procedure using a scoring tool (e.g. NIST SCLITE scoring package [8]). Then, by counting the numbers of CSIDs individually, the two standard *recognition accuracy* measures, i.e. the *percent correct* (*%Cor*) and the *word accuracy* (*WAcc*), for the target speech data are calculated as $\%\mathrm{Cor} = (\#\mathrm{C}/\#\mathrm{N}) \times 100$ [%] and $\mathrm{WAcc} = ((\#\mathrm{C} - \#\mathrm{I})/\#\mathrm{N}) \times 100$ [%] (the latter is now widely used), where #N is the true number of words in the recognition results given by $\#\mathrm{N} = \#\mathrm{C} + \#\mathrm{S} + \#\mathrm{D}$ [8]–[10].

To calculate the recognition accuracy for the target speech data, we have to prepare the reference transcriptions of the data. However, the cost of manual transcription is very high. Therefore, methods are desired for *estimating speech recognition accuracy without using reference transcriptions*. They are beneficial to the research and development of ASR technology. For example, when developing a practical application, we can efficiently judge whether or not the ASR technology can be used in the application by estimating the recognition accuracy for the speech data recorded assuming the application. In the research phase, from a large amount of speech data with no transcriptions, we can extract a small subset of the data whose estimated recognition accuracy is low. Then, using this subset (and if necessary, transcribing it), we can efficiently analyze the reasons for the low recognition accuracy and utilize the analysis results to improve the ASR technology.

Several methods have been proposed for estimating speech recognition accuracy without using reference transcriptions [3]–[7]. They share the following basic characteristics:

1) They focus on one or a few factors that affect the recognition accuracy, e.g. the additive noise [3]–[6], room reverberation [4], [7], and task complexity [6].

Manuscript received October 8, 2015; revised March 6, 2016, April 28, 2016, and August 10, 2016; accepted August 10, 2016. Date of publication August 26, 2016; date of current version September 29, 2016. This work was supported in part by JSPS KAKENHI under Grant 26280063. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karen Livescu.

A. Ogawa is with NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: ogawa.atsunori@lab.ntt.co.jp).

T. Hori is with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 USA (e-mail: thori@merl.com).

A. Nakamura is with the Graduate School of Natural Sciences, Nagoya City University, Aichi 467-8501, Japan (e-mail: atsushi@nsc.nagoya-cu.ac.jp).

Digital Object Identifier 10.1109/TASLP.2016.2603599

2) They find or create measures (i.e. features) that capture the changes in the factors of interest, e.g. the perceptual evaluation of speech quality (PESQ) [4]–[6], D value [7], sub-band dynamic range [3], and square mean root (SMR)-perplexity [6].
3) They design functions that convert the measured features to the recognition accuracy, e.g. WAcc $= f(x_1, x_2, \ldots)$ [%], where $x_1, x_2, \ldots$ are the features.

They do not need to perform ASR for the target speech data. They can accurately capture the changes in the recognition accuracy caused by the changes in the factors of interest. However, they have the following disadvantages:

1) There are various factors that affect the recognition accuracy under real conditions. Therefore, it is difficult for the above methods to capture the changes in the recognition accuracy caused by the changes in any factors other than the one or few factors being focused on. These methods have only been evaluated under artificially controlled conditions where the factors of interest were changed while the other factors were kept largely unchanged.
2) There are certain difficulties involved in obtaining the values of several features. For example, to obtain a PESQ score [4]–[6], a clean speech sample and its corrupted version in the target noisy condition are required. To obtain a D value [7], an impulse response in the target room is required. The task complexity of the target speech data is unknown and the SMR-perplexity [6] for the data cannot be obtained in advance.
3) They do not estimate the numbers of CSIDs. Therefore, they cannot characterize the recognition results. For example, even if they estimate that the accuracy of the recognition results is low, they cannot explain which type of error (i.e. S, I or D) is the main cause of this low accuracy. In addition, to estimate both %Cor and WAcc, they have to provide two individual functions for estimating %Cor and WAcc, respectively.

In contrast to the conventional methods described above, in this paper, we propose *recognition accuracy estimation methods based on error type classification* (*ETC*) [11], [12] (see Section II for details). ETC is an extension of confidence estimation [13]–[18]. In ETC, CSI classification and D detection are performed probabilistically without using the reference transcriptions, and the CSID probabilities for each of the words and inter-word positions in the recognition results are obtained. Two recognition accuracy estimation methods based on ETC are proposed. In the first easy-to-use method, ETC is performed by converting the recognition results represented as word confusion networks into *word alignment networks* (*WANs*). In the second and more accurate method, the WAN-based ETC results are refined with *conditional random fields* (*CRFs*) [19], [20] using various types of additional features extracted for each of the recognized words. Compared with the conventional methods, the proposed methods have the following characteristics:

1) They do not specify the factors that affect the recognition accuracy.
2) They (especially, the CRF-based method) utilize various types of features extracted for each word in the recognition results, e.g. the recognized word (itself), part-of-speech, acoustic log likelihood, language log probability, and confidence score.
3) They follow the formal procedure for calculating the recognition accuracy. They obtain the CSID probabilities word-by-word as a result of ETC. By summing these CSID probabilities individually, they estimate the numbers of CSIDs. Then, by using these numbers, they estimate the recognition accuracy.

They first have to perform ASR for the target speech data. However, compared with the conventional methods, they have the following advantages:

1) By utilizing various types of features, they can capture the changes in the recognition accuracy caused by the changes in various factors under real conditions.
2) The features they utilize can be easily extracted mainly from the recognition results and additionally from the models used for ASR, e.g. the acoustic, language and lexical models.
3) They estimate the numbers of CSIDs. Therefore, they can characterize the recognition results as, for example, those that have many/few S/I/D errors, by using the estimated numbers of SIDs. One of the applications of this characterization is to estimate the causes of recognition errors [21]. For example, assuming an ASR-based service through cellphones, if recognition results have many I errors, the user may be speaking in a noisy environment and the speech recognizer may misrecognize the noise segments in the user's utterances as inserted words. Conversely, if recognition results have many D errors, packet loss in the network may be occurring frequently and the user's utterances may be reaching the speech recognizer intermittently. The intermittent speech signals can cause D errors. Another application of the characterization is that the precision (recall) of a spoken document retrieval system can be improved by selecting recognition results with few I (D) errors [22]. In addition, by following the formal calculation procedure, they can easily estimate both %Cor and WAcc (and also other measures, e.g. S/I/D error rates [8]).

Experiments are conducted using two lecture speech corpora (Section III). The WAN-based method overestimates the recognition accuracy, whereas the CRF-based method estimates it accurately. The correlation coefficient and root mean square error between the lecture-level true WAccs calculated using the reference transcriptions and those estimated with the CRF-based method are 0.97 and lower than 2, respectively.

In our previous papers [11], [12], we have already proposed two recognition accuracy estimation methods, in particular the WAN- and CRF-based methods, and confirmed the superiority of the CRF-based method over the WAN-based method experimentally (see Sections II, III-B and III-C in this paper). This paper includes new investigations that we have made since our previous papers. The main points can be summarized as follows:

1) In addition to the English lecture speech corpus we used in our previous papers, we use a Japanese lecture speech

corpus that has different characteristics from the English corpus (Section III-A). Using these two corpora, we conduct basic experiments to evaluate the proposed recognition accuracy estimation methods. We also conduct a series of additional experiments and analyses to gain a better understanding of the proposed methods and to confirm their effectiveness more consistently.

2) Borrowing the framework of the conventional recognition accuracy estimation methods [3]–[7], we propose applying a linear regression to estimation results obtained with the WAN-based method (Section II-A). Experimental results show that, with the linear regression-based method, the overestimated WAN-based lecture-level WAccs can be accurately corrected (Section III-C). However, it is shown that it is difficult to correct the WAN-based word-level ETC results (Section III-E).

3) We conduct experiments to better understand the CRF-based method. Experimental results show that, with the CRF-based method, we cannot obtain good estimation performance when we use only a small number of features even though the features are important and we should use various types of additional features along with the important ones to obtain good performance (Section III-F). It is also shown that the CRF-based method has is very robust to CRF training data size. The training data size can be greatly reduced while maintaining good performance (Section III-G).

II. Recognition Accuracy Estimation Based on ETC

Error type classification (*ETC*) is an extension of confidence estimation [13]–[18]. In confidence estimation, each word in the recognition results for the target speech data is probabilistically classified into two categories: the correct recognition (C) and incorrect recognition. In ETC, the incorrect recognition category is further divided into two categories: the substitution error (S) and insertion error (I), and each recognized word is probabilistically classified into three categories: C, S and I (CSI classification). In addition, deletion errors (D) that can occur at inter-word positions in the recognition results are probabilistically detected (D detection). In the following, we describe two recognition accuracy estimation methods based on ETC. In the first easy-to-use method, ETC is performed with the WANs. In the second and more accurate method, the WAN-based ETC results are refined with the CRFs.

A. WAN-Based Method

In this paper, we use a speech recognizer that can provide word confusion networks (WCNs) as the recognition results for the target speech data. We estimate the recognition accuracy of the 1-best recognition results extracted from the WCNs. A WCN is a compact representation of multiple recognition hypotheses for an utterance in the target speech data. It can be obtained by converting a recognition lattice with consensus decoding [23]. An example of a WCN is shown in the first row of Fig. 1. A recognized word (or a null word) is represented as an arc, and all competing recognized words in a segment are represented as arcs that share the same start and end nodes (e.g. the words w_3^1, w_3^2, w_3^3 and the null word ε_3^4 are competing in s_3, i.e. Segment 3). Each competing word in a segment has a posterior probability ($P(w_3^1|s_3) = 0.4$, $P(w_3^2|s_3) = 0.3$, $P(w_3^3|s_3) = 0.2$ and $P(\varepsilon_3^4|s_3) = 0.1$) and their total is 1 ($P(w_3^1|s_3) + P(w_3^2|s_3) + P(w_3^3|s_3) + P(\varepsilon_3^4|s_3) = 1$).

A *word alignment network* (*WAN*), as shown in the second row of Fig. 1, can be converted from a WCN. The conversion procedure from a WCN to a WAN is divided into two parts. One is performed on the WCN segments that have an *actual* (not a null) word as the 1-best recognized word (Segments 1, 3, 4 and 7). We refer to these segments as the *actual word segments*. The other is performed on the WCN segments that have a null word (ε) as the 1-best recognized word (Segments 2, 5, 6 and 8). We refer to these segments as the ε *segments*.

From an actual word segment of a WCN (e.g. Segment 3), the actual word that has the highest posterior probability (the 1-best recognized word w_3^1) is selected as a word in the 1-best recognition result (shown in the third row of Fig. 1). The WAN arc that represents Correct recognition (arc "C") and its posterior probability (*correct recognition probability*$P(\mathrm{C}|s_3)$) are obtained by copying the WCN arc of the 1-best recognized word (w_3^1) and its posterior probability ($P(\mathrm{C}|s_3) = P(w_3^1|s_3) = 0.4$). This correct recognition probability ($P(\mathrm{C}|s_3)$) can be used as a *confidence measure* [23] that scores the reliability of the 1-best recognized word (or the segment). If the 1-best recognized word (w_3^1) is incorrect, it may be a substitution error. The WAN arc that represents Substitution error (arc "S") and its posterior probability (*substitution error probability*$P(\mathrm{S}|s_3)$) are obtained by merging the WCN arcs of competing actual words (w_3^2 and w_3^3) and by summing their posterior probabilities ($P(\mathrm{S}|s_3) = P(w_3^2|s_3) + P(w_3^3|s_3) = 0.5$). If this substitution error probability is high, a substitution error may occur in the segment. If the 1-best recognized word (w_3^1) is incorrect and the null word (ε_3^4) is correct, the 1-best word is an insertion error. The WAN arc that represents Insertion error (arc "I") and its posterior probability (*insertion error probability*$P(\mathrm{I}|s_3)$) are obtained by copying the WCN arc of the null word (ε_3^4) and its posterior probability ($P(\mathrm{I}|s_3) = P(\varepsilon_3^4|s_3) = 0.1$). If this insertion error probability is high, an insertion error may occur in the segment.

From an ε segment of a WCN (e.g. Segment 2), no recognized word (null word ε_2^1) is selected for the 1-best recognition result. If this null word (ε_2^1) is incorrect, a deletion error occurs in the segment. The WAN arc that represents Deletion error (arc "D") and its posterior probability (*deletion error probability*$P(\mathrm{D}|s_2)$) are obtained by merging the WCN arcs of the competing actual words (w_2^2 and w_2^3) and by summing their posterior probabilities ($P(\mathrm{D}|s_2) = P(w_2^2|s_2) + P(w_2^3|s_2) = 0.6$). If this deletion error probability is high, a deletion error may occur in the segment. As for the WCN arc of the null word (ε_2^1), it is copied to the WAN without any conversion.

As a result of the above conversion procedure, we can obtain a WAN with the correct recognition (C), substitution error (S), insertion error (I) and deletion error (D) probabilities (*CSID probabilities*) as shown in the second row of Fig. 1. A WAN

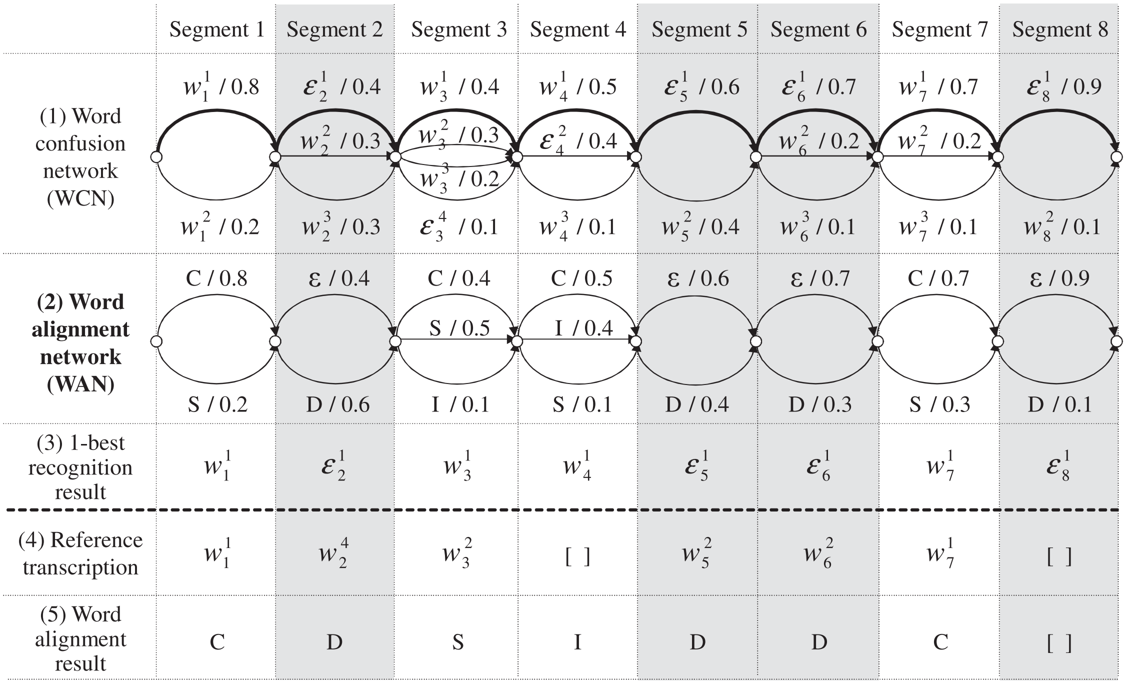


Fig. 1. Conversion from (1) a word confusion network (WCN) [23] to (2) a word alignment network (WAN). Segments 1, 3, 4 and 7 are the actual word segments and Segments 2, 5, 6 and 8 are the ε segments. The best path on the WCN is shown with the bold curved line and this is (3) the 1-best recognition result. If (4) the reference transcription is given, using a scoring tool (e.g. SCLITE [8]), (5) the (true) word alignment result is obtained. In this figure, w_j^n (ε_j^n) denotes the nth best word (null word) in the jth segment.

can be viewed as a *probabilistic word alignment* result obtained *without using* the reference transcription, i.e. the WAN itself is an ETC result.

If the reference transcription is given (shown in the fourth row of Fig. 1) for the 1-best recognition result, using a scoring tool (e.g. SCLITE [8]), the (*true*) word alignment result (fifth row) is obtained. By obtaining these word alignment results for all of the utterances in the target speech data and by following the formal calculation procedure, we obtain the *true* recognition accuracy for the data.

We can estimate the recognition accuracy for the target speech data from the WAN-based ETC results, i.e. the CSID probabilities, of the data *without using* the reference transcriptions. The estimated number of correct recognitions, i.e. $E(\#\mathrm{C})$, can be obtained as $E(\#\mathrm{C}) = \sum_i \sum_j P(\mathrm{C}|s_{i,j})$, where $s_{i,j}$ is the jth segment on the WAN for the ith utterance in the target speech data and $P(\mathrm{C}|s_{i,j})$ is the correct recognition probability of $s_{i,j}$. $E(\#\mathrm{S})$, $E(\#\mathrm{I})$ and $E(\#\mathrm{D})$ can be obtained in the same way. The estimated number of words, i.e. $E(\#\mathrm{N})$, can be obtained as $E(\#\mathrm{N}) = E(\#\mathrm{C}) + E(\#\mathrm{S}) + E(\#\mathrm{D})$. From these estimated numbers, by following the formal calculation procedure, the two standard recognition accuracy measures for the target speech data can be estimated as %Cor $= (E(\#\mathrm{C})/E(\#\mathrm{N})) \times 100$ [%] and WAcc $= ((E(\#\mathrm{C}) - E(\#\mathrm{I}))/E(\#\mathrm{N})) \times 100$ [%].

This WAN-based method is *easy-to-use* since it estimates the recognition accuracy almost directly from the recognition results and it does not require any model training, unlike the second CRF-based method described in Section II-B. In addition, it naturally detects deletion errors while, with the CRF-based method, we need a special procedure for detecting deletion errors.

However, a speech recognizer usually provides the 1-best recognition results with overestimation, and thus, in a segment of a WAN, the correct recognition probability tends to be overestimated while competing substitution and insertion error probabilities tend to be underestimated (e.g. as shown in Segments 1, 3, 4 and 7 of Fig. 1). In particular, the insertion error probability cannot be higher than 0.5 in a segment (e.g. Segments 3 and 4). As a result, with the WAN-based method, the recognition accuracy tend to be *overestimated*.

However, if the overestimated WAN-based recognition accuracy have a simple relationship with the true values (i.e. if a WAN-based recognition accuracy estimation result can be a good feature) the overestimated values can be corrected by designing a correction function as with the conventional recognition accuracy estimation methods [3]–[7]. We can actually design such a correction function based on a single *linear regression* as

$$y = ax + b, \tag{1}$$

where x is an overestimated WAN-based recognition accuracy and y is its corrected value. The two coefficients, i.e. the slope a and y-intercept b, can be analytically calculated using a large number of $(\bar{x}, \bar{y})$ pairs as training data. In the training phase, $\bar{y}$ is the true recognition accuracy.

With this correction, we can obtain accurate estimation results (see Section III-C). However, at the same time, some of the advantages of the WAN-based method are lost. We need to calculate the two coefficients in advance using training data. Thus, the method is no longer easy-to-use. The estimated numbers of CSIDs after applying this correction cannot be obtained.

If we want to estimate both %Cor and WAcc, we have to prepare an individual function for each of them.

B. CRF-Based Method

Our objective (the advantages of our methods) is the accurate estimation of *not only* the recognition accuracy *but also* the numbers of CSIDs. To achieve this objective, we have to perform more accurate ETC, i.e. CSID probability estimation, for each of the recognized words. We refine the WAN-based ETC results with *conditional random fields* (*CRFs*) [19], [20], using various types of additional features, e.g. the acoustic log likelihood, language log probability and lexical information, extracted for each of the recognized words that are not used with the WAN-based method. Our approach is basically the same as the recent trends in confidence estimation and is a promising one [13]–[18].

CRFs have been the most successful classification approach for addressing various types of sequential labeling problems in the fields of natural and spoken language processing. A CRF is a discriminative model that calculates the probability of an output label sequence given an input feature vector sequence. In the following, $\mathbf{x}_t$ is an input feature vector of a recognized word w_t. $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\#(\mathbf{X})}$ is a feature vector sequence extracted for a recognized word sequence $W = w_1, w_2, \ldots, w_{\#(\mathbf{X})}$ of length (number of words) $\#(\mathbf{X})$. y_t is an output label l that corresponds to $\mathbf{x}_t$. For example, in CSI classification, l denotes one of the three labels: C, S or I, i.e. $l \in \{\mathrm{C}, \mathrm{S}, \mathrm{I}\}$. $Y = y_1, y_2, \ldots, y_{\#(\mathbf{X})}$ is an output label sequence that corresponds to $\mathbf{X}$ and denotes a word-by-word CSI classification result. With the above definitions of $\mathbf{X}$ and Y, we can obtain a conditional probability $P(Y|\mathbf{X})$ by using a CRF as

$$P(Y|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{t=1}^{\#(\mathbf{X})} \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right), \quad (2)$$

where $f_k(y_t, y_{t-1}, \mathbf{x}_t)$ is the kth feature function (the state function or transition function) and λ_k is its weight, K is the number of feature functions, and $Z(\mathbf{X})$ is a normalization term defined as

$$Z(\mathbf{X}) = \sum_{Y'} \exp\left(\sum_{t=1}^{\#(\mathbf{X})} \sum_{k=1}^{K} \lambda_k f_k(y'_t, y'_{t-1}, \mathbf{x}_t)\right), \quad (3)$$

to satisfy $\sum_Y P(Y|\mathbf{X}) = 1$. The set of weights $\{\lambda_k\}_{k=1}^{K}$ can be estimated by the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and forward-backward algorithms with the L2 regularization using a large number of $(\bar{\mathbf{X}}, \bar{Y})$ pairs as training data. In the training phase, $\bar{Y}$ is the true output label sequence. In the evaluation phase, using a trained CRF, we obtain a marginal probability as

$$P(y_t = l|\mathbf{X}) = \sum_{y'_1,\ldots,y'_{t-1},y'_{t+1},\ldots,y'_{\#(\mathbf{X})}} P(y'_1, \ldots, y'_{t-1}, l, y'_{t+1}, \ldots, y'_{\#(\mathbf{X})}|\mathbf{X}). \quad (4)$$

This marginal probability can be efficiently calculated by the forward-backward algorithm. In CSI classification, $P(y_t = \mathrm{C}|\mathbf{X})$, $P(y_t = \mathrm{S}|\mathbf{X})$ and $P(y_t = \mathrm{I}|\mathbf{X})$ denote the correct recognition, substitution error and insertion error probabilities of the recognized word w_t, respectively. They are summed to 1, i.e. $P(y_t = \mathrm{C}|\mathbf{X}) + P(y_t = \mathrm{S}|\mathbf{X}) + P(y_t = \mathrm{I}|\mathbf{X}) = 1$.

TABLE I
THE 17 FEATURES OF A RECOGNIZED WORD

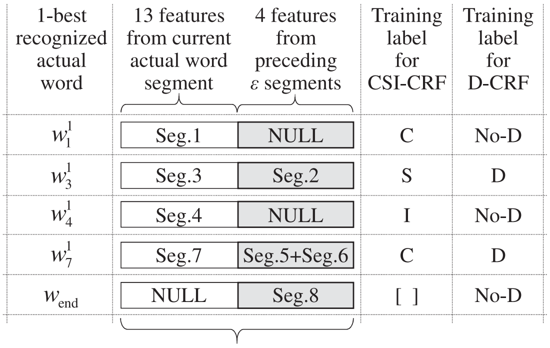
ID	Feature	Category
1	Recognized word (itself)	Word
2	Part-of-speech	
3	Acoustic log likelihood	Decoding
4	Unigram log probability	
5	Trigram log probability	
6	Number of frames	Supplementary
7	Number of phones	
8	Number of frames per phone	
9	Back-off behavior [16]	
10	Number of alternative hypotheses	
11	*Correct recognition probability*	WAN-CSI
12	*Substitution error probability*	
13	*Insertion error probability*	
14	*Sum of deletion error probabilities*	Preceding
15	Sum of ε probabilities	
16	Number of ε segments	
17	Sum of number of alternative hypotheses	

IDs 11 to 14 (italic) are WAN-based CSID probabilities that are refined with CRFs. IDs 1 to 13 are extracted from the current actual word segment. IDs 14 to 17 are extracted from preceding ε segments. The five feature categories used in Section III-F are also defined.

Table I lists the 17 features that we used in the experiments described in Section III. They consist of the basic features (IDs 1 to 10) used in the recent discriminative model based confidence estimation methods, the WAN-based ETC results, i.e. the CSID probabilities (IDs 11 to 14), and the features (IDs 14 to 17) extracted from the ε segments on a WCN and WAN that precede an actual word segment. The WAN-based CSID probabilities (IDs 11 to 14) are refined with CRFs.

The CRF-based ETC is performed with two CRFs. One performs CSI classification, i.e. the CRF classifies the 1-best recognized *actual* words (e.g. w_1^1, w_3^1, w_4^1 and w_7^1 in Fig. 1) into three categories: the correct recognition (C), substitution error (S) and insertion error (I). Hereafter, this CRF is referred to as a *CSI-CRF*. The other performs D detection, i.e. the CRF detects deletion errors (D) that may occur *between* the 1-best recognized actual words (*inter-word positions*, e.g. Segments 2, 5, 6 and 8 in Fig. 1). Hereafter, this CRF is referred to as a *D-CRF* (recently, a similar CRF-based deletion error detection method was also proposed in [24]).

Preparing feature vectors and labels for training a CSI-CRF is straightforward since one feature vector and one label are given to one recognized *actual* word. In contrast, preparing feature vectors and labels for training a D-CRF needs some consideration, since *multiple* (*consecutive*) deletion errors can occur at *arbitrary* inter-word positions in a recognized actual word sequence. This means that we have to solve two problems as regards D detection. One is the *position problem*, which requires us to find the inter-word positions that have deletion errors. The other is the *number problem*, which requires us to count the number of deletion errors in an inter-word position. In this paper, we try to solve the position problem only (in contrast to the CRF-based method, the WAN-based method naturally tries to solve both the number and position problems as described



1-best recognized actual word	13 features from current actual word segment	4 features from preceding ε segments	Training label for CSI-CRF	Training label for D-CRF
w_1^1	Seg.1	NULL	C	No-D
w_3^1	Seg.3	Seg.2	S	D
w_4^1	Seg.4	NULL	I	No-D
w_7^1	Seg.7	Seg.5+Seg.6	C	D
w_{end}	NULL	Seg.8	[ ]	No-D

Fig. 2. Feature vectors and labels for training a CSI-CRF and a D-CRF. This figure is based on Fig. 1.

in Section II-A). This is because, investigating the data used in the experiments described in Section III, we found that about 80% of the deletion errors were *singletons* and the others were consecutive. Our D-CRF detects inter-word positions that have one or more deletion errors, i.e. it classifies inter-word positions into two categories: those that have *one or more* deletion errors (labeled "D") and those that have *no* deletion errors (labeled "No-D").

As a solution to the position problem in D detection, we construct feature vectors and labels for training a D-CRF (and a CSI-CRF) based on the 1-best recognized *actual* words as shown in Fig. 2 (this figure is based on Fig. 1). We first look at the row for the recognized word w_3^1 (Segment 3). 13 features (IDs 1 to 13 in Table I) are extracted from the current actual word segment (Segment 3) and 4 features (IDs 14 to 17) are extracted from the *preceding* ε segment (Segment 2). Since w_3^1 is a substitution error word as shown in the *true* word alignment result in the fifth row of Fig. 1, the CSI-CRF is given the label S. Since one deletion error occurs between w_3^1 and w_1^1 (the inter-word position of the recognized actual two words that *precedes* w_3^1), the D-CRF is given the label D. With the row for w_7^1 (Segment 7), since there are two consecutive ε segments that precede w_7^1 (Segments 5 and 6), the 4 features are extracted by summing the two sets of 4 features extracted from these two ε segments, and the D-CRF is given the label D. With the row for w_4^1 (Segment 4), since there is no ε segment that precedes w_4^1, the 4 features are not extracted (represented as "NULL"), and, since there is no deletion error between w_4^1 and the preceding w_3^1, the D-CRF is given the label No-D. The last word w_{end} is a dummy word used for detecting the deletion errors that follow the last recognized actual word w_7^1.

All of the labels (C, S, I, No-D and D) shown in Fig. 2 are provided with a scoring tool (e.g. SCLITE [8]) by making a *word alignment* between the 1-best recognition result and its reference transcription (the *true* word alignment result shown in the fifth row of Fig. 1). We can assume that the 13 features (and especially the WAN-based CSI probabilities, i.e. IDs 11 to 13) are useful mainly for training a CSI-CRF, and the 4 features are useful mainly for a D-CRF. However, the 13 features may also be useful for the D-CRF and the 4 features may also be useful for the CSI-CRF (these are investigated in Section III-F). In addition, there is a case (not included in Fig. 2) where deletion errors occur at the inter-word positions that do not have the 4 features (like the rows for w_1^1 and w_4^1 in Fig. 2). In this case, we have to detect deletion errors using only the information from the 13 features. Therefore, we use the concatenated 17-dimension feature vector for both CSI-CRF and D-CRF. We also use the contextual features [16] and the n-gram dependency of the (output) labels in the experiments described in Section III.

By training a CSI-CRF and a D-CRF using the feature vectors and labels described above, the CSI-CRF can perform CSI classification for each of the recognized actual words and the D-CRF can perform D detection for each of the inter-word positions in the recognized actual word sequences. As a result of this CRF-based ETC, we can obtain the *refined* ETC results, i.e. the CSID probabilities. Then, using these refined CSID probabilities, we can estimate the recognition accuracy more accurately than with the WAN-based method described in Section II-A.

III. Experiments

We conducted basic experiments to evaluate the recognition accuracy estimation methods proposed in Section II. The results are described in Sections III-B and III-C. We also conducted a series of additional experiments and analyses to gain a better understanding of our methods and to confirm their effectiveness more consistently. The results are described in Sections III-D, III-E, III-F and III-G.

A. Experimental Settings

We used three tools. Our weighted finite-state transducer (WFST)-based speech recognition platform SOLON [25] was used to perform speech recognition. SCLITE [8] was used to obtain the true word alignment results between the 1-best recognition results from WCNs and their reference transcriptions. CRF++ [26] was used to train the CRFs and perform ETC with the trained CRFs.

We used two lecture speech corpora with different characteristics. One was the MIT lecture English speech corpus (hereafter, referred to as *MIT corpus*) [27], [28], which consists of speech data from regular university classes and invited talks. The other was the Corpus of Spontaneous Japanese (*CSJ corpus*) [29], [30], which consists of speech data from short lecture presentations in academic conferences. Each corpus contains speech data under various conditions, e.g. in terms of speakers (who provide good/bad speech recognition results) and in terms of recording environments (lecture room sizes, noises, reverberations, etc.). Since the experimental procedures are basically the same for both corpora, in the following, we mainly describe the procedure with the MIT corpus, and describe the corresponding values for the CSJ corpus in parentheses.

An HMM-based acoustic model was discriminatively trained by using 110 (250) hours of speech data with the differenced maximum mutual information (dMMI) criterion [31]. It had 2565 (2805) states optimized with the variational Bayesian estimation and clustering (VBEC) technique [32], and each state had 32 (32) -mixture Gaussian pdfs. A word trigram language model was trained by using 6.2 M (3.4 M) words of manually transcribed lecture speech. The vocabulary size of the lexicon was 16.5 k (59.9 k).

The CRF training data consisted of 238 (1335) lectures, 215 (277) hours, 114 k (220 k) utterances, and 2.0 M (3.5 M) words of speech. We performed speech recognition for this CRF training data, generated WCNs, converted the WCNs to WANs with the procedure described in Section II-A, and trained a CSI-CRF and a D-CRF with the procedure described in Section II-B. To estimate the labels (C, S, I, No-D and D) for the current actual word (see Fig. 2), we designed the state feature functions that consider the feature vectors within the preceding and succeeding 2 (3) words with up to 2 (3) contextual features [16]. We also designed the transition feature functions to consider the bi-gram dependency of the (output) labels. The number of feature functions (K in Eq. (2)) in the CSI-CRF was about 1.2 M (4.6 M) and that in the D-CRF was about 0.8M (3.0M). The CRF parameters (the set of weights $\{\lambda_k\}_{k=1}^{K}$ of the feature functions in Eq. (2)) were trained by L-BFGS and forward-backward algorithms with the L2 regularization.

The CRF development data consisted of 2 (10) lectures, 2.1 (2.1) hours, 3.5 k (1.6 k) utterances, and 24 k (34 k) words of speech. We applied the same procedures to this CRF development data as those applied to the training data as described above and used the processed data to determine two hyperparameters in the CRF training. One was the parameter (available with the `-c` option in CRF++) that controlled the fitness of the CRFs to the training data. With a larger value of this parameter, the CRFs tended to overfit the training data. We determined the value of this parameter at 1.5 (1.5). Another was the parameter (the `-f` option) that set the cutoff threshold of the feature functions. The feature functions that only occurred below the cutoff threshold were discarded. We determined the value of this parameter at 3 (3). The values of the other hyperparameters were set at their default values in CRF++.

The evaluation data consisted of 10 (30) lectures, 9.3 (5.7) hours, 9.5 k (3.2 k) utterances and 94 k (70 k) words of speech. We performed speech recognition for this evaluation data, generated WCNs, converted the WCNs to WANs, performed CSI classification and D detection by using the CSI-CRF and D-CRF (by obtaining the marginal probabilities described in Eq. (4)), respectively, and finally estimated the recognition accuracy from the WANs and CRF-based ETC results, respectively, *without using* the reference transcriptions based on the procedures described in Sections II-A and II-B. The true recognition accuracies were calculated from the true word alignment results, i.e. calculated using the reference transcriptions.

B. ETC Results

Before describing the recognition accuracy estimation results, we show in Table II the ETC results for *all* the evaluation data of the (a) MIT and (b) CSJ corpora. For comparison, we also show the confidence estimation results. WAN-based confidence estimation can be easily performed by summing the substitution and insertion error probabilities as the incorrect recognition probability at each of the actual word segments in WANs. To perform CRF-based confidence estimation, we trained a CRF that probabilistically classifies an actual recognized word into two categories: the correct recognition and incorrect recognition. In addition to the classification accuracies, Table II shows the F-scores for detecting each type of label. This is because the frequencies of the labels are highly *unbalanced*. In D detection, as described in Section II-B, if an inter-word position that has one or more deletion errors is detected, it is counted as a correct detection. Note that, as described in Section II-A, the linear regression-based method (described in Eq. (1)) only estimates the recognition accuracy. Thus, its (word-by-word based) ETC performance cannot be evaluated.

TABLE II
EXPERIMENTAL RESULTS OBTAINED WITH THE WAN- AND CRF-BASED METHODS FOR ALL THE EVALUATION DATA OF THE (A) MIT AND (B) CSJ CORPORA

(a) MIT results			
Confidence estimation		WAN	CRF
	Classification accuracy	81.19	**83.98**
F-score	Correct (77.07)	89.15	**90.06**
	Incorrect (22.93)	63.94	**65.96**
CSI classification		WAN	CRF
	Classification accuracy	79.38	**81.83**
F-score	Correct (77.07)	89.15	**90.06**
	Substitution (18.52)	55.92	**59.28**
	Insertion (4.41)	21.91	**35.83**
D detection		WAN	CRF
	Classification accuracy	94.56	**95.98**
F-score	No deletion (96.14)	98.03	98.05
	Deletion (3.86)	21.79	**31.73**
(b) CSJ results			
Confidence estimation		WAN	CRF
	Classification accuracy	87.56	**89.10**
F-score	Correct (85.44)	93.36	**93.83**
	Incorrect (14.56)	60.71	**62.95**
CSI classification		WAN	CRF
	Classification accuracy	86.91	**88.34**
F-score	Correct (85.44)	93.36	**93.93**
	Substitution (12.71)	55.89	**58.94**
	Insertion (1.86)	20.49	**31.02**
D detection		WAN	CRF
	Classification accuracy	95.73	**96.18**
F-score	No deletion (96.34)	98.14	98.19
	Deletion (3.66)	25.77	**32.51**

Classification accuracies [%] and F-scores [%] of confidence estimation and ETC (CSI classification and D detection). Values in parentheses are ratios [%] of the labels and are summed to 100% in each task. In each row, if the improvement of the best result from the corresponding result is statistically significant at the 1% level [33], the best result is shown in bold.

From Table II, we can confirm for each of the corpora that the CRF-based method performs better than the WAN-based method especially for infrequent labels, i.e. the insertion and deletion errors, in terms of the F-scores. This may be the effect of using various types of features with the CRFs including the CSID probabilities from the WANs as listed in Table I (Section III-F describes further experiments and analyses regarding the importance of each feature). However, the absolute performance of these infrequent labels remains at an unsatisfactory level. We can also confirm that the F-scores of correct recognition obtained with the CRF-based method in (three class)

TABLE III
EXPERIMENTAL RESULTS OBTAINED WITH THE WAN- AND CRF-BASED METHODS FOR ALL THE EVALUATION DATA OF THE (A) MIT AND (B) CSJ CORPORA

(a) MIT results

Method	#N	#C	#S	#I	#D	%Cor	WAcc
True	94449	72191	17345	4130	4913	76.43	72.06
WAN	**95449**	80173	10320	3173	**4956**	84.00	80.67
CRF	93420	**70990**	**18249**	**4427**	4181	**75.99**	**71.25**

(b) CSJ results

Method	#N	#C	#S	#I	#D	%Cor	WAcc
True	70206	58252	8663	1265	3291	82.97	81.17
WAN	**70180**	62043	5068	1068	**3067**	88.41	86.88
CRF	69457	**57602**	**9186**	**1392**	2669	**82.93**	**80.93**

The true numbers of NCSIDs and recognition accuracies (%Cors [%] and WAccs [%]) with their estimated values. In each column, the best estimation result is shown in bold.

CSI classification have the same level of accuracy as those obtained with (two class) confidence estimation (note that, with the WAN-based method, these F-scores are exactly the same).

C. Recognition Accuracy Estimation Results

Table III shows the recognition accuracy estimation results for *all* the evaluation data of the (a) MIT and (b) CSJ corpora. These recognition accuracies are calculated using estimated CSID numbers. For each of the corpora, we can confirm that, with the WAN-based method, #C is overestimated and, conversely, #S and #I are underestimated. As a result, the recognition accuracies are *overestimated* compared with their true values as described in Section II-A. We can also confirm that, with the CRF-based method, #D is underestimated. This is because, in D detection, we detect an inter-word position that has one or more deletion errors and count it as *one* deletion error even though the position has more than two deletion errors as described in Section II-B. However, with the CRF-based method, #C, #S and #I are estimated at a satisfactory level. As a result, the recognition accuracies are estimated at close to their true values. Note that, as described in Section II-A, the numbers of CSIDs cannot be estimated with the linear regression-based method (described in Eq. (1)), and thus the estimated recognition accuracies obtained with the method cannot be made available in Table III.

Fig. 3 shows the recognition accuracy estimation results for the *lecture-level* evaluation data of the (a) MIT and (b) CSJ corpora. For each of the corpora, we can confirm that the WAN-based method shows high WAcc estimation performance in terms of the correlation coefficient between the true WAccs because the estimated WAccs are distributed around a straight (chain) line (i.e. the regression line). However, the line shifts upward from the diagonal (dotted) line especially in the lower WAcc range. As a result, the root mean square error (RMSE) between the true and estimated WAccs is large.

The WAN-based lecture-level WAccs are distributed around the straight line for each of the two corpora. We corrected these WAccs with a *linear regression* as proposed in Section II-A. We used the same training data that we used for the CRF training. We obtained the WAN-based lecture-level WAccs ($\bar{x}$) for the training data and the corresponding true WAccs ($\bar{y}$), i.e. a large number Q of $(\bar{x}, \bar{y})$ pairs, $\{(\bar{x}^{(q)}, \bar{y}^{(q)})\}_{q=1}^{Q}$, where $Q = 238$ for the MIT corpus and 1335 for the CSJ corpus. Using these $(\bar{x}, \bar{y})$ pairs, we calculated the two coefficients, i.e. the slope a and y-intercept b in Eq. (1), for each of the two corpora.

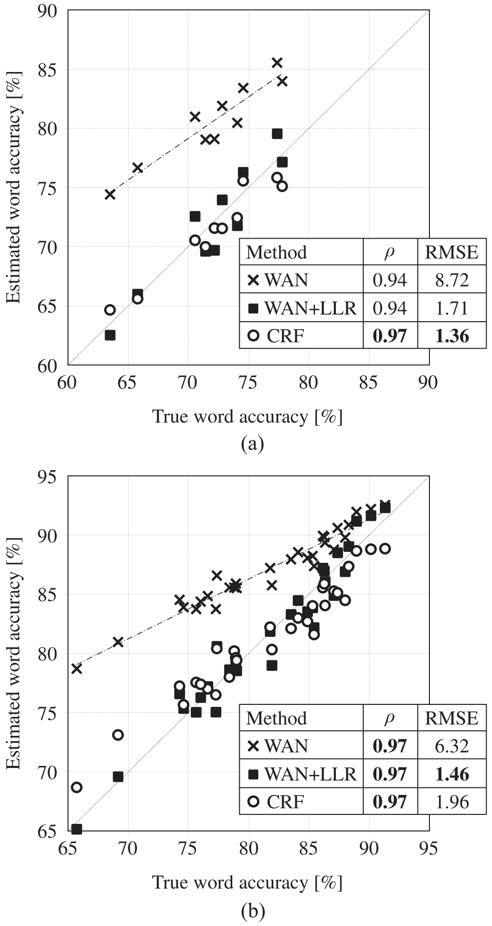
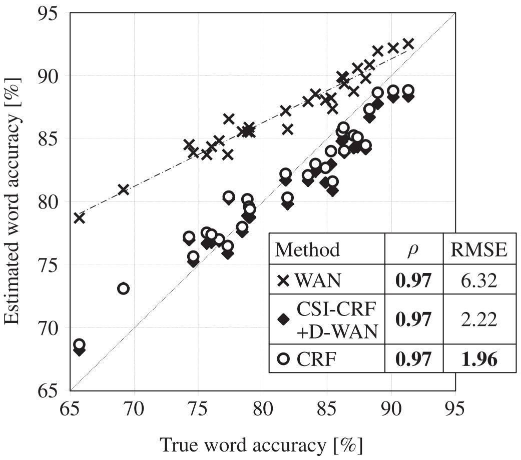


Fig. 3. Experimental results obtained with the WAN-based, WAN+LLR and CRF-based methods for the lecture-level evaluation data of the (a) MIT (10 lectures) and (b) CSJ (30 lectures) corpora. Correlations between the true WAccs [%] and estimated WAccs [%]. Correlation coefficients (ρ) and RMSEs [%] are also shown. The best results are shown in bold.

The resultant two coefficients were $(a, b) = (1.53, -51.23)$ for the MIT corpus and $(1.96, -89.40)$ for the CSJ corpus. a took values larger than 1 and these results meant that the correction worked to raise the slope of the regression lines (chain lines) that can be virtually imaged with the distributions of the WAN-based WAccs. b took negative values and these results meant that the correction worked to shift the WAN-based WAccs to the downward direction. Using these two calculated coefficients (a, b), we performed a linear regression according to Eq. (1) to the WAN-based lecture-level WAccs (x) estimated for the evaluation data and obtained their corrected values (y) for each of the two corpora.

The corrected lecture-level WAccs are plotted in Fig. 3 and the method is referred to as *WAN+LLR*, i.e. lecture-level linear regression for WAN-based WAcc estimation results. We can confirm that WAN+LLR provides the greatly improved (lower)



Method	ρ	RMSE
× WAN	**0.97**	6.32
◆ CSI-CRF +D-WAN	**0.97**	2.22
○ CRF	**0.97**	**1.96**

Fig. 4. Experimental results obtained with the WAN-based, CSI-CRF+D-WAN and CRF-based methods for the lecture-level evaluation data of the CSJ corpus (30 lectures). Correlations between the true WAccs [%] and estimated WAccs [%]. Correlation coefficients (ρ) and RMSEs [%] are also shown. The best results are shown in bold.

RMSE values compared with the WAN-based method (especially, for the CSJ corpus). As described in Section II-A, this linear regression-based method borrows the framework of the conventional recognition accuracy estimation methods [3]–[7]. However, its input feature is the proposed WAN-based (lecture-level) recognition accuracy estimation result. We can say that we found a good feature that reflects various factors that affect the recognition accuracy (see Section I) for the conventional recognition accuracy estimation framework.

We also confirm in Fig. 3 that the CRF-based method refines the WAN-based CSID probabilities by using many useful additional features as described in Section II-B and estimates WAccs very accurately. The correlation coefficient is 0.97 and the RMSE is lower than 2% for each of the two corpora. The performance of the CRF-based method is comparable to that of WAN+LLR. However, it should be noted that the CRF-based method can also estimate the numbers of CSIDs, which cannot be estimated with WAN+LLR.

Each corpus contains two lectures that provide lower true WAccs than the other lectures. This indicates that the ASR models (i.e. the acoustic, language and/or lexical models) did not match the speech data included in those lectures. The CRF-based method and WAN+LLR (especially, the latter) can accurately estimate (track) WAccs even for those *outlier* speech data.

D. Combination of CSI-CRF and D-WAN

The experimental results shown in Table III prompted us to combine the CRF-based CSI classification and the WAN-based D detection. This is because #C, #S and #I values are more accurately estimated with the CRFs (i.e. the CSI-CRFs) than with the WANs and, conversely, #D values are more accurately estimated with the WANs than with the CRFs (i.e. the D-CRFs). The combined method, hereafter referred to as *CSI-CRF+D-WAN*, is more convenient than the CRF-based method since it does not require the D-CRFs.

Fig. 4 shows the WAcc estimation results obtained with the WAN-based, CSI-CRF+D-WAN and CRF-based methods for the lecture-level evaluation data of the CSJ corpus. We can confirm that CSI-CRF+D-WAN shows (slightly) worse performance (in terms of the RMSE) than the CRF-based method. In CSI-CRF+D-WAN, CSI classification and D detection are performed *separately* by using the CSI-CRF and WAN, respectively. Also with the CRF-based method, they are performed separately by using the CSI-CRF and D-CRF, respectively. However, as described in Section II-B, these two CRFs use the *common* 17 features listed in Table I. Thus, the CRF-based method might have the ability to realize *balance* between each of the estimated numbers of CSIDs (specifically, there should be a relationship between #I and #D, i.e., with a reasonable ASR setting, there should not be a very large difference between these two numbers). These differences between CSI-CRF+D-WAN and the CRF-based method might be a reason for the slight difference in their performance.

E. Word-Level Linear Regression of WAN-Based ETC Results

As shown in Fig. 3, the WAN-based lecture-level WAcc estimation results are distributed around straight lines. These results also motivated us to correct the WAN-based ETC results at the *word-level* as with the lecture-level correction (linear regression) of the WAN-based WAcc estimation results (WAN+LLR), which we performed in Section III-C. A word-level correction method for the WAN-based ETC results is desirable since it can estimate the numbers of CSIDs as with the CRF-based method.

We performed a word-level single *linear regression* for the WAN-based ETC results. Hereafter, we refer to this method as *WAN+WLR*. Its basic framework is the same as that of WAN+LLR. We performed it for each of the labels in the WAN-based ETC results, *individually*, as

$$y_l = a_l x_l + b_l, \tag{5}$$

where l denotes one of the four labels: C, S, I or D (i.e. $l \in \{\mathrm{C}, \mathrm{S}, \mathrm{I}, \mathrm{D}\}$), x_l is a WAN-based ETC result, y_l is its corrected value, and (a_l, b_l) are the two coefficients, i.e. the slope and y-intercept, of the linear regression.

In CSI classification (i.e. $l \in \{\mathrm{C}, \mathrm{S}, \mathrm{I}\}$), x_l denotes a C probability (x_{C}), an S probability (x_{S}) or an I probability (x_{I}) for a recognized actual word as described in Section II-A, and satisfies $0 \leq x_l \leq 1$. y_l is also a probability. However, it may not satisfy the constraint, and thus we convert it to $\hat{y}_l$ as

$$\hat{y}_l = \begin{cases} 0, & y_l < 0, \\ 1, & y_l > 1, \\ y_l, & \text{otherwise}, \end{cases} \tag{6}$$

to satisfy $0 \leq \hat{y}_l \leq 1$. In addition, as x_{C}, x_{S} and x_{I} satisfy $x_{\mathrm{C}} + x_{\mathrm{S}} + x_{\mathrm{I}} = 1$, the sum of $\hat{y}_{\mathrm{C}}$, $\hat{y}_{\mathrm{S}}$ and $\hat{y}_{\mathrm{I}}$ should also be equal to 1. However, this constraint is not usually satisfied. To satisfy this constraint, we further convert $\hat{y}_l$ to $\tilde{y}_l$ as

$$\tilde{y}_l = \hat{y}_l \,/\, (\hat{y}_{\mathrm{C}} + \hat{y}_{\mathrm{S}} + \hat{y}_{\mathrm{I}}). \tag{7}$$

In D detection (i.e. $l = \mathrm{D}$), x_D denotes the *sum* of the D probabilities for an inter-word position between two recognized

TABLE IV
EXPERIMENTAL RESULTS OBTAINED WITH THE WAN-BASED, WAN+WLR AND CRF-BASED METHODS FOR THE EVALUATION DATA OF THE CSJ CORPUS

Method	CSI classification				D detection			WAcc estimation			
		F-score				F-score				Lecture-level data	
	Class. accuracy	C	S	I	Class. accuracy	No-D	D	All data (diff. from true)		ρ	RMSE
WAN	86.91	93.36	55.89	20.49	95.73	98.14	25.77	86.88	(+5.71)	**0.97**	6.32
WAN+WLR	87.56	93.38	56.11	20.51	96.31	98.14	25.83	79.45	(−1.72)	**0.97**	3.83
CRF	**88.34**	**93.93**	**58.94**	**31.02**	96.18	98.19	**32.51**	**80.93**	(−**0.24**)	**0.97**	**1.96**

(Left and center) Classification accuracies [%] and F-scores [%] of ETC (CSI classification and D detection) for all the evaluation data. In each column, if the improvement of the best result from the second best result is statistically significant at the 1% level [33], the best result is shown in bold. (Right) WAcc [%] estimation results for all the evaluation data (the true WAcc is 81.17%). Correlation coefficients (ρ) and RMSEs [%] between the true WAccs [%] and estimated WAccs [%] for the lecture-level evaluation data (30 lectures). In each column, the best result(s) is(are) shown in bold.

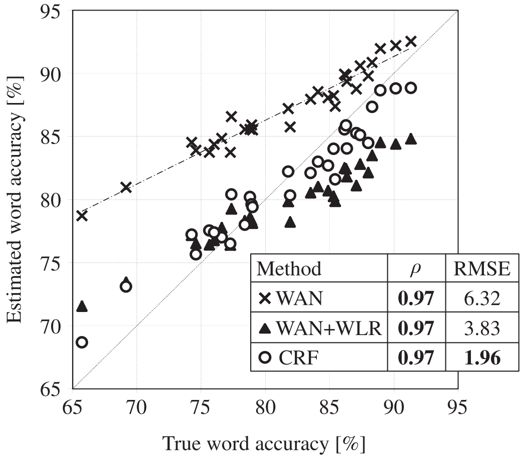


Fig. 5. Experimental results obtained with the WAN-based, WAN+WLR and CRF-based methods for the lecture-level evaluation data of the CSJ corpus (30 lectures). Correlations between the true WAccs [%] and estimated WAccs [%]. Correlation coefficients (ρ) and RMSEs [%] are also shown. The best results are shown in bold.

actual words, and satisfies $x_D \geq 0$. y_D is also a sum of probabilities. However, it may not satisfy the constraint, and thus we convert it to $\hat{y}_D$ as

$$\hat{y}_D = \begin{cases} 0, & y_D < 0, \\ y_D, & \text{otherwise}, \end{cases} \tag{8}$$

to satisfy $\hat{y}_D \geq 0$. Note that, we need these adjustments described in Eqs. (6), (7) and (8) for WAN+WLR (practically, the need of these adjustments depends on the label l). In contrast, we did not need any adjustment for WAN+LLR, which we performed in Section III-C, i.e. WAN+LLR was stable.

We evaluated WAN+WLR using the CSJ corpus. As with when we evaluated WAN+LLR, we used the same training data used for the CRF training described in Section III-A. We obtained the WAN-based ETC results ($\bar{x}_l$) for the training data and the corresponding true ETC results ($\bar{y}_l$), i.e. a large number (about) 3.5M of $(\bar{x}_l, \bar{y}_l)$ pairs, $\{(\bar{x}_l^{(r)}, \bar{y}_l^{(r)})\}_{r=1}^{3.5M}$, where $l \in \{\text{C, S, I, D}\}$. Using these $(\bar{x}_l, \bar{y}_l)$ pairs, we calculated the two coefficients (a_l, b_l) in Eq. (5) for each type of label $l \in \{\text{C, S, I, D}\}$.

The resultant two coefficients were (1.24, −0.29) for C, (1.21, 0.04) for S, (0.53, 0.01) for I, and (0.53, 0.03) for D. Focusing on the b_l values, we can understand that the correction worked to decrease the C probabilities and to increase the S, I and D probabilities. Using these two coefficients (a_l, b_l), we performed word-level linear regression according to Eq. (5) on the WAN-based ETC results (x_l) for the evaluation data and obtained their corrected values (y_l) for each type of label $l \in \{\text{C, S, I, D}\}$.

Table IV and Fig. 5 show the experimental results we obtained with the WAN-based, WAN+WLR, CRF-based methods for the CSJ evaluation data. We can confirm that WAN+WLR exhibits a steady performance improvement compared with the WAN-based method. These results indicate that the *simple* word-level linear regression for the WAN-based ETC results works well at a certain level, and the WAN-based ETC results can be good features for succeeding ETC and recognition accuracy estimation.

However, the performance of WAN+WLR does not reach that of the CRF-based method. These results indicate that there must be a *non-linear* relationship between the WAN-based ETC results and the corresponding true ETC results. The linear regression is insufficient to capture this non-linear relationship, and so we require a CRF that can model a non-linear relationship between input feature sequences and output label sequences (see Eq. (2)). Here, it should be noted (again) that, if we want to estimate the recognition accuracy only, the linear regression is sufficient. As described in Section III-C, WAN+LLR has the (lecture-level) recognition accuracy estimation performance comparable to that of the CRF-based method by focusing on correcting the WAN-based WAcc estimation results to approximate the true WAccs (and abandoning any estimation of the numbers of CSIDs).

The experimental results shown in Table IV and Fig. 5 also indicate the importance of using various types of additional features (listed in Table I) in the CRF-based method, which are not used in WAN+WLR. Further experiments and analyses are described in the next section that reveal the importance of each feature in the CRF-based method.

F. Importance Analyses of Features Used in CRFs

We have confirmed the effectiveness of a CRF-based method that uses various types of features. In this section, we conduct experiments to reveal which features are important in the CRF-based method for ETC and recognition accuracy estimation. To conduct the experiments, as shown in Table I, we group the 17 features into five categories.

The first category is referred to as *word*. The recognized word (itself) and its part-of-speech (IDs 1 and 2 in Table I) are grouped into this category. These features can be very important since, by using these features, the relationship between a recognized word and its error trend can be *directly* captured. However, at the same time, these are *sparse* (i.e. *task dependent*) features. Thus, if we want to decrease the task dependency (increase the task versatility) of the CRF-based method, we should not use these *word* features.

The second category is referred to as *decoding*. The *raw* scores obtained as a result of ASR decoding, i.e. the acoustic log likelihood and language log probabilities (IDs 3 to 5), are grouped into this category. The raw scores are first stored on a recognition lattice and, through the procedure described in Section II-A, converted to (more reliable) CSID (posterior) probabilities on a WAN. The WAN-based CSID probabilities are also used, and thus these *decoding* features may be redundant.

The third category is referred to as *supplementary*. The 5 features (IDs 6 to 10) that cannot be grouped into the other categories are gathered into this category. Compared with the other features, these features may only have *less direct* information about the recognition errors. We can assume that these *supplementary* features are not very important.

The fourth category is referred to as *WAN-CSI*. The WAN-based CSI probabilities (IDs 11 to 13) obtained from an actual word segment on a WAN are grouped into this category. As also assumed in Section II-B, we can assume that these *WAN-CSI* features are useful especially for CSI classification.

The fifth (last) category is referred to as *preceding*. The 4 features (IDs 14 to 17) extracted from the ε segments on a WCN and WAN that precede an actual word segment are grouped into this category. As also assumed in Section II-B, we can assume that these *preceding* features are useful especially for D detection.

Using the CSJ training data and the five feature categories defined above, we repeated the CRF training five times while changing the feature category to be deleted each time. The training procedure was the same as that described in Section III-A except for the deletion of one of the five feature categories. As a result, we obtained the five CRFs (both CSI-CRFs and D-CRFs) that did not have one of the five feature categories. We evaluated the five CRF-based methods based on these five CRFs using the CSJ evaluation data. Hereafter, we refer to the CRF-based method that does *not* use the feature category *X* as the *No-X* method.

Table V shows the experimental results for the CSJ evaluation data that we obtained with the five No-X methods along with the baseline CRF-based method that uses all of the 17 features. We can confirm that, in many cases, the No-decoding and No-supplementary methods show similar (or sometimes, slightly better) results than the baseline method. These results indicate that the *decoding* and *supplementary* features (especially, the former) are not very important as we assumed above.

In contrast, in many cases, the No-word, No-WAN-CSI and No-preceding methods provide worse results than the baseline method. These results indicate the importance of the *word*, *WAN-CSI* and *No-preceding* features as we assumed above. In particular, the *WAN-CSI* and *preceding* features are important in CSI classification and D detection. It is reasonable that the No-WAN-CSI method shows worse results in CSI classification (especially, in the F-scores for detecting S and I) and the No-preceding method shows worse results in D detection (especially, in the F-score for detecting D). However, it is interesting to see that the No-preceding method also shows worse results in CSI classification, i.e. the *preceding* features are also important for CSI classification (but, not vice versa, i.e. the *WAN-CSI* features may not be important for D detection). These results support our strategy of using the *common* features (including the *WAN-CSI* and *preceding* features) for both CSI-CRFs and D-CRFs as described in Section II-B.

From the experimental results obtained with the No-WAN-CSI and No-preceding methods, we can expect to obtain good performance by concatenating only the *WAN-CSI* and *preceding* features. We trained and evaluated a CRF (both CSI-CRF and D-CRF) that used only these two feature categories. Hereafter, we refer to the CRF-based method based on this CRF as *WAN-CSI+Preceding*. With the same expectation, we also trained and evaluated a CRF (both CSI-CRF and D-CRF) that used only the WAN-based CSID probabilities, i.e. the 4 features of IDs 11 to 14 in Table I. We refer to this CRF-based method as *WAN-CSID*.

The last two rows in Table V show the experimental results we obtained with WAN-CSI+Preceding and WAN-CSID. Contrary to our expectation, we cannot obtain good results with these two methods. These and the above results indicate that, to model the *non-linear* relationship between the WAN-based ETC results and the corresponding true ETC results (see Section III-E), the CRF-based method should use a *certain large number* of diverse features. If a CRF has a sufficient number (e.g. 17 in our experiments) of diverse features, we can eliminate a few unimportant features (e.g. the *decoding* or *supplementary* features) from them to maintain (or slightly improve) the performance. Conversely, we cannot obtain a good CRF with only a small number of features even though the features are important (e.g. the 7 features from the *WAN-CSI* and *preceding* feature categories, or the 4 WAN-based CSID probabilities). With the CRF-based method, we should use various types of additional features along with the important features.

WAN-CSID and WAN+WLR (see Section III-E) use the same 4 features, i.e. the WAN-based CSID probabilities. However, WAN-CSID performs much better than WAN+WLR (see Table IV). These results again indicate that, as described in Section III-E, that there must be a *non-linear* relationship between the WAN-based ETC results and the corresponding true ETC results. The linear regression is insufficient to capture this *non-linear* relationship, and thus the CRF is required.

G. Results With Reduced Size of CRF Training Data

We have confirmed that, with the CRF-based method, the recognition accuracy can be estimated accurately. However, in the experiments described above, we used a large amount of data to train the CRFs (see Section III-A). As described in Section I,

TABLE V
EXPERIMENTAL RESULTS OBTAINED WITH THE BASELINE CRF-BASED METHOD, THE FIVE NO-X METHODS, WAN-CSI+PRECEDING AND WAN-CSID FOR THE EVALUATION DATA OF THE CSJ CORPUS

Method (# of features used)	CSI classification				D detection			WAcc estimation			
		F-score				F-score				Lecture-level data	
	Class. accuracy	C	S	I	Class. accuracy	No-D	D	All data (diff. from true)		ρ	RMSE
CRF (baseline, all 17 features)	88.34	93.93	58.94	31.02	96.18	98.19	32.51	80.93	(−0.24)	0.97	1.96
No−word (15)	88.01	93.69	57.91	30.55	96.32	98.16	32.37	81.24	(+0.07)	0.97	2.13
No-decoding (14)	88.40	93.95	59.19	31.96	96.22	98.18	32.55	81.21	(+0.04)	0.98	1.92
No-supplementary (12)	88.25	93.82	59.05	30.58	96.24	98.19	32.84	80.99	(−0.18)	0.97	2.04
No-WAN-CSI (14)	87.93	93.60	56.74	27.65	96.17	98.18	32.83	81.09	(−0.08)	0.97	2.07
No-preceding (13)	87.80	93.78	57.29	28.47	96.11	98.17	26.19	81.00	(−0.17)	0.97	2.11
WAN-CSI+Preceding (7)	87.73	93.45	56.89	27.84	96.29	98.14	30.14	81.06	(−0.11)	0.97	2.25
WAN-CSID (4)	87.73	93.43	57.01	28.36	96.30	98.14	29.15	81.13	(−0.04)	0.97	2.28

(Left and center) Classification accuracies [%] and F-scores [%] of ETC (CSI classification and D detection) for all the evaluation data. In each column, if the degradation of a No-X, WAN-CSI+Preceding or WAN-CSID compared with the baseline is statistically significant at the 1% level [33], its result is shown with an underline. (Right) WAcc [%] estimation results for all the evaluation data (the true WAcc is 81.17%). Correlation coefficients (ρ) and RMSEs [%] between the true WAccs [%] and estimated WAccs [%] for the lecture-level evaluation data (30 lectures). In each column, if a No-X, WAN-CSI+Preceding or WAN-CSID is worse than the baseline, its result is shown with an underline.

TABLE VI
EXPERIMENTAL RESULTS OBTAINED WITH THE CRF-BASED METHOD FOR THE EVALUATION DATA OF THE CSJ CORPUS AS A FUNCTION OF THE CRF TRAINING DATA SIZE

Training data size			WAcc estimation			
					Lecture-level data	
hours	#utts	#words	All data (diff. from true)		ρ	RMSE
277	220 k	3.5 M	80.93	(−0.24)	0.97	1.96
200	159 k	2.5 M	81.02	(−0.15)	0.97	1.85
100	80 k	1.3 M	81.00	(−0.17)	**0.98**	1.83
50	40 k	629 k	**81.27**	**(+0.10)**	**0.98**	**1.76**
20	16 k	251 k	81.34	(+0.17)	0.97	1.94
10	8 k	125 k	80.96	(−0.21)	0.97	1.91

WAcc [%] estimation results for all the evaluation data (the true WAcc is 81.17[%]). Correlation coefficients (ρ) and RMSEs [%] between the true WAccs [%] and estimated WAccs [%] for the lecture-level evaluation data (30 lectures). In each column, the best result(s) is(are) shown in bold.

the cost of manual transcription is very high. To confirm whether or not we can reduce the cost for preparing the CRF training data, we evaluated the CRF-based method while reducing the training data size.

We conducted the experiments using the CSJ corpus. The CRF training data consisted of 1335 lectures, 277 hours, 220 k utterances, and 3.5 M words of speech. We gradually reduced the data size from 277 hours (the full data size) to 10 hours and obtained six training data sets (including the full size data). We listed all the utterances in the training data and selected utterances from this list with an equal interval. For example, with 50 hours, we selected an utterance every 5.5 utterances. This means that a larger size of data does not completely include a smaller size of data. With the procedure described in Section III-A, we trained the six CRFs (both CSI-CRFs and D-CRFs) using these six sets of CRF training data and evaluated them using the CSJ evaluation data.

Table VI shows the WAcc estimation performance of the CRF-based method as a function of the CRF training data size. We can confirm that the training data size can be greatly reduced while maintaining the WAcc estimation performance. The performance of the 50 hours of data is slightly better than that of the other sizes of data. There is no reasonable way to explain these results. We assume that the characteristics of the 50 hours of data coincidentally match those of the evaluation data.

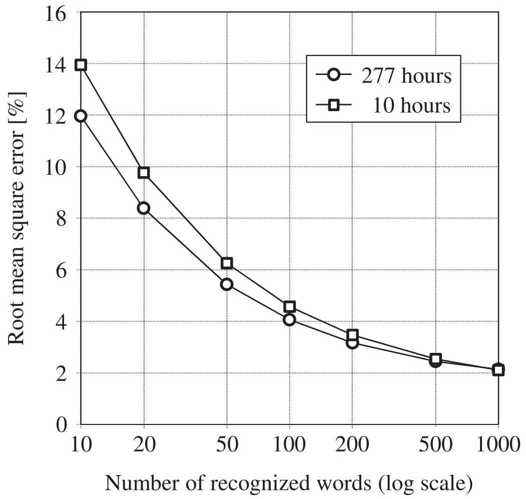


Fig. 6. Experimental results obtained with the CRF-based method for the lecture-level evaluation data of the CSJ corpus (30 lectures). RMSEs [%] between the true WAccs [%] and estimated WAccs [%] as a function of the number of recognized words for calculating WAccs. Two CRFs are used: one trained with 277 hours (full) of data and the other with 10 hours of data.

One reason for the performance robustness against the reduced size of the CRF training data shown in Table VI is attributable to the number of recognized words used for calculating (summarizing) the recognition accuracy. In the CSJ corpus, one lecture consisted of an average of 2.3 k words (see Section III-A and Table III) and thus, in the experiments described above, the lecture-level recognition accuracies were calculated using the CSID probabilities estimated for about 2.3 k of recognized words. We can expect to obtain more stable estimation results by using more recognized words. We evaluated the convergence properties of the recognition accuracy estimation performance while increasing the number of recognized words when we calculated the recognition accuracy. We conducted the experiments using the CRF trained with 277 hours (full) of data and that trained with 10 hours of data while gradually increasing the number of recognized words from 10 to 1000.

Fig. 6 shows the convergence properties of the WAcc estimation performance in terms of the RMSE. We can confirm that the performance converges very quickly as the number of recognized words increased. We can also confirm a slight performance degradation of the CRF trained with 10 hours of data compared with that trained with 277 hours of data for small numbers of recognized words.

IV. Conclusion and Future Work

In this paper, we have proposed recognition accuracy estimation methods based on error type classification (ETC). In ETC, each word in the recognition results is probabilistically classified into three categories: the correct recognition (C), substitution error (S) and insertion error (I). Deletion errors (D) that can occur at inter-word positions in the recognition results are also probabilistically detected. By summing these CSID probabilities individually, the numbers of CSIDs and, as a result, the two standard recognition accuracy measures, i.e. the percent correct and word accuracy (WAcc), of the recognition results can be estimated without using the manually transcribed references.

Two recognition accuracy estimation methods based on ETC have been proposed. The first method performs ETC based on word alignment networks (WANs). It is an easy-to-use method, but it overestimates the recognition accuracy. The second method is based on conditional random fields (CRFs). It refines the WAN-based ETC results, i.e. the CSID probabilities, using various types of additional features and estimates the recognition accuracy accurately.

In experiments using the two lecture speech corpora, the correlation coefficient and root mean square error between the lecture-level true word accuracies calculated using the reference transcriptions and those estimated with the CRF-based method were 0.97 and lower than 2%, respectively. A series of additional experiments and analyses were also conducted to better understand the effectiveness of the CRF-based method.

We are planning to conduct the following studies as future work:

1) The two lecture speech corpora that we used in this paper have very diverse features. Therefore, the obtained experimental results ensure that the proposed recognition accuracy estimation methods are robust to such variation. However, to confirm the consistent effectiveness of our methods, we need to evaluate them using different types of speech corpora such as a voice search corpus and a spoken dialogue corpus, which contain shorter (a few words of) utterances (responses) than the relatively longer utterances included in the lecture speech corpus.
2) We used only our speech recognizer in the experiments. For the reason described above, we will evaluate our methods using other speech recognizers (e.g. HTK [34] and Kaldi [35]) or by changing the settings (e.g. the acoustic/language models and/or decoding parameters) of our speech recognizer to provide more (or less) accurate recognition results.
3) We will improve the ETC performance of the infrequent labels (i.e. I and D errors shown in Table II) by using more powerful models (e.g. neural networks) than CRFs (we reported our first trial in [36]) and solve the *number problem* in D detection (see Section II-B).

References

[1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning approach*. London, U.K.: Springer-Verlag, 2015.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[3] M. Kondo, K. Takeda, and F. Itakura, "Predicting the degradation of speech recognition performance from sub-band dynamic ranges," *J. Inf. Process. Soc. Jpn.*, vol. 43, no. 7, pp. 2242–2248, Jul. 2002.

[4] H. Sun, L. Shue, and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. I-865–I-868.

[5] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006.

[6] T. Yamada, T. Nakajima, N. Kitawaki, and S. Makino, "Performance estimation of noisy speech recognition considering task complexity," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2042–2045.

[7] T. Fukumori, M. Morise, and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria RSR-D_n with acoustic parameters," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 562–565.

[8] *Speech recognition scoring toolkit (SCTK) version 2.4.0*, 2009. [Online] Available: http://www.itl.nist.gov/iad/mig/tools/

[9] K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*. Norwell, MA, USA: Kluwer, 1989.

[10] HTKBook. 2015. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[11] A. Ogawa, T. Hori, and A. Nakamura, "Error type classification and word accuracy estimation using alignment features from word confusion network," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4925–4928.

[12] A. Ogawa, T. Hori, and A. Nakamura, "Recognition rate estimation based on word alignment network and discriminative error type classification," in *Proc. Workshop Spoken Lang. Technol.*, 2012, pp. 113–118.

[13] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Commun.*, vol. 45, no. 4, pp. 455–470, Apr. 2005.

[14] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 809–812.

[15] C. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence estimation, OOV detection, and language ID using phone-to-word transcription and phone-level alignments," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4085–4088.

[16] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros, "CRF-based combination of contextual features to improve a posteriori word-level confidence measures," in *Proc. Annu. Conf. Int. Speech, Commun. Assoc.*, 2010, pp. 1942–1945.

[17] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2461–2473, Nov. 2011.

[18] M. S. Seigel, "Confidence estimation for automatic speech recognition hypotheses," Ph.D. dissertation, Univ. of Cambridge, Cambridge, U.K., Dec. 2013. [Online]. Available: http://mi.eng.cam.ac.uk/mss46/papers/seigel_thesis.pdf

[19] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[20] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, Nov. 2012.

[21] A. Ogawa and A. Nakamura, "Joint estimation of confidence and error causes in speech recognition," *Speech Commun.*, vol. 54, no. 9, pp. 1014–1028, Nov. 2012.

[22] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proc. Spec. Interest Group Inf. Retrieval*, 2006, pp. 51–58.

[23] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 373–400, Oct. 2000.

[24] M. S. Seigel and P. C. Woodland, "Detecting deletions in ASR output," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 2302–2306.

[25] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1352–1365, May2007.

[26] *CRF++0.58*, 2013. [Online]. Available: http://taku910.github.io/crfpp/

[27] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 2553–2556.

[28] H. A. Chang and J. R. Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 4481–4484.

[29] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. Workshop Spontaneous Speech Process. Recognit.*, 2003, pp. 7–12.

[30] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous Japanese," in *Proc. Workshop Spontaneous Speech Process. Recognit.*, 2003, pp. 135–138.

[31] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4894–4897.

[32] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering for large vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 855–872, May 2006.

[33] S. Nakagawa and H. Takagi, "Statistical methods for comparing pattern recognition algorithms and comments on evaluating speech recognition performance," *J. Acoust. Soc. Jpn.*, vol. 50, no. 10, pp. 849–854, Oct. 1994.

[34] C. Zhang and P. Woodland, "A general artificial neural network extension for HTK," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3581–3585.

[35] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Workshop Autom. Speech Recognit. Understanding*, 2011.

[36] A. Ogawa and T. Hori, "ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4370–4374.

Atsunori Ogawa received the B.E. and M.E. degrees in information engineering, and the Ph.D. degree in information science from Nagoya University, Aichi, Japan, in 1996, 1998, and 2008, respectively. Since 1998, he has been with Nippon Telegraph and Telephone (NTT) Corporation. He has been engaged in researches on speech recognition and speech enhancement at NTT Cyber Space Laboratories and NTT Communication Science Laboratories. He is a Member of the International Speech Communication Association, Institute of Electronics, Information, and Communication Engineers, Information Processing Society of Japan, and Acoustical Society of Japan (ASJ). He received the ASJ Best Poster Presentation Awards in 2003 and 2006, respectively.

Takaaki Hori received the B.E. and M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan, in 1994 and 1996, respectively, and the Ph.D. degree in system and information engineering from Yamagata University in 1999. From 1999 to 2015, he had been engaged in researches on speech recognition and spoken language understanding at Cyber Space Laboratories and Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan. Since 2015, he has been a Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. He was a Visiting Scientist at the Massachusetts Institute of Technology from 2006 to 2007. He received the 22nd Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the IPSJ Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2012, and the 58th Maejima Hisoka Award from Tsushinbunka Association in 2013. He is a Member of the Institute of Electronics, Information, and Communication Engineers and the ASJ.

Atsushi Nakamura received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, Japan, in 1985, 1987, and 2001, respectively. In 1987, he joined Nippon Telegraph and Telephone Corporation (NTT), where he was engaged in the research and development of network service platforms, including studies on the application of speech processing technologies to network services, at Musashino Electrical Communication Laboratories, Tokyo, Japan. From 1994 to 2000, he was with the Advanced Telecommunications Research Institute, Kyoto, Japan, as a Senior Researcher, undertaking research on spontaneous speech recognition, the construction of spoken language databases, and the development of speech translation systems. From April 2000 to March 2014, he was with NTT Communication Science Laboratories, Kyoto, Japan. Since April 2014, he has been with the Graduate School of Natural Sciences, Nagoya City University, Aichi, Japan. His research interests include the acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and the application of learning theories to signal analysis, and modeling. He served as a Member of the IEEE Machine Learning for Signal Processing Technical Committee, and as the Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a Member of the Institute of Electronics, Information and Communication Engineering (IEICE) and the Acoustical Society of Japan. He received the IEICE Paper Award in 2004, and twice received the TELECOM System Technology Award of the Telecommunications Advancement Foundation, in 2006 and 2009.