INTERSPEECH 2009 BRIGHTON

Simultaneous Estimation of Confidence and Error Cause in Speech Recognition Using Discriminative Model

Atsunori Ogawa & Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

Abstract

Since recognition errors are unavoidable in speech recognition, confidence scoring, which accurately estimates the reliability of recognition results, is a critical function for speech recognition engines. In addition to achieving accurate confidence estimation, if we are to develop speech recognition systems that will be widely used by the public, speech recognition engines must be able to report the causes of errors properly, namely they must offer a reason for any failure to recognize input utterances. This paper proposes a method that simultaneously estimates both confidences and causes of errors in speech recognition results by using discriminative models. We evaluated the proposed method in an initial speech recognition experiment, and confirmed its promising performance with respect to confidence and error cause estimation.

Index Terms: speech recognition, confidence, error cause, discriminative model

10.21437/Interspeech.2009-347

1. Introduction

As a result of the significant progress that has been made on speech recognition technology, practical speech recognition systems have been developed for certain applications, e.g. the closed-captioning of news broadcasts [1] and recording of parliamentary minutes [2]. However, there are almost no systems that are widely used by the public.

The reason for this is mainly attributable to two problems [3, 4]. The first is that users are not familiar with the proper usage of speech recognition systems and cannot understand the behavior of the systems when they use them improperly. The second problem is that system developers do not know the proper way to use speech recognition engines or how to exploit their performance. To solve these problems, in [3, 4], a framework is proposed for developing speech recognition systems that involves close cooperation between engine builders, system developers and users.

To enable users to use speech recognition systems properly and to enable system developers to exploit the performance of speech recognition engines adequately, we focus on enhancing the functions of speech recognition engines. Since errors are unavoidable in speech recognition, confidence scoring, which accurately estimates the reliability of the recognition results, is a critical function for speech recognition engines. Many studies have attempted to develop good confidence measures [5]. In addition to achieving accurate confidence estimation, it is important for speech recognition engines to report the causes of errors properly, namely they must offer a reason for any failure to recognize input utterances (Section 2).

In this paper, we propose a method that simultaneously estimates both confidences and causes of errors in speech recognition results by using discriminative models. Error handling has been actively studied in the research area of spoken dialogue systems [6]. In these studies, confidence measures are used to detect the *occurrences* of errors in human-machine dialogues. In contrast, the proposed method not only detects the occurrences of errors but also estimates their *causes* directly. The estimation of confidences and error causes is formulated as a discrimination problem (Section 3). We evaluate the proposed method in an initial speech recognition experiment, and confirm its promising performance as regards confidence and error cause estimation (Section 4). The proposed method has the potential to be utilized in various research and development themes concerned with speech recognition (Sections 5 and 6).

2. Effects of error cause estimation

Here we consider a simple question-answering spoken dialogue task, namely "receive a question about the weather of a Japanese city uttered by an adult male in a clean environment and provide a forecast". Fig. 1 shows examples of two speech recognition systems that can execute this task. System A is based on a speech recognition engine that can estimate the confidences of recognition results. System B is based on an engine that can estimate both the confidences and error causes of recognition results. We use these examples to consider the effects of direct error cause estimation for users.

If users employ the systems properly (U0), the answers from both systems are probably the same. In this case, recognition results are obtained with high confidence scores in both systems, and users could obtain proper answers (A0 and B0). In contrast, if the users employ the systems improperly, recognition results are probably obtained with low confidence scores in both systems, and the answers from the two systems will differ. Improper uses include out-of-vocabulary (OOV) words are employed, e.g. the system is presented with unknown city names (U1), the system is used in noisy environments, e.g. in a moving vehicle (U2), or the system is used by unexpected users, e.g. females, children or the elderly (U3).

System A can ask users to repeat their utterances (AX). However, the users cannot understand the reason for this request. Or the system may remain silent (AY). In this case, the users cannot even know whether or not the system is working. However, even if the system provides a detailed technical explanation for rejecting the users' utterances (AZ), the users' only response will be silence (these technical explanations may be useful for system developers).

On the other hand, because system B can estimate the causes of recognition result errors, it can provide proper requests to users (B1, B2 and B3) according to the type of improper use (U1, U2 and U3). The users can then use the system properly in accordance with the system requests and thus the recognition performance of the system is improved.

3. Proposed method

We propose a conditional random fields (CRF) [7] based method for the simultaneous estimation of the confidences and error causes of speech recognition results. CRF is a discriminative model and has been applied to natural language processing tasks such as parsing, tagging and segmentation.

In the following, $\boldsymbol{x}_i$ denotes an input observation vector and $\boldsymbol{y}_i$ denotes an output label vector corresponding to $\boldsymbol{x}_i$. In addition, $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{L(\boldsymbol{X})}$ denotes a sequence of $\boldsymbol{x}_i$ of length $L(\boldsymbol{X})$ and $\boldsymbol{Y} = \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{L(\boldsymbol{X})}$ denotes a sequence of $\boldsymbol{y}_i$ corresponding to $\boldsymbol{X}$. We can obtain a conditional probability $p(\boldsymbol{Y}|\boldsymbol{X})$ by using CRF as follows:

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp \sum_{i=1}^{L(\boldsymbol{X})} \sum_{k=1}^{K} \lambda_k f_k(\boldsymbol{x}_i, \boldsymbol{y}_i)}{\sum_{\boldsymbol{Y}'} \exp \sum_{j=1}^{L(\boldsymbol{X})} \sum_{l=1}^{K} \lambda_l f_l(\boldsymbol{x}_j, \boldsymbol{y}_j{}')}, \quad (1)$$

where λ_k is the weight of the k-th feature function $f_k(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and K is the number of weights and feature functions. The set

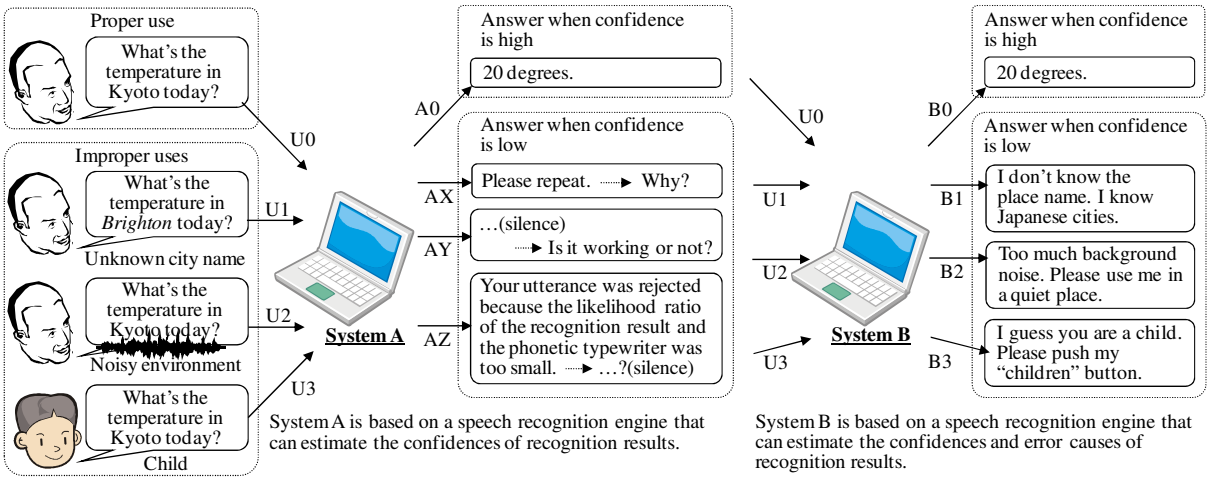

Figure 1: *Examples of the speech recognition systems to consider the effects of direct error cause estimation for users.*

Table 1: *Definition of the four elements in* $\boldsymbol{y}$.

y_d	Meaning when $y_d = 0$	Meaning when $y_d = 1$
y_0	The recognized word is correct.	The recognized word is incorrect.
y_1	The uttered word is IV.	The uttered word is OOV.
y_2	The system is used in a clean environment.	The system is used in a noisy environment.
y_3	The system is used by an adult male (expected gender).	The system is used by an adult female (unexpected gender).

Table 2: *Possible IDs which* $\boldsymbol{y}$ *takes and corresponding meanings (IDs are given by bit operations).*

ID of $\boldsymbol{y}$	y_0	y_1	y_2	y_3	Meaning of $\boldsymbol{y}$
0	0	0	0	0	There is no error cause that is focused on, and the recognized word is correct.
1	0	0	0	1	The system is used by an adult female, however the recognized word is correct.
2	0	0	1	0	The system is used in a noisy environment, however the recognized word is correct.
8	1	0	0	0	There is no error cause that is focused on, however the recognized word is incorrect.
9	1	0	0	1	The system is used by an adult female, therefore the recognized word is incorrect.
10	1	0	1	0	The system is used in a noisy environemt, therefore the recognized word is incorrect.
12	1	1	0	0	The uttered word is OOV, therefore the recognized word is incorrect.

of weights $\{\lambda_k\}_{k=1}^K$ is estimated by a quasi-Newton method such as L-BFGS [8] and a forward-backward algorithm using training data $\{(\boldsymbol{X}_m, \boldsymbol{Y}_m)\}_{m=1}^M$.

In the proposed method, as with conventional confidence estimation methods [5], we define $\boldsymbol{x}_i$ as a set of features that are related to a word in a recognized word sequence. Features concerned with a recognized word are, for example, the average acoustic likelihood, the average phoneme duration, the linguistic likelihood and the posterior probability. These features are collected in the main recognition process and certain additional processes.

The definition of $\boldsymbol{y}_i$ is the key feature of the proposed method. We define $\boldsymbol{y}_i$ as a D-dimensional vector in which each element $y_d (d = 0, 1, \cdots, D-1)$ takes binary digits 0 or 1. One of the elements, y_0 for convenience, denotes that the recognized word is *correct* ($y_0 = 0$) or *incorrect* ($y_0 = 1$). And each remaining $D-1$ element $y_d (d = 1, 2, \cdots, D-1)$ denotes the *nonexistence* ($y_d = 0$) or *existence* ($y_d = 1$) of each of $D-1$ error cause on which we focus. For example, if we define y_1 as an element that is related to *OOV*, $y_1 = 0$ denotes *in-vocabulary (IV)* and $y_1 = 1$ denotes *OOV*. Conventional confidence estimation methods [5] estimate only the reliability of the recognized word (i.e. the value of y_0 in the above definition). In contrast, based on the above definition of $\boldsymbol{y}_i$, the proposed method can estimate not only the reliability of the recognized word but also its error causes *simultaneously*.

In Section 4, we evaluate the proposed method in an initial isolated word utterance recognition experiment. For the experiment, because $L(\boldsymbol{x}) = 1$, we shrink $\boldsymbol{X}$ and $\boldsymbol{Y}$ to $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Accordingly, we shrink Eq. (1) as follows:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_{k=1}^K \lambda_k f_k(\boldsymbol{x}, \boldsymbol{y})}{\sum_{\boldsymbol{y}'} \exp \sum_{l=1}^K \lambda_l f_l(\boldsymbol{x}, \boldsymbol{y}')}. \quad (2)$$

Eq. (2) is a conditional probability obtained by using a maximum entropy model (MaxEnt) [9], which is also a discriminative model. Thus, we evaluate the proposed method based on MaxEnt in the next section.

4. Speech recognition experiment

We assume a simple speech recognition system that "recognizes a Japanese city name uttered by an adult male in a clean environment". We evaluate the proposed method based on this system.

4.1. Definition of $\boldsymbol{y}$

We focus on three error causes; OOV, use in a noisy environment and use by an adult female (unexpected gender). Accordingly, the four elements in $\boldsymbol{y}$ are defined as shown in Table 1. In addition, we assume that there is no more than one error cause at a time. Thus, possible IDs taken by $\boldsymbol{y}$ and corresponding meanings are defined as shown in Table 2.

As shown in Table 2, the recognized word could be correct even if the system is used by adult females (ID=1) or in noisy environments (ID=2). However, obviously, the recognized word cannot be correct if an OOV word is uttered (ID=4 is impossible).

The case ID=8 has the potential to play an important role. Here we focus on only three error causes. However, we expect that, by defining ID=8, we will be able to cover other error causes, such as too large or too small a volume of utterance or a user with unexpected age (e.g. a child or an elderly person). Because ID=8 says that "there is no error cause that is focused on, however the recognized word is incorrect" in Table 2, in other words, "there may be some error causes that are not focused on, therefore the recognized word is incorrect".

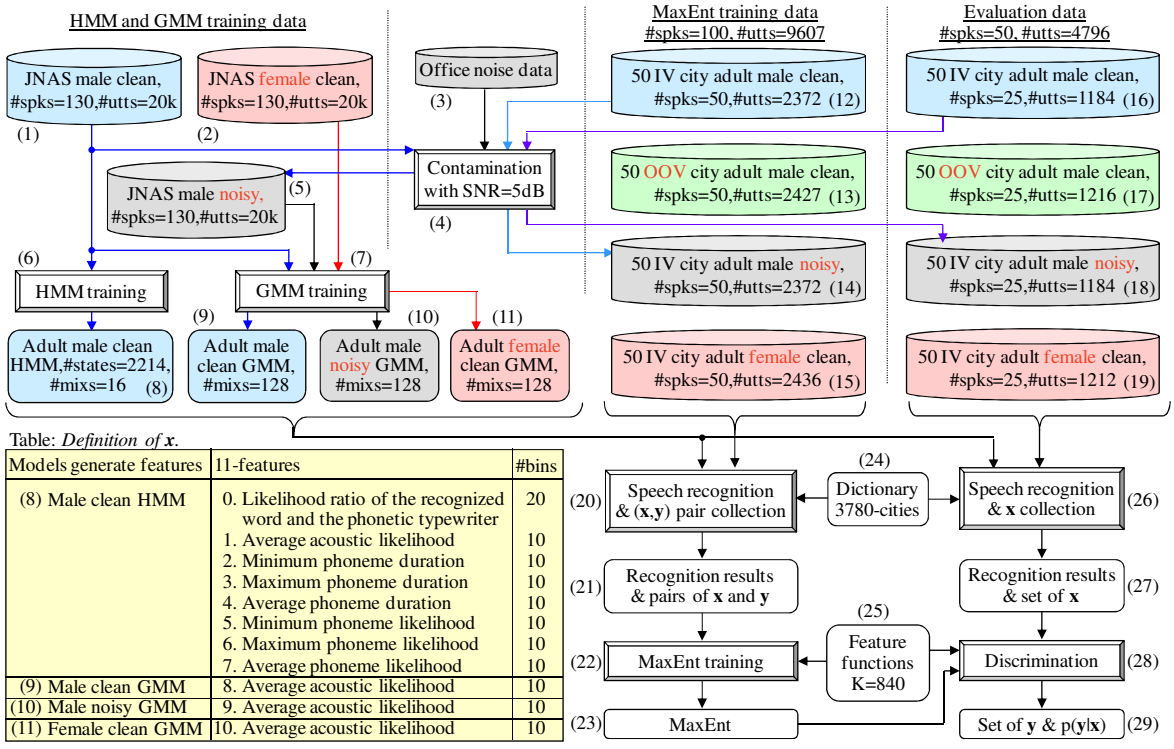

Table: *Definition of* $\boldsymbol{x}$.

Models generate features	11-features	#bins
(8) Male clean HMM	0. Likelihood ratio of the recognized word and the phonetic typewriter	20
	1. Average acoustic likelihood	10
	2. Minimum phoneme duration	10
	3. Maximum phoneme duration	10
	4. Average phoneme duration	10
	5. Minimum phoneme likelihood	10
	6. Maximum phoneme likelihood	10
	7. Average phoneme likelihood	10
(9) Male clean GMM	8. Average acoustic likelihood	10
(10) Male noisy GMM	9. Average acoustic likelihood	10
(11) Female clean GMM	10. Average acoustic likelihood	10

Figure 2: *Procedures of speech recognition experiment.*

4.2. Experimental procedure

Fig. 2 shows the experimental procedures. We trained a hidden Markov model (HMM)-based acoustic model and Gaussian mixture models (GMMs), which we used for the speech recognition and $(\boldsymbol{x},\boldsymbol{y})$ pair collection process. An adult male clean speech HMM-based acoustic model (8) was trained (6) using adult male clean speech data (1) from the Japanese Newspaper Article Sentences (JNAS) speech corpus [10]. The adult male clean speech data (1) were contaminated (4) by adding office noise data (3) with a signal to noise ratio (SNR) of 5 dB and became adult male noisy speech data (5). Then, using adult male clean speech data (1), adult male noisy speech data (5) and adult female clean speech data (2), we trained (7) an adult male clean GMM (9), an adult male noisy GMM (10), and an adult female clean GMM (11), respectively.

We used 100 Japanese city name speech data from The Japan Electronic Industry Development Association's Common Speech Data (JCSD) corpus (LDC96S64) [11] to train a MaxEnt and to evaluate the proposed method based on the MaxEnt. 75 adult males and 75 adult females were divided into groups of 50 adult males and 50 adult females for the MaxEnt training and 25 adult males and 25 adult females for the evaluation. 100 city names were divided into 50 IV city names and 50 OOV city names. Based on the above divisions, we prepared 50 IV city adult male clean speech data (12 and 16), 50 OOV city adult male clean speech data (13 and 17) and 50 IV city adult female clean speech data (15 and 19) for the MaxEnt training and evaluation purposes. In addition, 50 IV city adult male clean speech data (12 and 16) were contaminated (4) by adding office noise data (3) with an SNR of 5 dB and became 50 IV city adult male noisy speech data (14 and 18). As a result, four sets of 100 city name speech data were prepared for the MaxEnt training (12, 13, 14 and 15) and evaluation (16, 17, 18 and 19) purposes. And by adding 3730 dummy city names, which were different from the 50 OOV city names, to the 50 IV city names, a 3780 city name dictionary (24) was created for speech recognition.

Using the HMM (8), three GMMs (9, 10 and 11) and the 3780 city name dictionary (24), we performed speech recognition and $(\boldsymbol{x},\boldsymbol{y})$ pair collection procedures (20) for the four sets of 100 city name speech data for the MaxEnt training (12, 13, 14 and 15), and obtained speech recognition results and the pairs of $\boldsymbol{x}$ and $\boldsymbol{y}$; $\{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{9607}$ (21). We defined $\boldsymbol{x}$ as the 11-dimensional feature vector shown in the table in the bottom left of Fig. 2. The 0-th feature "likelihood ratio of the recognized word and the phonetic typewriter" is a conventional confidence measure [12]. With reference to [13, 14], we defined 840 feature functions (25), which executed the nonlinear quantization of each element in $\boldsymbol{x}$ with the number of bins shown on the right-hand side of the table. We used these feature functions (25) and the pair data $\{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{9607}$ (21) to estimate (22) the set of weights $\{\lambda_k\}_{k=1}^{840}$ of the MaxEnt (23).

In the evaluation, we used the HMM (8), three GMMs (9, 10 and 11) and the 3780 city name dictionary (24) to perform speech recognition and $\boldsymbol{x}$ collection procedures (26) for the four sets of 100 city name speech data for the evaluation (16, 17, 18 and 19), and obtained speech recognition results and $\{\boldsymbol{x}_n\}_{n=1}^{4796}$ (27). Then using $\{\boldsymbol{x}_n\}_{n=1}^{4796}$ (27), we obtained the corresponding $\{\boldsymbol{y}_n\}_{n=1}^{4796}$ with conditional probabilities $\{p(\boldsymbol{y}_n|\boldsymbol{x}_n)\}_{n=1}^{4796}$ (29) by using the discrimination processes (28) of the MaxEnt (23).

We trained the HMM (8) and three GMMs (9, 10 and 11) and performed the speech recognition by using *SOLON* [15]. As described above, this time there were closed conditions, namely that the office noise data (3) used for contamination (4) and the 50 OOV city names were the same in the MaxEnt training data (13 and 14) and the evaluation data (17 and 18).

4.3. Experimental result

The left hand side of Fig. 3 shows the receiver operator characteristics (ROC) curve of the confidence estimation accuracy obtained with the proposed method for all the evaluation data. When plotting this ROC curve, we focused solely on the value of y_0, i.e. *correct* ($y_0 = 0$) or *incorrect* ($y_0 = 1$) and ignored the values of the remaining three elements $y_d (d = 1, 2, 3)$, which denote the *nonexistence* ($y_d = 0$) or *existence* ($y_d = 1$) of the three error causes. This ROC curve is plotted by varying the acceptance / rejection threshold continuously from 0.0 to 1.0 and comparing it with the conditional probabilities of $\boldsymbol{y}$ (29). For comparison, we also plot the confidence estimation accuracy of the 0-th element of $\boldsymbol{x}$, i.e. the conventional confidence measure "likelihood ratio of the recognized word and the phonetic typewriter [12]" in Fig. 3 (left). It is clear that the con-

Table 3: *Confidence and error cause estimation EERs with the phonetic typewriter based method and the proposed method.*

Set of the evaluation data	Word Recog. Rate [%]	Phonetic typewriter	Proposed	
		Conf. est. EER [%]	Conf. est. EER [%]	Err. cause est. EER [%]
All	52.48	28.93	22.88	25.59
(16) 50 IV city adult male clean	86.06	26.76	26.61	51.74
(17) 50 *OOV* city adult male clean	0.00	—	—	14.71
(18) 50 IV city adult male *noisy*	68.07	26.32	23.97	32.32
(19) 50 IV city adult *female* clean	57.10	40.25	34.00	29.25

fidence estimation accuracy of the proposed method is higher than that provided by the phonetic typewriter based confidence estimation method. The absolute improvement is about 6% in terms of the equal error rate (EER).

Table 3 shows confidence estimation EERs of the phonetic typewriter based method and the proposed method for all of the evaluation data and the four sets of the evaluation data (16, 17, 18 and 19). Word recognition rates are also shown. It is impossible to calculate the EER for the 50 OOV city adult male clean speech data (17) because it is impossible to calculate the false rejection rate for this data. This table confirms that the improvement in the confidence estimation accuracy provided by the proposed method compared with that of the phonetic typewriter based method is mainly obtained for the 50 IV city adult female clean speech data (19). We assume that this improvement is obtained as a result of the effect of the female GMM (11) used in the proposed method.

The right hand side of Fig. 3 shows the ROC curve of the error cause estimation accuracy obtained with the proposed method for the evaluation data whose recognition results are estimated as *incorrect* ($y_0 = 1$) by the proposed method itself. This ROC curve is also plotted by using the same thresholding procedures used for plotting the ROC curves in Fig. 3 (left). As shown in this figure, we obtained an EER of about 26% in terms of error estimation accuracy.

Table 3 shows the error cause estimation EERs of the proposed method for the evaluation data whose recognition results are estimated as *incorrect* ($y_0 = 1$) by the proposed method, and its four sets corresponding to (16), (17), (18) and (19). This table confirms that error cause estimation for the 50 IV city adult male clean speech data (16), i.e. the estimation of $\boldsymbol{y}$ when it takes ID=8 in Table 2, is very difficult. This is because the amount of training data when $\boldsymbol{y}$ takes ID=8 was very small and the MaxEnt was not well trained to estimate this case. However, as described in Section 4.1, ID=8 has the potential to play an important role, thus we have to improve the estimation accuracy. We also confirmed that the error cause estimation accuracy for the 50 OOV city adult male clean speech data (17) is very high. However, this high accuracy is obtained because of the closed condition whereby the 50 OOV city names are the same in the MaxEnt training data (13) and the evaluation data (17) as described in Section 4.2. It has been reported that OOV word detection is an essentially more difficult task [5].

5. Relationship with other work

The proposed method can be categorized with recently reported confidence estimation methods, e.g. [13, 14], which are based on discriminative models and use many features of the recognition results.

The proposed method can be incorporated in spoken dialogue systems [6] and used to develop more sophisticated dialogue strategies.

The proposed method is also related to *analyses* of speech recognition errors, e.g. [16, 17, 18]. In these studies, based on the assumption that misrecognitions are mainly attributable to the weaknesses of the speech recognition engines, detailed analyses are conducted mainly for utterances that should be correctly recognized but that are misrecognized. In contrast, the proposed method does not *analyze* but *detects* error causes for any input utterances assuming that misrecognitions are attributable to both the weaknesses of the speech recognition engine and improper operation by users (in the system construction stage, failures by system developers could cause errors).

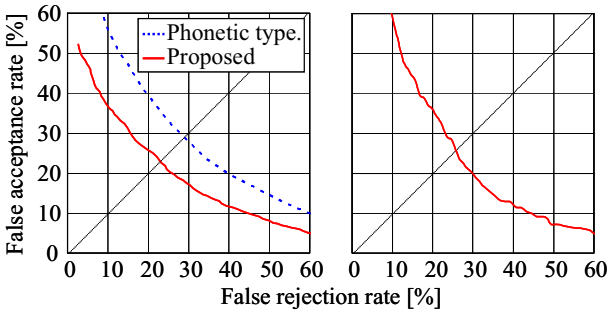

Figure 3: *(Left) Confidence estimation ROC curves of the phonetic typewriter based method and the proposed method. (Right) Error cause estimation ROC curve of the proposed method.*

Therefore, it is expected that error cause *detection* with the proposed method could constitute the preprocessing of the detailed error *analysis* of the existing error analysis methods.

6. Conclusion and future work

We proposed a method for simultaneously estimating the confidences and error causes of speech recognition results using discriminative models, and confirmed its promising performance in an initial speech recognition experiment.

We are planning to undertake the following studies: First, we will fix the closed experimental conditions with respect to the office noise data and 50 OOV city names described in Section 4.2. Second, to relax the limited assumption described in Section 4.1, we will assume the condition that there could be more than two error causes at a time. Under this condition, the estimation procedure is formulated as a multi-category labeling problem [19]. Third, we will improve the estimation accuracy of the proposed method by devising better estimation strategies. For example, we can prepare individual discriminative models for confidence estimation and each error cause estimation respectively, and merge their estimation results to form a final estimation result. Finally, by using the proposed method, we will develop a discriminative training technique for acoustic models [20] that explicitly takes the error causes in the competing recognition hypotheses into account.

7. References

[1] T. Imai et al., Proc. Interspeech, pp. 1602–1605, 2006.
[2] L. Lamel et al., Proc. ICASSP, pp. 997–1000, 2007.
[3] T. Nakano et al., Proc. IEEE ASRU, pp. 601–606, 2007.
[4] T. Kobayashi, IPSJ SIG Tech. Rep. 2008-SLP-74(19).
[5] H. Jiang, Speech Communication, vol. 45, pp. 455–470, 2005.
[6] Editorial, Speech Communication, vol. 45, pp. 207–209, 2005.
[7] J. Lafferty et al., Proc. ICML, pp. 282–289, 2001.
[8] D.C. Liu et al., Math. Prog., vol. 45, pp. 503–528, 1989.
[9] A.L. Berger et al., Comput. Ling., vol. 22, no. 1, 1996.
[10] K. Itou et al. Proc. ICSLP, pp. 3261–3264, 1998.
[11] LDC web site, http://www.ldc.upenn.edu/
[12] L. Jiang et al., Proc. ICSLP, paper 0625, 1998.
[13] C. White et al., ICASSP, pp. 809–812, 2007.
[14] A. Kobayashi et al., Proc. Interspeech, pp. 1453–1456, 2005.
[15] T. Hori et al., IEEE Trans. ASLP, vol. 15, no. 4, 2007.
[16] H. Nanjo et al., Proc. Interspeech, pp. 1027–1030, 2000.
[17] S. Furui et al., Proc. 8th Int. Conf. on TSD, pp. 9–22, 2005.
[18] N. Duta et al., IEEE Trans. ASLP, vol. 14, no. 5, Sept. 2006.
[19] N. Ueda et al., Proc. NIPS, pp. 721–728, 2002.
[20] E. McDermott et al., Proc. Interspeech, pp. 2398–2401, 2008.