N-BEST RESCORING BY ADABOOST PHONEME CLASSIFIERS FOR ISOLATED WORD RECOGNITION

Hiroshi Fujimura #1, Masanobu Nakamura #, Yusuke Shinohara #, Takashi Masuko #

Corporate Research and Development Center, Toshiba Corporation
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan
1 hiroshi4.fujimura@toshiba.co.jp

***Abstract*—This paper proposes a novel technique to exploit generative and discriminative models for speech recognition. Speech recognition using discriminative models has attracted much attention in the past decade. In particular, a rescoring framework using discriminative word classifiers with generative-model-based features was shown to be effective in small-vocabulary tasks. However, a straightforward application of the framework to large-vocabulary tasks is difficult because the number of classifiers increases in proportion to the number of word pairs. We extend this framework to exploit generative and discriminative models in large-vocabulary tasks. N-best hypotheses obtained in the first pass are rescored using AdaBoost phoneme classifiers, where generative-model-based features, i.e. difference-of-likelihood features in particular, are used for the classifiers. Special care is taken to use context-dependent hidden Markov models (CDHMMs) as generative models, since most of the state-of-the-art speech recognizers use CDHMMs. Experimental results show that the proposed method reduces word errors by 32.68% relatively in a one-million-vocabulary isolated word recognition task.**

I. INTRODUCTION

Many applications that exploit automatic speech recognition (ASR) demand better performance for large vocabulary ASR with the rise of machine power. Some of those applications require ASR engines to recognize several million words. However, it becomes difficult to discriminate target words from others because the large vocabulary includes similar word sets. Language models can help to discriminate similar words in the speech recognition process. Nevertheless, there are still some similar word sets that have a similar meaning and are used in similar contexts. In this case, the system has to discriminate the target phonemes from others because only a few different phonemes are the clues to discriminate similar words. Therefore we focus on phoneme discrimination in the speech recognition. As the first step, we consider the isolated word recognition to improve the "acoustic" phoneme discrimination performance.

The hidden Markov model (HMM) is the principal technique for automatic speech recognition. However, since its performance is not sufficient for speech recognition, many efforts have been made to compensate for the weakness of HMM. One of the well-known problems of HMM is that it is a generative model whereas the word recognition task is a discrimination problem. To cope with this gap, some techniques employed discriminative models. Ganapathiraju *et al.*[1] combined HMM and the support vector machine (SVM) which is one of the effective discriminative models. Their technique applies SVMs to phoneme segments obtained from alignment of HMM and N-best hypotheses are rescored by scores of SVMs. Gales and Longworth[2] also proposed combining the scores of HMMs and SVMs with generative kernels to improve the performance of the continuous digit task. Padrell-Sendra *et al.*[4] applied SVM to the frame-level discrimination combined with token passing instead of HMM, and obtained better performance than a standard HMM. Gunawardana *et al.*[5] applied hidden conditional random fields (HCRFs) to the speech recognition instead of HMMs.

Among previous studies of using discriminative models for speech recognition, we especially focus on the combination of generative and discriminative models in [2] [3]. This technique can also be applied to image recognition [6] [7]. This framework can easily apply the developed techniques of speech recognition to generative models because features of discriminative models are calculated from generative models. Gales and Longworth [2] apply SVMs with generative kernels to pairs of words. Classification scores are calculated from the SVM outputs. Finally the classification score is combined with the HMM score, which has led to improved performance on a continuous digit task. However it is difficult to model all pairs of words because the number of pairs is very large when it is applied to large vocabulary speech recognition. Therefore we cannot directly use this method for large vocabulary tasks. We would like to apply the framework of combining generative and discriminative models to large vocabulary tasks. The previous method could be applied to pairs of phonemes. However it is not realistic to prepare models for all pairs of all triphones. In this paper, we propose a combination method of generative and discriminative models for large vocabulary tasks. The combination is realized by N-Best rescoring by discriminative scores. In our technique, classification scores are calculated

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

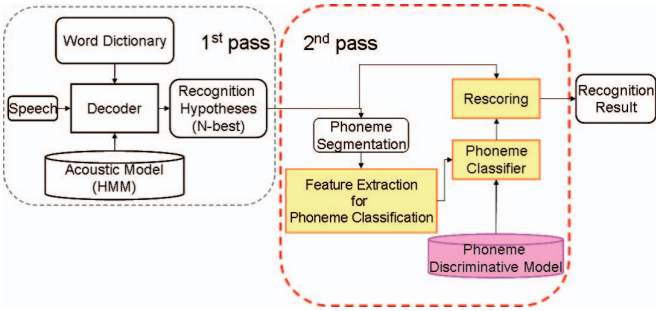

Fig. 1. *The framework of the proposed method.*

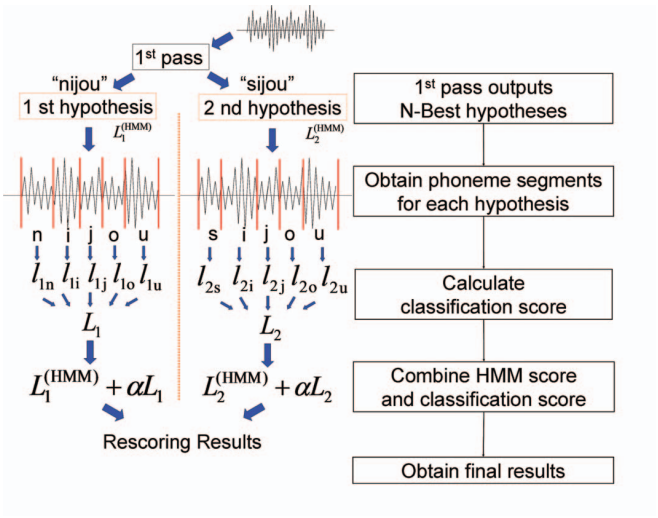

Fig. 2. *The flowchart of the second pass.*

using AdaBoost classifiers [8] [9] for the phoneme segments of N-best hypotheses given by HMMs, and N-best hypotheses are re-ordered based on the classification scores. AdaBoost classifies a phoneme of a segment into the correct phoneme or not as a binary classifier. For AdaBoost feature, we use difference-of-likelihood feature which is a part of generative process of HMM. It can handle the triphone framework for the discriminative model feature. All of these techniques are for the scalability for large vocabulary tasks.

The reminder of this paper is organized as follows: Section II introduces the proposed rescoring method by AdaBoost. Section III shows experimental setups and results of the proposed rescoring method. Conclusions are presented in section IV.

II. Rescoring Using AdaBoost Classifiers

A. Overview of the Rescoring Process

This section describes the proposed method to rescore the N-best hypotheses using AdaBoost phoneme classifiers. Our method consists of the first pass and the second pass shown in Fig.1. The first pass outputs N-best hypotheses with scores by a conventional speech recognition system using HMMs. The process of the second pass is shown in Fig.2. First, phoneme segments are obtained by forced alignment of phoneme HMMs for each hypothesis. Then, the difference-of-likelihood features are extracted from each phoneme segment, and classification scores are calculated by AdaBoost classifiers using the segment features. Finally, classification scores are added to the N-best scores and N-best hypotheses are re-ordered.

B. Segment Feature

In our approach, difference-of-likelihood of HMM as feature of generative process is used for segment feature of AdaBoost. Difference-of-likelihood of HMM $\phi_{1,2}(\mathbf{X};\boldsymbol{\lambda})$ between phonme class $p1$ and $p2$ is defined by the following equation for input feature vectors $\boldsymbol{x}$

$$\phi_{1,2}(\boldsymbol{x};\boldsymbol{\lambda}) = \frac{1}{T}\left[\log\left(P(\boldsymbol{x};\boldsymbol{\lambda}^{(p1)})\right) - \log\left(P(\boldsymbol{x};\boldsymbol{\lambda}^{(p2)})\right]\right, \quad (1)$$

where T is the number of frames in $\boldsymbol{x}$, and $P(\boldsymbol{x};\boldsymbol{\lambda}^{(p1)})$ and $P(\boldsymbol{x};\boldsymbol{\lambda}^{(p2)})$ are the likelihoods from HMMs with parameters $\boldsymbol{\lambda}$ associated with phoneme class $p1$ and $p2$. This calculation is applied to all phoneme combinations and they are concatenated to a vector $\boldsymbol{f}(\boldsymbol{x})$.

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \phi_{1,2}(\boldsymbol{x};\boldsymbol{\lambda}) \\ \phi_{1,3}(\boldsymbol{x};\boldsymbol{\lambda}) \\ \cdots \\ \phi_{N-2,N}(\boldsymbol{x};\boldsymbol{\lambda}) \\ \phi_{N-1,N}(\boldsymbol{x};\boldsymbol{\lambda}) \end{bmatrix}. \quad (2)$$

Fig.3 shows the process of extracting features. First phoneme segmentations are obtained by forced alignment based on triphone HMMs for N-Best hypotheses. Secondly, likelihood of each segment is calculated by each triphone HMM, where the left and right contexts of the triphone are defined by the phoneme sequence of the hypothesis. Thirdly, differences of likelihood are calculated for all phoneme combinations. Finally ${}_NC_2$ dimensional feature vector is extracted from the segment. This feature can handle the variations in the left and right contexts. Therefore the latter discriminator can ignore the contexts.

C. AdaBoost Score

Assume that there are N training samples $(x_1, y_1), ..., (x_N, y_N)$, where x_i and y_i denote a feature vector and a class label of sample i. AdaBoost classifiers are trained by the following steps:

- Step 1. Initialize sample weight distribution $D_0(i)$ by the following equation:

$$D_0(i) = \begin{cases} \dfrac{1}{2\sum_{j:y_j=1} 1}, & y_i = 1, \\ \dfrac{1}{2\sum_{j:y_j=-1} 1}, & y_i = -1. \end{cases} \quad (3)$$

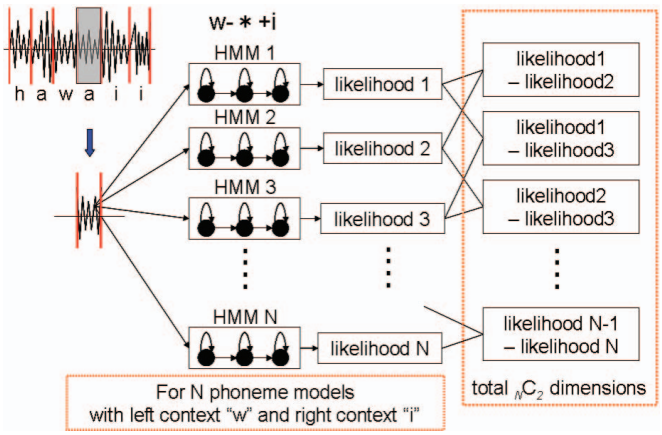

Fig. 3. *The process of extracting the difference-of-likelihood feature.*

- Step2. Train a weak classifier $h_t(x_i)$ minimizing error rate ε_t on sample weight distribution D_t,

$$\varepsilon_t = \sum_{i:y_i \neq h_t(x_i)} D_t(i). \tag{4}$$

- Step3. Compute voting weight α_t as

$$\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}. \tag{5}$$

- Step4. Recompute sample weight distribution as

$$\hat{D}_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i h_t(x_i)). \tag{6}$$

- Step5. Normalize the summation of sample weights to 1

$$D_{t+1}(i) = \frac{\hat{D}_{t+1}(i)}{Z_{t+1}}, \tag{7}$$

where

$$Z_{t+1} = \sum_{i=1}^{N} \hat{D}_{t+1}(i). \tag{8}$$

- Step6. Iterate from step 2 to step 5 T times, and obtain T weak classifiers.

Finally the strong classifier $H(x)$ is obtained by weighted sum of T weak classifiers as

$$H(x) = \text{sign}\left\{ \sum_{t=1}^{T} \alpha_t h_t(x) \right\}, \tag{9}$$

where each weak classifier outputs 1 (true) or -1 (false) by comparing a threshold and values of a selected dimension of the segment feature vectors.

A phoneme classifier is constructed for each phoneme by the AdaBoost algorithm. Each phoneme classifier discriminates whether the given segment is the phoneme or not, and outputs the AdaBoost score for the segment. As described in [10], the AdaBoost score $S_p(x)$ of a phoneme p for segment feature x is derived from the following equation,

$$S_p(x) = \frac{1}{\sum_{t=1}^{T} \alpha_t^{(p)}} \sum_{t=1}^{T} \alpha_t^{(p)} h_t^{(p)}(x). \tag{10}$$

Note that the range of $S_p(x)$ is $-1 \leq S_p(x) \leq 1$ due to the normalization.

TABLE I
Training and evaluation data

Language	Japanese
Training data	120 hours
#Evaluation utterances	5800 (58 speakers)
#Phonemes	39 (including a silence phoneme)
#Phoneme segments in evaluation utterances	43848 (including silences)
Evaluation task	1 million isolated words recognition
Evaluation vocabulary	city and town names station names, people's names, nouns from the Web
Feature for HMM	12-dimensional MFCCs + power, their Δs and $\Delta\Delta$s (totally 39 dimensions)
Feature for AdaBoost #dimensions	difference-of-likelihood $_{39}C_2 = 741$

D. Classification Score

Classification scores are calculated from the AdaBoost scores. First the classification score is calculated at each phoneme segment in all N-best hypotheses. Classification score l_i of i-th phoneme segment in a hypothesis is obtained by the following equation:

$$l_i = \log\left(S_{ip_i} + 1\right), \tag{11}$$

where p_i denote the phoneme of the i-th segment in the hypothesis. In order to guarantee that the normalized score is positive, a constant 1 is added to all AdaBoost scores S_{ip_i}.

The classification score L for a hypothesis is obtained by averaging l_i in the hypothesis,

$$L = \frac{1}{K} \sum_{i=1}^{K} l_i, \tag{12}$$

where K is the number of phonemes in the hypothesis. Finally, the rescored score L_{re} is obtained by weighted sum of L and HMM score $L^{(\text{HMM})}$ of the hypothesis,

$$L_{\text{re}} = (1-\alpha)L^{(\text{HMM})} + \alpha L \quad (0 \leq \alpha \leq 1), \tag{13}$$

and the N-best hypotheses are re-ordered using the score L_{re}.

III. Experiments

A. Experimental Setup

1) Evaluation Data: A one-million-word recognition experiment was conducted using a Japanese speech database. A one-million-word grammar was used where the one-million-word vocabulary consisted of Japanese city names, town names, station names, people's names and nouns from the Web. Evaluation data were recorded under clean conditions. All utterances in the evaluation data were Japanese city names in one-million-words such as 'Sapporo' and 'Meguro.' The number of speakers is 58 and each speaker utters 100 words. Hence, 5800 utterances were used for the evaluation.

2) HMM training: The HMMs were trained by 120-hour Japanese speech data recorded under clean conditions. The basic structure of the HMM is three-state continuous density triphones that share 3000 states with 20 Gaussian mixture components. All triphones have a simple left-to-right topology. A feature vector for the HMM consisted of 12 MFCCs (Mel-frequency cepstral coefficients), a power, and their Δs and $\Delta\Delta$s (39 dimensions). HTK[11] was used for the HMM training. HLDA (heteroscedastic linear discriminant analysis)[12] without nuisance dimensions and MPE (minimum phone error) training[13] were applied to the HMM after the MLE training.

3) AdaBoost Training: AdaBoost was trained using the same data as HMMs'. The feature parameters for AdaBoost were obtained by the method described in Sec.II-B based on forced alignment of HMMs which was trained in Sec.III-A2. AdaBoost classifiers were trained for 39 phonemes which include a silence phoneme. Consequently, there were 741 dimensions for AdaBoost features ($_{39}C_2 = 741$) to take differences of likelihood for 39 phonemes combinations. Each AdaBoost classifier discriminates if a given segment is the phoneme or not. The positive samples for phoneme A are samples which have label A in training data, and negative samples are others. Each selection of weak classifier step selects the best dimension and its threshold for the current discrimination.

B. Rescoring Experiment

1) Recognition Performance: Performance of the proposed technique was evaluated on the isolated word recognition task shown in Sec.III-A1. In this experiment, the first pass output 32-best hypotheses, and they were rescored by the AdaBoost scores in Eq.(12) in the second pass. 1-best and 5-best results are shown in this experiment. Fig. 4 and Fig. 5 show the 1-best and 5-best performance of the rescoring, respectively. The horizontal axis shows the number of weak classifiers. $\mathrm{Weak\ Classifiers} = 0$ means the original result without rescoring. "Correct" means the correct rate which is calculated from the following equation:

$$Correct\ [\%] = 100 \times \frac{\#correct}{\#sentence}, \tag{14}$$

They show that the performance increases as the number of weak classifiers increases. Both 1-best and 5-best were almost saturated where the number of weak classifiers is 400. The original performance of 1-best and 5-best were 66.67% and 90.76%, respectively. The 1-best and 5-best performance of rescoring by 400 weak classifiers were 70.84% and 93.78%, respectively. They show that the proposed technique improves the performance of the isolated word recognition.

Fig. 6 shows the Relative Error Rate Reduction (RERR) of the proposed technique. RERR is calculated from the following equation:

$$RERR\ [\%] = 100 \times \frac{Correct_{rescore} - Correct_{original}}{100 - Correct_{original}}, \tag{15}$$

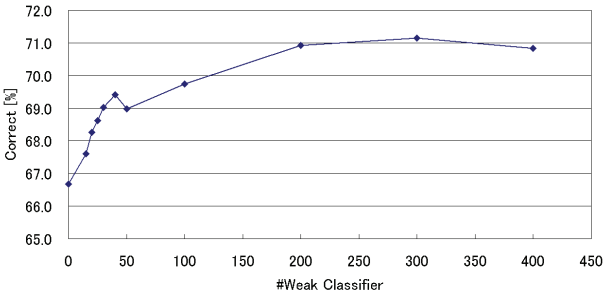

Fig. 4. *1-best result of rescoring by the proposed technique.*

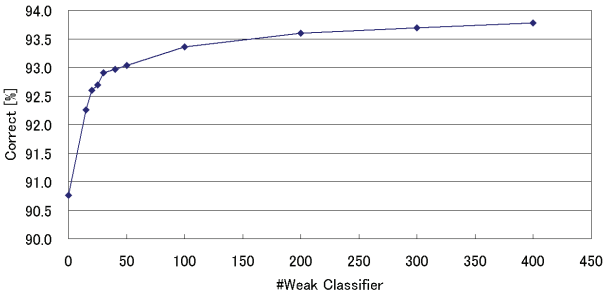

Fig. 5. *5-best result of rescoring by the proposed technique.*

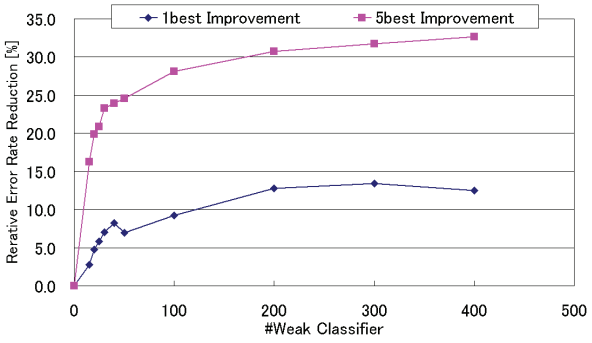

Fig. 6. *Relative Error Rate Reduction of the proposed technique.*

RERR of 5-best was higher than 1-best for any number of weak classifiers. We calculated $1-31$ best results for $\#Weak\ Classifiers = 400$ in order to obtain more details about N-Best results. Fig. 7 and Fig. 8 show the $1-32$ best results and their RERR. They show that the proposed technique is especially effective for $4-16$ best when 32 best results are rescored. These results obtained more than 30.0% RERR. We conclude that this framework using difference-of-likelihood feature is more effective for $4-16$ best recognition.

We show the relation between the parameter α and the peformance. Fig. 9 shows the best parameter α for each

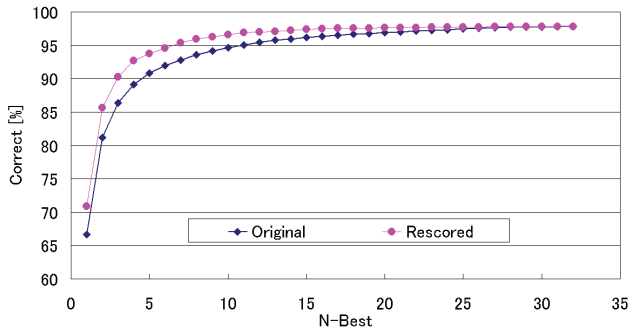

Fig. 7. *The 32-best performance of* $\#Weak\ Classifiers = 400$.

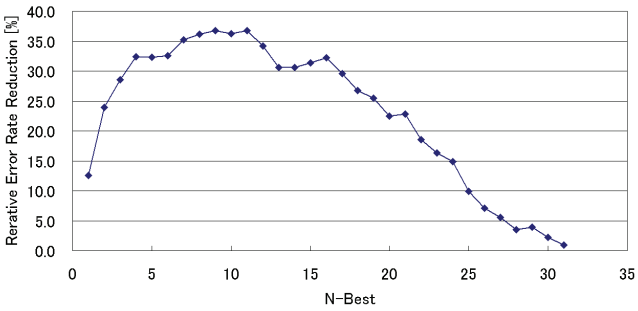

Fig. 8. *The 32-best performance of* $\#Weak\ Classifiers = 400$.

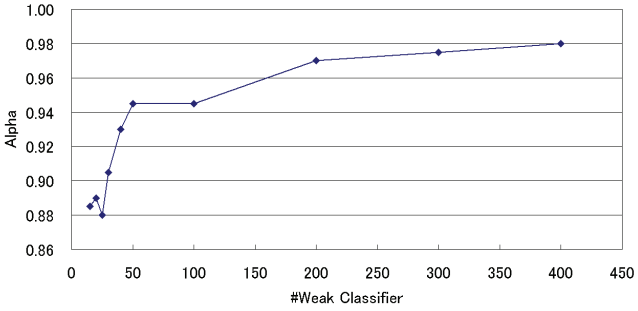

Fig. 9. *The best coefficient* α *for the number of weak classifiers.*

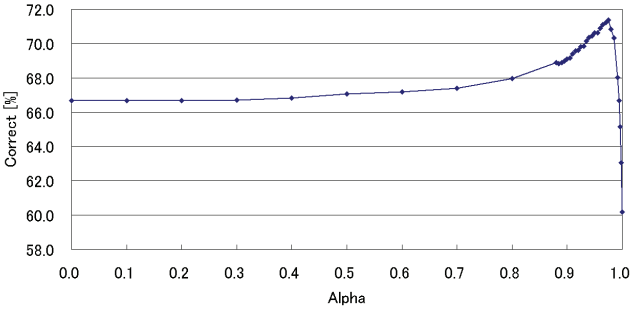

Fig. 10. *The 1-best performance by the control of coefficient* α *(*$\#Weak\ Classifiers = 400$*).*

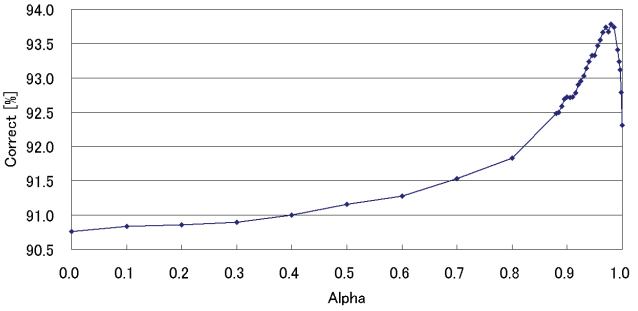

Fig. 11. *The 5-best performance by the control of coefficient* α *(*$\#Weak\ Classifiers = 400$*).*

number of weak classifiers. The α which means the weight of AdaBoost score increased as the number of weak classifiers increased. It is natural because AdaBoost becomes strong discriminator. Fig. 10 and Fig. 11 show 1-best and 5-best results, respectively when the number of weak classifiers is fixed at 400 and parameter α is changed. $\alpha = 0.0$ denotes the result without rescoring by AdaBoost score. On the other hand, $\alpha = 1.0$ denotes the result using AdaBoost score only. In our experiment the range of HMM score and AdaBoost score were not normalized. Therefore absolute value of α has no meaning. The performance of rescoring is better than original's for any value of α in the case of 5-best results. On the other hand, the performance degrades when α is near 1.0. However trends of both 1-best and 5-best are the same, and a certain range of α for contribution to improving performance exists.

IV. Conclusions

We proposed a method of combining generative and discriminative models for large vocabulary tasks. The combination is realized by N-best rescoring by classification scores. In our technique, classification scores are calculated using AdaBoost phoneme classifiers for the phoneme segments of N-best hypotheses given by HMMs, and N-best hypotheses are re-ordered based on the weighed sum of the HMM score and the classification score. We use difference-of-likelihood feature which is a part of generative process of HMM for AdaBoost. A phoneme classifier is constructed for each phoneme by the AdaBoost algorithm. Each phoneme classifier discriminates whether the given segment is the phoneme or other phonemes, and outputs the AdaBoost score for the segment as the classification score. Experimental results showed that the proposed technique consistently improved the isolated word recognition performance. RERRs for 1-best and 5-best results were 12.51% and 32.68%, respectively when the number of weak classifiers was 400. This framework has many possibilities because we can try other classification scores using other discriminators. Furthermore it can be applied to LVCSR tasks because it is based on the phoneme-based discrimination models and it can handle the framework of triphones.

pass rescoring in the future.

REFERENCES

[1] A. Ganapathiraju, J. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *Signal Processing, IEEE Trans*. Vol. 52, Issue 8, 2004.

[2] M.J.F. Gales and C. Longworth, "Discriminative classifiers with generative kernels for noise robust ASR," *INTERSPEECH*, 2008.

[3] M.J.F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Computer Speech and Language*, Vol.24, pp. 648-662, 2010.

[4] J. Padrell-Sendra, D. Martin-Iglesias, and F. Diaz-de-Maria, "Support vector machines for continuous speech recognition," In *Proceedings of the 14th European Signal Processing Conference*, Florence, Italy, 2006.

[5] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification," *INTERSPEECH*, 2005.

[6] F. Perronnin, C. Dance, "Fisher kernels on visual vocabularies for image categorization," In *CVPR*, 2007

[7] F. Perronnin, J. Sanchez and T. Mensink, "Improving the Fisher kernel for large-scale image classification," In *ECCV*, 2010.

[8] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and Science*, Vol. 55, pp. 119-139, 1997.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. CVPR*, pp. 511-518, 2001.

[10] K. Shutte and J. Glass, "Speech Recognition with Localized Time-Frequency Pattern Detectors," *ASRU*, pp. 341-346, 2007.

[11] S. Young *et al.*, "The HTK Book," Cambridge University Engineering Department, 2009.

[12] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, " *Speech Communication*, Vol. 26, pp. 283-297, 1998.

[13] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," *Proc. ICASSP-02*, pp.I-105-I-108, April 2002.