DISCRIMINATIVE RECOGNITION RATE ESTIMATION FOR N-BEST LIST AND ITS APPLICATION TO N-BEST RESCORING

Atsunori Ogawa, Takaaki Hori and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation
{ogawa.atsunori,hori.t,nakamura.atsushi}@lab.ntt.co.jp

ABSTRACT

Techniques for estimating recognition rates without using reference transcriptions are essential if we are to judge whether or not speech recognition technology is applicable to a new task. We have proposed a *discriminative recognition rate estimation (DRRE)* method for 1-best recognition hypotheses and shown its good estimation performance experimentally. In this paper, we extend our DRRE to N*-best lists* of recognition hypotheses by modifying its feature extraction procedures and efficiently selecting N-best hypotheses for its discriminative model training. In addition, we apply our extended DRRE to N*-best rescoring*. In the experiments, the extended DRRE also showed good estimation performance for the N-best lists. And using the estimated recognition rates, the 1-best word accuracy was significantly improved by N-best rescoring from the baseline.

Index Terms— Speech recognition, discriminative recognition rate estimation, N-best list, N-best rescoring

1. INTRODUCTION

If we are to judge whether or not speech recognition technology is applicable to a new task, a crucial factor is recognition accuracy for the task. To calculate the recognition rates for a task, we have to prepare reference transcriptions for a large amount of speech data of the task. However, the cost for manual transcription is very high. Therefore, developing methods that can *estimate speech recognition rates without using reference transcriptions* is very important for applying speech recognition technology to various tasks with minimal cost.

Some methods have been proposed for estimating speech recognition rates without using reference transcriptions, e.g. [1, 2, 3, 4]. With these methods, the recognition rate, i.e. the percent correct (%Cor) or word accuracy (WAcc), for a task is estimated as a function of the factors that affect the recognition rate, e.g. WAcc=$f(\alpha_1, \alpha_2, \cdots)$ [%], where $\alpha_1, \alpha_2, \cdots$ are the factors such as the quality of the input speech and the task complexity (e.g. perplexity [4]). Good estimation performance is reported in the literature. However, these methods require us to have some prior knowledge about the target task. For example, we have to select a few important factors from many possible factors that affect the recognition rate of the task. And we have to know in advance the actual values of some factors, e.g. the task complexity [4], that are essentially unknown for a new task.

In contrast to the conventional methods, we have proposed a recognition rate estimation method based on *error type classification (ETC)* [5, 6]. In ETC, recognized words are classified into four categories, namely, correct (C), substitution error (S), insertion error (I) and deletion error (D) along with their probabilities. By individually summing these CSID probabilities over the recognition results for a task, we can obtain the estimated numbers of CSIDs (i.e. #C, #S, #I and #D). Then using these numbers, we can estimate the recognition rates of the task as %Cor=(#C/#N)×100 [%] and WAcc=(#C−#I)/#N×100 [%], where #N is the estimated number of words in the recognition results obtained as #N=#C+#S+#D. ETC is a simple extension of *confidence estimation* and does not require any prior knowledge for estimating the recognition rates of a new task. In addition, we can use *discriminative models* for ETC with many types of features, as with the recent trends in confidence estimation (and out-of-vocabulary word detection), e.g. [7, 8, 9, 10]. We have shown experimentally that the recognition rates can be accurately estimated with our *discriminative recognition rate estimation (DRRE)* method based on ETC [5, 6].

The conventional methods and our DRRE have been developed for estimating the recognition rates of *1-best hypotheses*. However, 1-best hypotheses sometimes contain many errors. Therefore, many techniques and systems related to speech recognition technology have been developed based on the forms of *multiple recognition hypotheses*. For example, spoken term detection and spoken document retrieval techniques exploit word (and/or sub-word) based lattices and/or confusion networks, e.g. [11, 12]. Spoken dialogue systems exploit N-best lists for dialogue modeling, e.g. [13, 14]. And natural language processing applications such as machine translation and parsing also exploit N-best lists, e.g. [15, 16]. These facts indicate that, if the recognition rate estimation methods are also applicable to multiple hypotheses rather than just to 1-best hypotheses, they have great potential to enhance the performance of the above techniques and systems, e.g. by selecting and/or reranking the hypotheses using the estimated recognition rates.

In this paper, we extend our DRRE to N*-best lists*. Because of a wide range of feature values of N-best hypotheses, it is difficult to directly apply our previous 1-best hypotheses based method [5, 6] to N-best lists. Therefore, we *modify the feature extraction procedures* for N-best hypotheses (Section 2.1). And the size of the training data is vastly increased, we develop a *hypothesis selection method* for efficient discriminative model training using N-best lists (Section 2.2). Our hypothesis selection method is inspired by that proposed in [17] for discriminative language modeling. In addition, we apply our extended DRRE to N*-best rescoring* [18] (Section 2.3). In the experiments, the extended DRRE also showed good estimation performance for the N-best lists (Section 3.2). And using the estimated recognition rates, the 1-best WAcc was significantly improved by N-best rescoring from the baseline (Section 3.3).

2. DISCRIMINATIVE RECOGNITION RATE ESTIMATION FOR N-BEST LIST

We describe methods for extending our DRRE to N-best lists and applying it to N-best rescoring. To save space, here we partially describe the experimental setup (the rest is described in Section 3.1) along with explanations of the methods.

2.1. Extraction of N-best Word Alignment Features

Our ETC is based on conditional random fields (CRF) [19], i.e. a discriminative model, and therefore, an important point to conduct

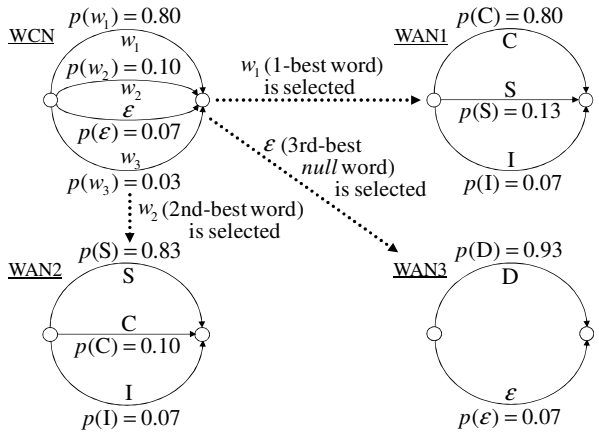

Fig. 1. Extraction of WAFs by selecting a word hypothesis as the recognition result from the competing word hypotheses on a WCN (the WCN and WANs have only one segment for simplicity).

accurate ETC is introducing effective features as with recent confidence estimation (and out-of-vocabulary word detection) methods based on discriminative models, e.g. [7, 8, 9, 10]. For those features, we have proposed *word alignment features (WAFs)* [5, 6] and confirmed their effectiveness for ETC experimentally [5].

Figure 1 shows the WAF extraction procedures for N-best word hypotheses. These procedures are common in both the training and evaluation phases. We assume a speech recognizer that can provide a word confusion network (WCN) with word posterior probabilities as the multiple recognition hypotheses [20]. In accordance with our previous proposal [5, 6], in Fig. 1, we only select the 1-best word hypothesis on the WCN, i.e. w_1, as the recognition result. By this selection, the WCN is converted to a *word alignment network (WAN)*, i.e. WAN1, with the *correct (C), substitution error (S) and insertion error (I) probabilities*, i.e. $p(\mathrm{C})$, $p(\mathrm{S})$ and $p(\mathrm{I})$ ($p(\mathrm{C})$ is a confidence measure [12]). And these probabilities (and the *deletion error (D) probability*, i.e. $p(\mathrm{D})$) are the WAFs extracted for the selected word hypothesis (the conversion procedures from a WCN to a WAN, i.e. the WAF extraction procedures, are detailed in [5, 6]).

To extend our DRRE to an N-best list, we also select word hypotheses lower than the 1-best hypothesis on the WCN as the recognition results. For example, if we select w_2, i.e. the second best word, the WCN is converted to WAN2 with the WAFs. And if we select ε, i.e. the third best *null* word, the WCN is converted to WAN3 with $p(\mathrm{D})$.

If the 1-best word hypothesis is selected, its $p(C)$ tends to be high (close to 1) and, accordingly, $p(S)$ and $p(I)$ tend to be low (close to 0) as shown in WAN1 [5, 6]. In contrast, if the lower-rank word hypotheses are selected, their $p(C)$, $p(S)$ and $p(I)$ have a wide range of values between 0 to 1 as shown in WAN2 ($p(I)$ is low in WAN2, however, there is a case where $p(I)$ becomes high). These differences obviously indicate that, in the CRF training described below, we have to use the WAFs of the lower-rank hypotheses in addition to the WAFs of the 1-best hypotheses to train the CRFs that can capture a wide range of WAF values in the N-best hypotheses.

WAF extraction can be viewed as an ETC. And in the evaluation phase, with the WAFs, i.e. the CSID probabilities on the WANs, we can estimate the recognition rates for the evaluation data *without using* the reference transcriptions based on the procedures described in Section 1. However, we have experimentally confirmed that the estimation performance with the WAFs alone is not very high [5, 6]. Therefore, we *refine* the CSID probabilities with CRFs using many types of features in addition to the WAFs.

Table 1. 18 features used in the CRFs. IDs from 1 to 4 are the WAFs.

ID	Feature	ID	Feature
1	*Correct recog. prob.*	10	Acoustic log like.
2	*Substitution error prob.*	11	Unigram log like.
3	*Insertion error prob.*	12	Trigram log like.
4	*Deletion error prob.*	13	Back-off behavior
5	Recog. word itself	14	# of alternative hyps.
6	Part-of-speech	15	Rank in competing hyps.
7	Number of frames	16	# of preceding ε segs.
8	Number of phones	17	Sum. of ε probs.
9	# of frames per phone	18	Sum. of # of alt. hyps.

Table 2. The CRFs and their conditions of the hypothesis selection.

CRF name	Diff. rate thrsh. [%]	# of train. smpls. [k]	CRF size [MB]	
			CSI-CRF	D-CRF
1B	—	114	1.7	1.5
NBx2	60	239	2.0	1.9
NBx4	50	468	2.9	2.6
NBx8	40	878	4.5	4.2
NBx18	30	2073	8.9	8.3
NBx80	20	9073	23.2	21.4

Table 1 lists 18 features used in the experiments described in Section 3. In the training phase, we trained the CRFs using these features for each word and the corresponding reference CSID labels obtained with a NIST SCLITE scoring tool [21] using the reference transcriptions. We quantized the features [7] and also used the contextual features [10]. We trained two types of CRFs; one was *CSI-CRF*, which refines the CSI probabilities and the other was *D-CRF*, which refines the D probabilities. D-CRF is especially needed since deletion errors can occur at arbitrary inter-word positions in a recognition hypothesis with arbitrary numbers [6]. Then in the evaluation phase, we *refined* the CSID probabilities using the CSI-CRF and D-CRF and, with these refined probabilities, we estimated the recognition rates for the evaluation data.

2.2. Hypothesis Selection for CRF Training

As described in Section 2.1, to accurately estimate the recognition rates of each recognition hypothesis (recognized sentence) in an N-list, we should train the CRFs by using N-best lists as the training data. We are assuming a WCN to be the multiple recognition hypotheses provided by a speech recognizer. And an N-best list of the hypotheses can be converted from a WCN. Each word in the N-best hypotheses can be linked to its original position in the source WCN. And thus we can extract the WAFs (and other features) for each word with the procedures described in Section 2.1.

An N-best list is the simplest form of the multiple recognition hypotheses. However, compared with a WCN, i.e. the most compact form, an N-best list is a very *inefficient* form. In an N-best list, the difference between the r-th and $(r+1)$-th ranks of hypotheses is usually only one word and many similar hypotheses are listed in several consecutive ranks. Therefore, the training of the CRFs using N-best lists will also be very inefficient, and the size of the resultant CRFs will be redundantly large.

To solve the problem of training using N-best lists, a *hypothesis selection* method is proposed in [17] for discriminative language modeling. The goal of the discriminative language model in [17] is to differentiate the recognition hypothesis that has the fewest errors in an N-best list from its competitors with more errors. And in the literature, it is concluded that this goal can be achieved by using only the most errorful hypothesis in the N-best list as the competitor.

In contrast to the case described in [17], we think that the key point of the hypothesis selection for our CRF training is to collect the *varieties* of recognition hypotheses from an N-best list. And we have developed the following method:

(i) We first put the 1-best hypothesis into the selected hypothesis set and continue the following procedures until we reach the lowest rank of the N-best list.

(ii) We compare the current rank of the hypothesis with each of the already selected hypotheses and find the hypothesis that is most similar to the current one. Then we calculate the *difference rate* of these two hypotheses, i.e. the ratio of the number of different words divided by the sentence length (their sentence lengths are the same since they are extracted from the same WCN). And if the difference rate is higher than the previously determined threshold, we add the current hypothesis to the selected hypothesis set.

Table 2 lists the CRFs used in the experiments described in Section 3. "1B" is the CRF (CSI-CRF and D-CRF set) trained using only the 1-best hypotheses from the N-best lists, i.e. the CRF trained based on our previous proposal [5, 6]. And "NBxm" are the CRFs trained using the hypothesis sets generated by our hypothesis selection method from the N-best lists. We set N at 5000 with reference to [17] and changed the difference rate threshold at 60, 50, 40, 30 and 20 [%]. When the threshold is set higher, the differences between the selected hypotheses increase, the number of selected hypotheses decrease, and the sizes of the trained CRFs also decrease. Conversely, when the threshold is set lower, the opposite results are obtained. m in NBxm denotes the ratio of the number of hypotheses divided by that of 1B. For example, on average, eight recognition hypotheses were selected from a 5000-best list and used to train NBx8.

2.3. Application to N-best Rescoring

We can *rerank* the recognition hypotheses in an N-best list, i.e. conduct *N-best rescoring* [18], using their recognition rates estimated with our DRRE. In the following, w_i^r is the i-th word in the r-th rank of hypothesis $\mathbf{w}^r$ in an N-best list, L is the length of (i.e. the number of words in) $\mathbf{w}^r$, $p(w_i^r)$ is the posterior probability of w_i^r provided by the source WCN of the N-best list, and $a(\mathbf{w}^r)$ is the WAcc of $\mathbf{w}^r$ estimated by the DRRE. We calculate the score of $\mathbf{w}^r$ taking $a(\mathbf{w}^r)$ into account as

$$s(\mathbf{w}^r) = (1-\lambda)\exp\left(\frac{\sum_{i=1}^{L}\log p(w_i^r)}{L}\right) + \lambda\frac{a(\mathbf{w}^r)}{\beta}. \quad (1)$$

Where the first term is the posterior probability of $\mathbf{w}^r$. In the second term, $a(\mathbf{w}^r)$ is divided by a coefficient β so as to balance its range with that of the first term. This time β is set at 100. And λ is the interpolation coefficient of the first and second terms ($0 \le \lambda \le 1$). As λ is set larger, $a(\mathbf{w}^r)$, i.e. the DRRE estimation result, becomes more emphasized.

We calculate this score for all ranks of the recognition hypotheses in an N-best list, and we rerank the hypotheses using these scores. If the DRRE estimation performance is high, the 1-best hypothesis in an N-best list will be replaced by another hypothesis that has a higher WAcc. And as a result, the WAcc of the 1-best hypotheses for the evaluation data will be improved. Note that we can conduct N-best rescoring based on accuracy or error measures other than WAcc, e.g. %Cor and the substitution, insertion and deletion error rates defined as (#S/#E)×100 [%], (#I/#E)×100 [%] and (#D/#E)×100 [%] (#E is the estimated number of errors obtained as #E=#S+#I+#D) [12].

Table 3. Recognition rate estimation results for the 1-best hypotheses of the entire evaluation data obtained by 1B and NBx8.

CRF	#N	#C	#S	#I	#D	%Cor	WAcc
True	94449	72191	17345	4130	4913	76.43	72.06
1B	93337	71110	18120	4436	4107	76.19	71.43
NBx8	93328	71177	17848	4641	4303	76.27	71.29

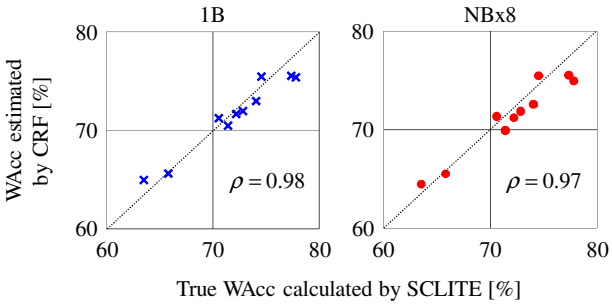

Fig. 2. Correlation of lecture-level WAccs for 1-best hypotheses calculated by SCLITE (true WAccs) and those estimated by 1B or NBx8.

3. EXPERIMENTS

We conducted experiments to evaluate the recognition rate estimation performance of our DRRE for N-best lists and its applicability to N-best rescoring. We performed all the experiments with our speech recognition platform SOLON [22] using the MIT lecture speech corpus [23, 24].

3.1. Experimental Setup

An HMM-based acoustic model was discriminatively trained by using 110 hours (104 lectures) of speech data with a differenced maximum mutual information (dMMI) criterion [25]. It had 2565 states optimized by the variational Bayesian estimation and clustering (VBEC) technique [26] and each state had 32-mixture Gaussian pdfs. A word trigram language model was trained by using 6.2M words of manually transcribed lecture speech. The vocabulary size of the lexicon was 16.5k.

The CRF training data consisted of 215 hours (238 lectures) of speech data (114k utterances and 2.0M words). We trained the CRFs listed in Table 2 using this training data with the procedures described in Sections 2.1 and 2.2. The evaluation data consisted of 9.3 hours (10 lectures) of speech data (9450 utterances, 94k words and a 2.87% out-of-vocabulary rate). We performed the feature extraction for this evaluation data with the procedures described in Section 2.1, estimated the recognition rates using the CRFs listed in Table 2, and conducted N-best rescoring with the method described in Section 2.3. With reference to [17], the N value of the N-best lists was set at 5000 in both the training and evaluation phases. In the evaluation phase, the total number of hypotheses in all the N-best lists was 14.7M. The *true* recognition rates were calculated by SCLITE [21] *using* the reference transcriptions. Note that this time we slightly increased both the training and evaluation data compared with those used in our previous experiments [5, 6].

3.2. Recognition Rate Estimation Results

We first show the recognition rate estimation results *for the 1-best hypotheses* obtained by 1B and NBx8. The estimation results obtained by NBxm CRFs other than NBx8 are omitted since they are similar to those obtained by NBx8. Table 3 shows the recognition rate estimation results obtained for the entire evaluation data. We

Table 4. Correlation coefficients (ρ) between utterance-level WAccs for N-best hypotheses calculated by SCLITE (true WAccs) and those estimated by each of the CRFs.

CRF	1B	NBx2	NBx4	NBx8	NBx18	NBx80
ρ	0.64	0.80	0.80	**0.81**	0.80	0.79

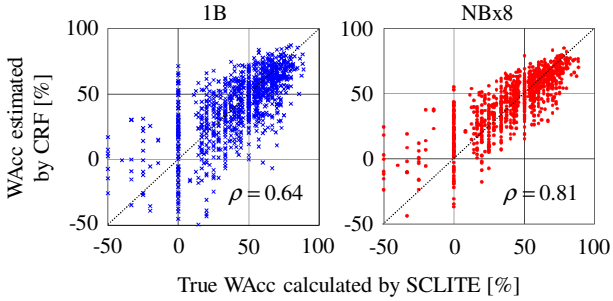

Fig. 3. Correlation of utterance-level WAccs for N-best hypotheses calculated by SCLITE (true WAccs) and those estimated by 1B or NBx8 (2k points are randomly sampled from 14.7M points).

can confirm that #C, #S, #I, #D, and #N estimated by 1B and NBx8 are reasonably close to their true values, and as a result, the %Cor and WAcc are estimated accurately, i.e. with an error rate of less than 1%. Figure 2 shows the correlation between the lecture-level true WAccs and their estimated values. We can confirm the high estimation performance of 1B and NBx8, i.e. their correlation coefficients ρ are 0.97 and 0.96, respectively. At the utterance-level, these values degrade to 0.70 and 0.69, respectively. However, these degraded values still appear to be reasonable since it is inherently difficult to estimate the WAcc for each utterance, especially for short utterances.

We next show the recognition rate estimation results obtained *for the N-best hypotheses*. Table 4 shows the correlation coefficients between true utterance-level WAccs and their values estimated by each of the CRFs. Figure 3 shows the results obtained by 1B and NBx8. We can confirm that NB$x m$ CRFs perform better than 1B. This is because, as described in Section 2.1, 1B only captures a *narrow* range of WAF values in the 1-best hypotheses. In contrast, NB$x m$ CRFs capture a *wide* range of WAF values in the N-best hypotheses. As regards m, based on Table 4, eight (i.e. 40% of the difference rate threshold as shown in Table 2) is slightly better than the other m values. This result indicates that the hypothesis selection described in Section 2.2 is also effective for our CRF training as reported in [17] for discriminative language modeling.

From the above results, we can confirm that 1B performs well only for the 1-best hypotheses, in contrast, NBx8 perform well *both* for the 1-best and N-best hypotheses. The correlation coefficient ρ by NBx8 for the N-best hypotheses is 0.81 as shown in Table 4 and Fig. 3 and this is higher than 0.69, i.e. the value for the 1-best hypotheses. This means that, with NBx8, the estimation for the 1-best hypotheses is more difficult than that for the hypotheses lower than the second best. This is an interesting result and we guess that this is because the 1-best recognition hypotheses tend to be *overestimated* by a speech recognizer [6].

3.3. N-best Rescoring Results

Figure 4 shows the N-best rescoring results obtained by 1B and NBx8. The correlation coefficient ρ by NBx8 for the N-best hypotheses is 0.81 as shown in Table 4 and Fig. 3. With this NBx8 estimation performance, the WAcc of the 1-best hypotheses is steadily

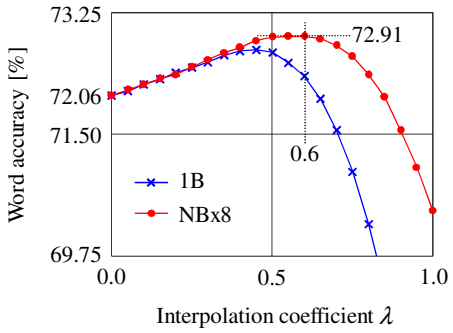

Fig. 4. N-best rescoring results obtained by 1B and NBx8, i.e. WAcc for the 1-best hypotheses of the entire evaluation data as a function of the interpolation coefficient λ.

improved by gradually changing the interpolation coefficient λ and it reaches its highest value 72.91% at $\lambda = 0.6$. This WAcc improvement from 72.06% (Table 3), i.e. the baseline, is statistically significant at the 1% level [27]. It is surprising that the WAcc is also improved even by 1B ($\rho = 0.64$).

4. RELATION TO PRIOR WORK

In Section 1, we have already discussed the relation (difference) between our DRRE and the function estimation type of conventional recognition rate estimation methods, e.g. [1, 2, 3, 4]. There is another type of method [28]. It is also an interesting approach that estimates the classification error rate based on using only the HMM parameters without running recognition experiments. However, the method considers only substitution errors and only isolated word/phone experiments were conducted.

Here we emphasize again that, in contrast to the conventional methods described above, our DRRE can estimate #C, #S, #I, #D, and #N as shown in Table 3. We believe that this feature of the DRRE can be applied to various techniques related to speech recognition technology. As one of the applications, the precision (recall) of a spoken document retrieval system can be improved by selecting recognition hypotheses with few insertion (deletion) errors [12].

N-best rescoring [18] is a traditional multiple-pass search strategy that is still frequently used in many techniques related to speech recognition technology. One such technique is discriminative language modeling, e.g. [17]. With discriminative language modeling, sophisticated objective functions are defined and optimized to generate the reranking models. In contrast, our N-best rescoring approach described in Section 2.3 is more *direct* since we directly estimate the recognition rates of the hypotheses in an N-best list so as to rerank them. We believe that our DRRE can be used in *unsupervised* discriminative language modeling, e.g. [17] (and also acoustic modeling, e.g. [29, 30]).

5. CONCLUSION AND FUTURE WORK

We have extended our discriminative recognition rate estimation (DRRE) method to N-best lists and applied it to N-best rescoring. In the experiments, our DRRE showed good estimation performance for the N-best lists and, using the estimated recognition rates, the 1-best word accuracy was significantly improved by N-best rescoring from the baseline. Future work will include improving the performance of our DRRE by using more efficient features, e.g. [7, 8, 9, 10, 31], and its application to, e.g., spoken document retrieval [12] and unsupervised discriminative language/acoustic modeling [17, 29, 30].

6. REFERENCES

[1] M. Kondo, K. Takeda, and F. Itakura, "Predicting the degradation of speech recognition performance from sub-band dynamic ranges," *Journal of Information Processing Society of Japan (IPSJ)*, vol. 43, no. 7, pp. 2242–2248, 2002.

[2] H. Sun, L. Shue, and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," in *Proc. ICASSP*. IEEE, 2004, vol. 1, pp. 865–868.

[3] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2006–2013, 2006.

[4] T. Yamada, T. Nakajima, N. Kitawaki, and S. Makino, "Performance estimation of noisy speech recognition considering task complexity," in *Proc. Interspeech*. ISCA, 2010, pp. 2042–2045.

[5] A. Ogawa, T. Hori, and A. Nakamura, "Error type classification and word accuracy estimation using alignment features from word confusion network," in *Proc. ICASSP*. IEEE, 2012, pp. 4925–4928.

[6] A. Ogawa, T. Hori, and A. Nakamura, "Recognition rate estimation based on word alignment network and discriminative error type classification," in *Proc. Workshop on Spoken Language Technology (SLT)*. IEEE, 2012.

[7] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. ICASSP*. IEEE, 2007, pp. 809–812.

[8] L. Burget, P. Schwarz, P. Matějka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Černocký, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in *Proc. ICASSP*. IEEE, 2008, pp. 4081–4084.

[9] 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. ICASSP*. IEEE, 2008, pp. 4085–4088.

[10] 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. Interspeech*. ISCA, 2010, pp. 1942–1945.

[11] C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[12] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proc. Special Interest Group on Information Retrieval (SIGIR)*. ACM, 2006, pp. 51–58.

[13] J.D. Williams, "Exploiting the ASR N-Best by tracking multiple dialog state hypotheses," in *Proc. Interspeech*. ISCA, 2008, pp. 191–194.

[14] J.D. Williams and S. Balakrishnan, "Estimating probability of correctness for ASR N-Best lists," in *the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL2009)*. ACL, 2009, pp. 132–135.

[15] L. Shen, A. Sarkar, and F. Och, "Discriminative reranking for machine translation," in *Proc. HLT-NAACL*. ACL, 2004, pp. 177–184.

[16] M. Collins and T. Koo, "Discriminative reranking for natural language processing," *Computational Linguistics*, vol. 31, no. 1, pp. 25–70, 2005.

[17] T. Oba, T. Hori, and A. Nakamura, "Efficient training of discriminative language models by sample selection," *Speech Communication*, vol. 54, no. 6, pp. 791–800, 2012.

[18] R. Schwartz, L. Nguyen, and J. Makhoul, "Multiple-pass search strategies," in *Automatic Speech and Speaker Recognition*, C.H. Lee, F.K. Soong, and K.K. Paliwal, Eds., pp. 57–81. Klewer Academic Publishers, Norwell, MA, 1996.

[19] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning (ICML)*. ACM, 2001, pp. 282–289.

[20] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, October 2000.

[21] "NIST SCLITE Scoring Package Version 1.5," http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm.

[22] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[23] 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. Interspeech*. ISCA, 2007, pp. 2553–2556.

[24] H.-A. Chang and J.R. Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," in *Proc. ICASSP*. IEEE, 2009, pp. 4481–4484.

[25] 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. ICASSP*. IEEE, 2010, pp. 4894–4897.

[26] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 855–872, 2006.

[27] S. Nakagawa and H. Takagi, "Statistical methods for comparing pattern recognition algorithms and comments on evaluating speech recognition performance," *Journal of the Acoustical Society of Japan*, vol. 50, no. 10, pp. 849–854, 1994.

[28] C.-S. Huang, H.-C. Wang, and C.-H. Lee, "A study on model-based error rate estimation for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 581–589, 2003.

[29] M. Delcroix, A. Ogawa, T. Nakatani, and A. Nakamura, "Dynamic variance adaptation using differenced maximum mutual information," in *Proc. Symposium on Machine Learning in Speech and Language Processing (MLSLP), available online*, 2012.

[30] M. Gibson and T. Hain, "Correctness-adjusted unsupervised discriminative acoustic model adaptation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2648–2656, 2012.

[31] P. Matějka M. Karafiát S. Kombrink, L. Burget and H. Hermansky, "Posterior-based out of vocabulary word detection in telephone speech," in *Proc. Interspeech*. ISCA, 2009, pp. 80–83.