



A Novel Confidence Measure Based on Marginalization of Jointly Estimated Error Cause Probabilities

Atsunori Ogawa & Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation

Abstract

We propose a novel confidence measure based on the marginalization of jointly estimated error cause probabilities. Conventional confidence measures directly score the reliability of recognition results. In contrast, our method first calculates *joint* confidence and *error cause* probabilities and then sums them with respect to the error cause patterns to obtain the *marginal* confidence probability. We show experimentally that, the confidence estimation accuracy obtained with the proposed method is significantly improved compared with that obtained with the conventional confidence measure.

Index Terms: speech recognition, confidence measure, cause of error, joint estimation, marginalization

1. Introduction

Recently, speech recognition systems that assume the general public as the users, e.g. voice search applications [1, 2], have been actively developed. However, so far, most people are unfamiliar with these systems and they do not know well the proper way to use the systems. In this situation, confidence estimation [3, 4] that scores the reliability of recognition results has become more important function for speech recognition engines.

We think that, as well as realizing further improvements in confidence estimation accuracy, we need additional extended functions for speech recognition engines if we are to develop speech recognition systems that are more familiar to the general public. And as the first step in extending the functions of speech recognition engines, we have proposed a method for estimating the confidence of recognition results while *jointly estimating the causes of recognition errors* that affect the reliability of recognition results (e.g. out-of-vocabulary (OOV) words and noisy environments) using a discriminative model, and showed its potential experimentally [5, 6]. Error cause estimation will allow users to employ speech recognition systems properly.

Recently proposed confidence measures that show good performance, e.g. [4], receive many features related to the decoded word and score the reliability of it by using discriminative models, e.g. a maximum entropy model (MaxEnt) [7] and a conditional random fields (CRF) [8]. Our joint confidence and error cause estimation method is also based on this framework. The confidence of the decoded word is naturally affected by error causes [3, 4]. And the joint estimation method directly captures the *co-occurrence* of the correctly/incorrectly decoded word and the nonexistence/existence of error causes. Thus we expected that our method could improve the confidence estimation accuracy with the help of the error cause estimation. However, in our previous experiments [6], its confidence estimation accuracy was no better than that obtained with the conventional confidence measure.

In this paper, we modify the confidence estimation procedure in our joint estimation method so that the confidence is computed via *marginalization* of joint confidence and error cause probabilities over the error cause patterns, and show

its significantly improved confidence estimation accuracy in speech recognition experiments.

2. Proposed methods

In this section, we describe our joint and separate confidence and error cause estimation methods [5, 6]. We modify the confidence estimation procedure in the joint estimation method with marginalization. The separate estimation method involves the conventional confidence measure as a component.

2.1. Joint estimation method

Our joint confidence and error cause estimation method is based on a CRF [8]. In the following, \mathbf{x}_i denotes an input observation vector and \mathbf{y}_i denotes an output label vector corresponding to \mathbf{x}_i . In addition, $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{L(\mathbf{X})}$ denotes a sequence of \mathbf{x}_i of length $L(\mathbf{X})$, and $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{L(\mathbf{X})}$ denotes a sequence of \mathbf{y}_i corresponding to \mathbf{X} . We can obtain a conditional probability $p(\mathbf{Y}|\mathbf{X})$ by using a CRF as follows:

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left(\sum_{i=1}^{L(\mathbf{X})} \sum_{k=1}^K \lambda_k f_k(\mathbf{x}_i, \mathbf{y}_i) \right), \quad (1)$$

where λ_k is the weight of the k -th feature function $f_k(\mathbf{x}_i, \mathbf{y}_i)$, K is the number of feature functions, and $Z(\mathbf{X})$ is a normalization term. The set of weights $\{\lambda_k\}_{k=1}^K$ is estimated by a quasi-Newton method and a forward-backward algorithm using a large number N of (\mathbf{X}, \mathbf{Y}) pairs, i.e. $\{(\mathbf{X}_n, \mathbf{Y}_n)\}_{n=1}^N$ as the training data.

With the proposed method, as with conventional confidence measures [3, 4], we define \mathbf{x}_i as a set of features that are related to a word w_i in a decoded word sequence $\mathbf{W} = w_1, w_2, \dots, w_{L(\mathbf{X})}$. Features related to a decoded word are, for example, the frame normalized acoustic likelihood, the average phoneme duration, the linguistic prior probability, and the posterior probability. These features are obtained by the main speech recognition process and certain additional processes.

The definition of \mathbf{y}_i is the key feature of the proposed method. We define \mathbf{y}_i as a D -dimensional *joint* confidence and error cause vector $\mathbf{y}_i = (y_{0,i}, y_{1,i}, \dots, y_{D-1,i})^T$ in which each element $y_{d,i}$ ($d = 0, 1, \dots, D - 1$) takes a binary digit 0 or 1. The first element $y_{0,i}$ denotes the binary confidence of the decoded word w_i , i.e. w_i is *correct* ($y_{0,i} = 0$) or *incorrect* ($y_{0,i} = 1$). The remaining $D - 1$ elements constitute an error cause vector $\mathbf{y}_{e,i} = (y_{1,i}, y_{1,i}, \dots, y_{D-1,i})^T$ (i.e. $\mathbf{y}_i = (y_{0,i}, \mathbf{y}_{e,i}^T)^T$) that represents the $D - 1$ error cause types of interest and their combinations. An element $y_{d,i}$ ($d = 1, 2, \dots, D - 1$) in $\mathbf{y}_{e,i}$ denotes that an error cause *does not exist* ($y_{d,i} = 0$) or *exists* ($y_{d,i} = 1$). For example, if we define $y_{1,i}$ as an element that is related to OOV words, $y_{1,i} = 0$ denotes that w_i is an *in-vocabulary (IV)* word and $y_{1,i} = 1$ denotes that the uttered word is *OOV*. Conventional confidence

measures [3, 4] estimate only the reliability of the decoded word (i.e. the value of $y_{0,i}$). In contrast, based on the above definition of \mathbf{y}_i , the proposed method can *jointly* estimate the confidence of the decoded word and its error causes.

As a result, a *joint* confidence and error cause sequence is obtained as $\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} p(\mathbf{Y}|\mathbf{X})$ with the conditional probability $p(\hat{\mathbf{Y}}|\mathbf{X})$. The confidence of the decoded word is naturally affected by error causes [3, 4]. Therefore, this joint estimation is a reasonable method since it directly models *co-occurrence* of the correctly/incorrectly decoded word and the nonexistence/existence of error causes. Hereafter, we refer to this method as ‘‘JNT’’. A graphical representation of its estimation procedure is shown on the left in Fig. 1.

$\hat{\mathbf{Y}}$ can be decomposed into a confidence and $D - 1$ error cause sequences $\hat{\mathbf{Y}} = (\hat{\mathbf{Y}}_0, \hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_{D-1})^\top$. In our previous experiments [5, 6], we simply extracted the confidence sequence $\hat{\mathbf{Y}}_0$ from $\hat{\mathbf{Y}}$ with the approximated conditional probability $p(\hat{\mathbf{Y}}_0|\mathbf{X}) \approx p(\hat{\mathbf{Y}}|\mathbf{X})$. However, with this procedure, the confidence estimation accuracy was no better than that obtained with the conventional confidence measure [6].

To improve the confidence sequence estimation accuracy, we have to obtain the exact confidence sequence probability rather than its approximated value. To this end, we calculate the *marginal* confidence sequence probability $p_m(\mathbf{Y}_0|\mathbf{X})$ by summing the joint probabilities of the binary confidence $y_{0,i}$ and error cause vector $\mathbf{y}_{e,i}$ with respect to $\mathbf{y}_{e,i}$ word by word as follows:

$$p_m(\mathbf{Y}_0|\mathbf{X}) = \prod_{i=1}^{L(\mathbf{x})} \sum_{\mathbf{y}_{e,i}} p(\mathbf{y}_i|\mathbf{X}) = \prod_{i=1}^{L(\mathbf{x})} \sum_{\mathbf{y}_{e,i}} p(y_{0,i}, \mathbf{y}_{e,i}|\mathbf{X}), \quad (2)$$

where $p(\mathbf{y}_i|\mathbf{X}) (= p(y_{0,i}, \mathbf{y}_{e,i}|\mathbf{X}))$ is the joint confidence and error cause probability for the decoded word w_i given an input observation vector sequence \mathbf{X} . From Eq. (2), we can obtain the confidence sequence as $\tilde{\mathbf{Y}}_0 = \arg \max_{\mathbf{Y}_0} p_m(\mathbf{Y}_0|\mathbf{X})$ with the conditional probability $p_m(\tilde{\mathbf{Y}}_0|\mathbf{X})$. We can employ the same marginalization procedure to obtain each error cause sequence as shown on the left in Fig. 1.

In JNT, the number of classes of \mathbf{y}_i to be discriminated word by word is 2^D (this value can be reduced by taking account of the fact that OOV word utterances can never be correctly decoded). Therefore, when D is large, the classification problem becomes difficult, and so, the estimation accuracy of $\hat{\mathbf{Y}}$ might degrade.

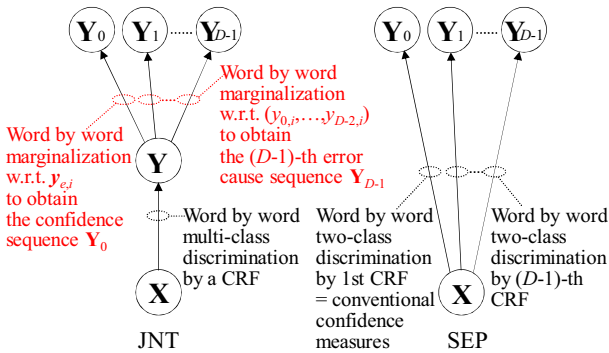


Figure 1: Graphical representations of proposed methods.

2.2. Separate estimation method

To alleviate the classification problem in JNT, we have proposed a *separate* estimation method [6]. With this method, we estimate a confidence and $D - 1$ error cause sequences *separately* with each of the D -CRFs that provide the following conditional probability:

$$p(\mathbf{Y}_d|\mathbf{X}) = \frac{1}{Z_d(\mathbf{X})} \exp \left(\sum_{i=1}^{L(\mathbf{x})} \sum_{k=1}^{K_d} \lambda_{d,k} f_{d,k}(\mathbf{x}_i, y_{d,i}) \right), \quad (3)$$

for $d = 0, 1, \dots, D - 1$, where $\{\lambda_{d,k}, f_{d,k}(\mathbf{x}_i, y_{d,i}), K_d, Z_d(\mathbf{X})\}$ have the same meanings as those in Eq. (1). Then, a confidence and $D - 1$ error cause sequences are obtained as $\hat{\mathbf{Y}}_d = \arg \max_{\mathbf{Y}_d} p(\mathbf{Y}_d|\mathbf{X})$ with the conditional probabilities $p(\hat{\mathbf{Y}}_d|\mathbf{X})$ ($d = 0, 1, \dots, D - 1$). Hereafter, we refer to this separate estimation method as ‘‘SEP’’. A graphical representation of its estimation procedure is shown on the right in Fig. 1.

With SEP, each CRF solves the two-class discrimination problem word by word and this is easier than the multi-class discrimination problem solved by a CRF in JNT. Therefore, we can expect to obtain each sequence $\hat{\mathbf{Y}}_d$ ($d = 0, 1, \dots, D - 1$) accurately. The confidence sequence $\hat{\mathbf{Y}}_0$ estimation procedure in SEP is equivalent to recently proposed confidence measures, e.g. [4], which are based on discriminative models using many features of the decoded word.

3. Experiments

We assumed a simple speech recognition task that ‘‘recognizes an *isolated-word* uttered by a *male* in a *clean environment*’’, and evaluated the two proposed methods described in Section 2 with respect to the confidence and each error cause estimation accuracies, respectively. All the experiments were performed with our speech recognition platform *SOLOON* [9].

3.1. Simplifications

For the isolated-word speech recognition task, since $L(\mathbf{x}) = 1$, we shrunk \mathbf{X} and \mathbf{Y} in Eqs. (1) and (2) to \mathbf{x} and \mathbf{y} (indices i are also omitted) and obtain:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}, \mathbf{y}) \right), \quad (4)$$

$$p_m(y_0|\mathbf{x}) = \sum_{\mathbf{y}_e} p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{y}_e} p(y_0, \mathbf{y}_e|\mathbf{x}). \quad (5)$$

Equation (4) is the conditional probability of JNT provided by a MaxEnt [7]. Applying the same shrinking procedure to Eq. (3), we can obtain the MaxEnt-based conditional probability for SEP. These simplifications do not undermine the generalities of the proposed methods.

3.2. Definition and meanings of \mathbf{y}

We focused on three error causes; *OOV word utterances*, *use by a female (gender mismatch)* and *use in a noisy environment*. Accordingly, the four elements (y_0, y_1, y_2, y_3) in \mathbf{y} are defined as shown in Table 1. In our previous experiments [5, 6], we restricted the experimental condition by assuming that ‘‘there was no more than one error cause at a time’’. In this paper, we lifted this restriction and considered the experimental condition that ‘‘there could be multiple error causes at a time’’. Thus, possible class indices taken by \mathbf{y} ($\text{CID}(\mathbf{y})$) and their corresponding meanings are defined as shown in Table 2. As shown in Table

Table 1: Definition of the four elements in \mathbf{y} .

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

Table 2: Possible class indices taken by \mathbf{y} ($CID(\mathbf{y})$ given by bit operations) and their corresponding meanings.

y_0	y_1	y_2	y_3	$CID(\mathbf{y})$	Meaning of \mathbf{y}
0	0	0	0	0	There is no error cause of interest, and the decoded word is correct.
0	0	0	1	1	The system is used in a noisy environment, however the decoded word is correct.
0	0	1	0	2	The system is used by a female, however the decoded word is correct.
0	0	1	1	3	The system is used by a female in a noisy environment, however the decoded word is correct.
1	0	0	0	8	There is no error cause of interest, however the decoded word is incorrect.
1	0	0	1	9	The system is used in a noisy environment, therefore the decoded word is incorrect.
1	0	1	0	10	The system is used by a female, therefore the decoded word is incorrect.
1	0	1	1	11	The system is used by a female in a noisy environment, therefore the decoded word is incorrect.
1	1	0	0	12	An OOV word is uttered, therefore the decoded word is incorrect.
1	1	0	1	13	An OOV word is uttered in a noisy environment, therefore the decoded word is incorrect.
1	1	1	0	14	An OOV word is uttered by a female, therefore the decoded word is incorrect.
1	1	1	1	15	An OOV word is uttered by a female in a noisy environment, therefore the decoded word is incorrect.

2, the decoded word could be correct even if the system is used by females and/or in noisy environments, i.e. $CID(\mathbf{y}) = 1, 2, 3$ are possible. However, it is obvious that the decoded words can never be correct if OOV word utterances are input, i.e. $CID(\mathbf{y}) = 4, 5, 6, 7$ are impossible. The case $CID(\mathbf{y}) = 8$ plays an important role. Here we focus on only three error causes. However, by defining $CID(\mathbf{y}) = 8$, we can cover all other error causes, such as utterance volumes that are too large or too small or utterance speeds that are too fast or too slow, since we can paraphrase $CID(\mathbf{y}) = 8$ as “there may be some error causes that are not focused on, therefore the decoded word is incorrect”.

3.3. Data and experimental procedure

We prepared isolated-word speech data consisting of eight subsets for the MaxEnt training and eight subsets for the evaluation as shown in Table 3. These data subsets correspond to the combinations in the nonexistence/existence of the three error causes of interest (i.e. $y_d = 0/1$ for $d = 1, 2, 3$). The OOV words, speakers, and the noise conditions of the evaluation data are different from those of the training data. Some speakers uttered in the multiple subsets.

Figure 2 shows the experimental procedure. We prepared a hidden Markov model (HMM)-based male clean acoustic model, four Gaussian mixture models (GMMs); a male clean GMM, a male noisy GMM, a female clean GMM and a female noisy GMM, and a 3830 word dictionary (3). In the following, we explain the JNT procedure, which is essentially the same as the SEP procedure.

Using the HMM, four GMMs and the 3830 word dictionary (3) we performed speech recognition and (\mathbf{x}, \mathbf{y}) pair collection (4) for the eight subsets of the MaxEnt training data (1), and obtained the decoded words and pairs consisting of the input feature vector \mathbf{x} and the reference joint confidence and error cause vector \mathbf{y} ; $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{55632}$ (5). We defined \mathbf{x} as an 18-dimension feature vector comprising the frame normalized HMM/GMM likelihoods, the average phoneme duration, the posterior probability, and so on. For the MaxEnt, we defined $K = 1080$ feature functions (for the SEP MaxEnts, $K_d = 180$ for $d = 0, 1, 2, 3$) (6). We used these feature functions (6) and the pair data (5) to estimate (7) the set of weights $\{\lambda_k\}_{k=1}^{1080}$ of the MaxEnt (8).

In the evaluation, we again used the HMM, four GMMs and the 3830 word dictionary (3) to perform speech recog-

nition and \mathbf{x} collection (9) for the eight subsets of the evaluation data (2) and obtained the decoded words and the input feature vector set $\{\mathbf{x}_r\}_{r=1}^{18612}$ (10). The speech recognition rate for each of the evaluation data subsets and that averaged over all subsets are shown on the right in Table 3. Then using $\{\mathbf{x}_r\}_{r=1}^{18612}$ (10), we obtained the corresponding joint confidence and error cause vector set with the conditional probabilities $\{\mathbf{y}_r, p(\mathbf{y}_r|\mathbf{x}_r)\}_{r=1}^{18612}$ (11) by the discrimination processes (11) of the MaxEnt (8) defined by Eq. (4). Applying the marginalization procedure (13) defined by Eq. (5) to each of $\{\mathbf{y}_r, p(\mathbf{y}_r|\mathbf{x}_r)\}_{r=1}^{18612}$ (11), we obtained the estimation result sets of the confidence and each error cause with the marginal conditional probabilities $\{y_{d,r}, p_m(y_{d,r}|\mathbf{x}_r)\}_{r=1}^{18612}$

Table 3: MaxEnt training data and evaluation data (#spks = # of speakers, #utts = # of utterances, RR = recognition rate [%]).

Combination of the three error causes	Training		Evaluation		
	#spks	#utts	#spks	#utts	RR
IV male clean	50	4799	25	2400	95.58
IV male noisy	50	4799	25	2400	63.29
IV female clean	50	4860	25	2416	63.20
IV female noisy	50	4860	25	2416	28.02
OOV male clean	55	9079	27	2243	0.00
OOV male noisy	55	9079	27	2243	0.00
OOV female clean	55	9078	27	2247	0.00
OOV female noisy	55	9078	27	2247	0.00
Total	210	55632	104	18612	32.33

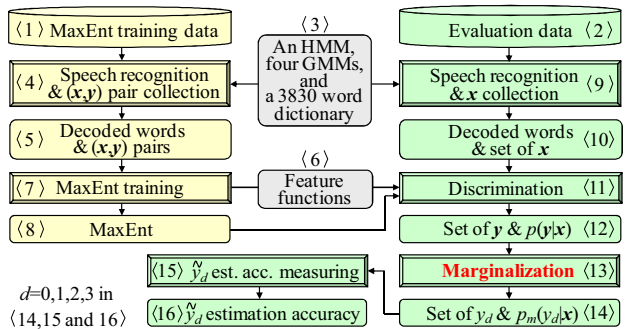


Figure 2: Experimental procedure (yellow components represent the MaxEnt training procedure, and green components represent the evaluation procedure).

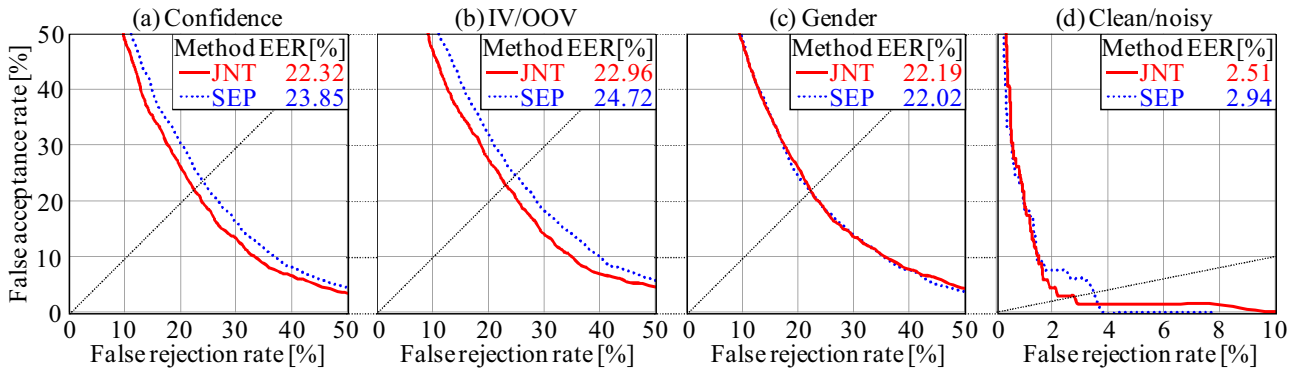


Figure 3: ROC curves for (a) confidence (correct/incorrect), (b) IV/OOV, (c) gender (male/female), and (d) clean/noisy environment estimation accuracies with JNT and SEP. EERs are also shown.

($d = 0, 1, 2, 3$) (14). Finally, from these estimation result sets we measured (15) the confidence \hat{y}_0 and each error cause \hat{y}_d ($d = 1, 2, 3$) estimation accuracies (16), respectively (in the SEP procedure, we repeated all the steps until (12) described above to obtain the confidence \hat{y}_0 and each error cause \hat{y}_d ($d = 1, 2, 3$) estimation results, respectively, and measured the accuracy for each of them).

3.4. Experimental results

Figure 3 shows the receiver operator characteristics (ROC) curves for the confidence and each of the three error cause estimation accuracies obtained with JNT and SEP. Equal error rates (EERs) are also shown. These ROC curves are plotted by varying the acceptance/rejection threshold continuously from 0.0 to 1.0 and comparing it with the conditional probabilities attached to the estimation results.

As shown in Fig. 3 (a), the confidence estimation accuracy obtained with JNT is higher than that obtained with SEP, which is, as described in Section 2.2, equivalent to recently proposed confidence measures, e.g. [4]. The EER reduction from SEP (23.85%) to JNT (22.32%) is statistically significant at the 1% level. Similar results are obtained for the IV/OOV estimation as shown in Fig. 3 (b). The EER reduction from SEP (24.72%) to JNT (22.96%) is again statistically significant at the 1% level. As described in Section 3.2, the correlation between confidence and OOV word utterances is very strong since OOV word utterances can never be correctly decoded. JNT directly models this correlation and *mutually* enhances its confidence and IV/OOV estimation performance. As a result, with JNT, we can obtain the improved confidence and IV/OOV estimation accuracies compared with those obtained with SEP as shown in Fig. 3 (a) and (b).

From Fig. 3 (d), we can confirm that the clean/noisy environment estimation is very easy under this experimental condition, and the accuracies obtained with the two methods are very high. Also in this case, JNT outperforms SEP in the EERs. The EER reduction from SEP (2.94%) to JNT (2.51%) is statistically significant at the 5% level. Utterances in noisy environments also degrade speech recognition performance. However, in contrast to OOV word utterances, they could be correctly decoded. In other words, the correlation between confidence and utterances in noisy environments is weak compared with that between confidence and OOV word utterances. As a result, the accuracy improvement in the clean/noisy environment estimation is smaller than that in the IV/OOV estimation.

In contrast to the other three estimation results, as shown in Fig. 3 (c), the gender estimation accuracies obtained with JNT and SEP are almost the same. The difference between their EERs is not statistically significant even at the 20%

level. As shown in Table 3, the recognition rates for the female (mismatched gender) utterances are almost the same with those for the utterances in noisy environments (the recognition rates for the evaluation data subsets (IV, male, noisy), (IV, female, clean) and (IV, female, noisy) are 63.29%, 63.20% and 28.02%, respectively). Therefore, we can expect the accuracy improvement in the gender estimation comparable to that in the clean/noisy environment estimation. However, we cannot obtain it. One reason of this result may be attributable to the gender identification performance by the GMMs shown in Fig. 2.

4. Conclusion and future work

We proposed a novel confidence measure based on the marginalization of jointly estimated error cause probabilities, and showed its good confidence estimation performance in the speech recognition experiments. To obtain further improvements in the confidence and each error cause estimation accuracies, we are planning to use more efficient features, e.g. those improve both the confidence estimation and OOV detection accuracies proposed in [10]. We are also planning to evaluate our methods in continuous speech recognition experiments.

5. References

- [1] Special Session, "Voice search technology and applications," Proc. ICASSP, SS-6, 2008.
- [2] Special Session, "Lessons and challenges deploying voice search," Proc. Interspeech, Wed-Ses-S1, 2009.
- [3] H. Jiang, "Confidence measures for speech recognition: A survey," Speech Communication, vol. 45, pp. 455–470, 2005.
- [4] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," Proc. ICASSP, pp. 809–812, 2007.
- [5] A. Ogawa and A. Nakamura, "Simultaneous estimation of confidence and error cause in speech recognition using discriminative model," Proc. Interspeech, pp. 1199–1202, 2009.
- [6] A. Ogawa and A. Nakamura, "Discriminative confidence and error cause estimation for extended speech recognition function," Proc. ICASSP, pp. 4454–4457, 2010.
- [7] A.L. Berger, S.A. Della Pietra, and V.J. Della Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics, vol. 22, no. 1, pp. 39–71, 1996.
- [8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," Proc. ICML, pp. 282–289, 2001.
- [9] 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. ASLP, vol. 15, no. 4, pp. 1352–1365, May 2007.
- [10] C. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence estimation, OOV detection and language ID using phone-to-word transduction and phone-level alignments," Proc. ICASSP, pp. 4085–4088, 2008.