REJECTION OF OUT-OF-VOCABULARY WORDS USING PHONEME CONFIDENCE LIKELIHOOD

Takatoshi JITSUHIRO , Satoshi TAKAHASHI , Kiyoaki AIKAWA

NTT Human Interface Laboratories
1-1, Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239 Japan
{jitsu, taka, aik}@nttspch.hil.ntt.co.jp

ABSTRACT

The rejection of unknown words is important in improving the performance of speech recognition. The anti-keyword model method can reject unknown words with high accuracy in small vocabulary and specified task. Unfortunately, it is either inconvenient or impossible to apply if words in the vocabulary change frequently. We propose a new method for task independent rejection of unknown words, where a new phoneme confidence measure is used to verify partial utterances. It is used to verify each phoneme while locating candidates. Furthermore, the whole utterance is verified by a phonetic typewriter. This method can improve the accuracy of verification in each phoneme, and improve the speed of candidate search. Tests show that the proposed method improves the recognition rate by 4% compared to the conventional algorithm at equal error rates. Furthermore, a 3% improvement is obtained by training acoustic models with the MCE algorithm.

1. INTRODUCTION

It is important to reject out-of-vocabulary utterances in practical speech recognition. The wrong actions might triggered by an unknown word, a noise, a breath sound, or an unnecessary word. Therefore, many methods have been proposed to the rejection of out-of-vocabulary words.

There are two types of rejection methods. One depends on the task, the other does not. Typical task-dependent methods are reported in [1][2][3], etc. The models such as anti-keywords and anti-subwords are made to match the task, and these are used to normalize candidate likelihood. These methods suit tasks with small vocabularies and fail to handle vocabularies that are large or that change.

Task independent methods operate in the speech recognition system which has no vocabulary restriction [4][5]. Such methods are called vocabulary-free speech recognition or the phonetic typewriter. Vocabulary-free speech recognition uses the loop network of the syllables. This method compares the likelihood by the vocabulary-based recognition to the likelihood by the vocabulary-free recognition, and decides whether the obtained candidate should be rejected. This is independent of the task, and so it supports tasks that use changeable vocabularies.

The performance of vocabulary-free speech recognition is, however, not so high so the rejection accuracy is not so high. Furthermore, it does not work effectively for slightly different words because it verifies entire utterances.

Our goal is a task independent rejection method that offers high rejection performance. We use the verification of partial utterances in addition to likelihood normalization by a vocabulary-free recognition system. The phoneme confidence measure is defined for partial verification; it is based on the difference of likelihood between the current phoneme and the other phonemes. Moreover, the likelihood of the phoneme confidence measure is added to the accumulated acoustic likelihood of each Viterbi path to impose a penalty on unknown words. The proposed rejection method is, therefore, more accurate than the phonetic typewriter, and is independent of the task if the acoustic model is independent of the task. Furthermore, candidate search efficiency can be increased.

First, in Section 2, we define the phoneme confidence likelihood. Second, in Section 3, we briefly explain MCE. Finally, in Section 4, experiments and experimental results are described.

2. REJECTION BY PHONEME CONFIDENCE LIKELIHOOD

2.1. Recognition Processing Using Phoneme Confidence Likelihood

We aim at improving the accuracy of rejection by using the likelihood distributions of voice segments. The relative likelihoods are calculated for each phoneme, and the logarithm of the relative likelihoods are added to the accumulated log likelihoods of each Viterbi path. This means that this process imposes a penalty on the confidence likelihood of each phoneme. The method requests these relative likelihood distributions beforehand, and calculates likelihoods from the distributions at recognition step. Here, the relative likelihood of each phoneme is defined as the phoneme confidence likelihood.

As a result, it becomes possible to delete phonemes that have small phoneme confidence likelihoods in the search process. Also, the likelihoods of candidates for unknown words will be lower at the end of utterance even if they are not deleted until that time. Therefore, it will be easier to reject unknown words than if we use only vocabulary-free recognition.

After the logarithm of phoneme confidence likelihood $p_i(X_1^2)$ is calculated at the end of each phoneme, as shown

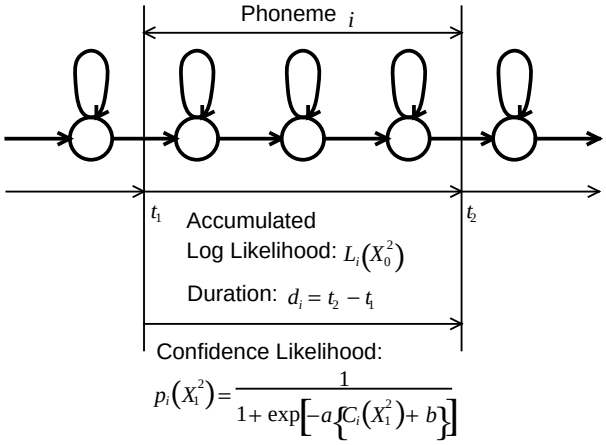

Figure 1: Calculation of confidence likelihood.

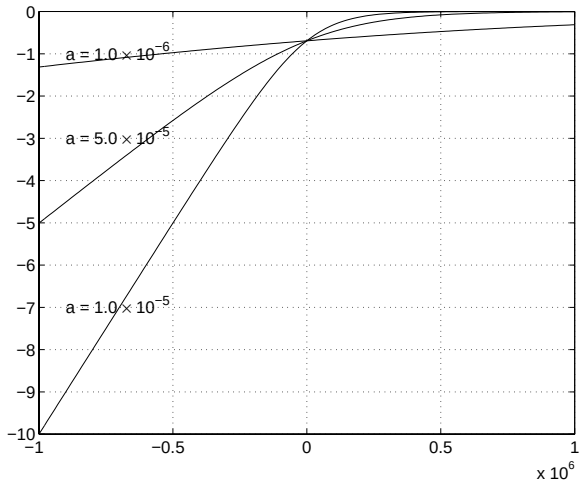

Figure 2: Logarithm of Sigmoid Function $y = -\log[1 + \exp\{-a(x+b)\}]$.

in Figure 1, it is multiplied by the constant α, and added to accumulated acoustic logarithm likelihood $L_i(X_0^2)$ at that time.

$$\hat{L}_i(X_0^2) = L_i(X_0^2) + \alpha \times \log\{p_i(X_1^2)\},$$

where X_1^2 is the feature vector from time t_1 to t_2, and α is constant. $\hat{L}_i(X_0^2)$ becomes the accumulated logarithm likelihood of the path, its weight matches the reliability of the phoneme.

In addition, recognition candidate's likelihood is normalized by the accumulated logarithm likelihood obtained from the vocabulary-free speech recognition system at the end of the utterance. The candidates are decided by their normalized likelihood values whether they should be rejected.

2.2. Definition of Phoneme Confidence Likelihood

The phoneme confidence measure is defined as follows:

$$C_i(X_1^2) = \frac{1}{d_i}\sum_{t=t_1}^{t_2}\left[g_i(X_t) - \frac{1}{N-1}\sum_{j,j\neq i} g_j(X_t)\right],$$

where $g_i(X_t)$ is the logarithm likelihood of the i phoneme model of the candidate for feature X_t of the input voice at time t, and N is the total number of phoneme models, and $d_i = t_2 - t_1$ is the duration. The right second term is defined as the mean of likelihoods of the other phonemes. We found that this was stable after we tested some another definitions, for example, using maximum likelihood among phonemes. The phoneme confidence likelihood $p_i(X_1^2)$ is defined by using the sigmoid function as follows.

$$p_i(X_1^2) = \frac{1}{1+\exp[-a\{C_i(X_1^2)+b\}]},$$

where a and b are constant. $p_i(X_1^2)$ is taken between 0 and 1, it approaches 1 when the likelihood of the current phoneme model is relatively larger than that of other phoneme models, it approaches 0 otherwise. Constant a in the sigmoid function means an inclination, and is set from the experiments. Figure 2 shows the logarithm of the sigmoid function for a few values of a. This function is generally 0 in the positive value area of the horizontal axis for any a but has smaller negative value in the negative value area of the horizontal axis for larger a. Therefore, a can control the weight of the confidence likelihood. About constant b, statistics of confidence measure were taken from actual speech data, and the minimum value of each phoneme model was set as b.

2.3. Use Past Records of Confidence Likelihood

Some indistinct phonemes might be included in a utterance even if it is possible to understand the whole utterance. Therefore, phoneme confidence likelihood does not necessarily obtain accurate value. Since it is dangerous to use only the reliability of the phoneme, we also use weighting the phoneme confidence likelihood by some records of confidence likelihood for each phoneme.

The confidence likelihood obtained in each phoneme is maintained, and its record is left by propagating simultaneously with accumulated logarithm likelihood. The accumulated logarithm likelihood is weighted by the records of confidence likelihood at each phoneme terminal.

$$\hat{L}_i(X_0^2) = L_i(X_0^2) + \alpha \times \frac{1}{M+1}\sum_{j=0}^{M} l_{i-j},$$

where l_{i-j} is the jth past record of the ith phoneme, and M is the number of records. The record of confidence likelihood is not used when $M = 0$.

3. MINIMUM ERROR DISCRIMINATIVE TRAINING

We used Minimum Error Discriminative Training (MCE) [6] to obtain more powerful acoustic models. MCE trains acoustic models to be able to discriminate between one another. On the other hand, the performance of the phoneme

confidence likelihood is more accurate if acoustic models are more distinguishable from one other. Therefore, MCE training might make more accurate for the phoneme confidence likelihood.

Here, we explain about MCE algorithm in brief. In parameter set Λ of acoustic model, the discriminant function is defined as $g_k(X,\Lambda)$, logarithm likelihood for class k of observation vector X is identified, and the misclassification function is defined as

$$d_k(X,\Lambda) = -g_k(X,\Lambda) + G_k(X,\Lambda),$$

where

$$G_k(X,\Lambda) = \log\left[\frac{1}{K-1}\sum_{j,j\neq k}\exp\{\eta g_j(X,\Lambda)\}\right]^{1/\eta},$$

is according the log likelihood of competition candidates for class k, where K is a number of competition candidates, and η is constant. The class loss function is defined as the sigmoid function

$$l_k(X;\Lambda) = \ell(d_k) = 1/(1+e^{-\beta(d_k+\gamma)}),$$

where β and γ are constant. The parameter Λ is updated by

$$\Lambda_{t+1} = \Lambda_t - \epsilon_t V_t \nabla l_k(X;\Lambda)|_{\Lambda=\Lambda_t},$$

where ϵ_t is a small positive real number, and V_t is a positive definite matrix. The parameter Λ is updated by controlling small change $\nabla l_k(X;\Lambda)|_{\Lambda=\Lambda_t}$ with ϵ_t and V_t.

In training process, N-best candidates of vocabulary-free speech recognition are used as competition candidates. Each candidates are divided into phonemes, and the parameters Λ are estimated and updated for each phonemes.

4. EXPERIMENTS

4.1. Experimental Setup

For experimental condition, we used the frame length of 32 ms and the frame shift of 8 ms at 12 kHz sampling frequency. 16th selective LPC cepstrum and 16th Δ cepstrum and Δ power were used with the analysis condition. Context dependent phoneme models including 450 states with four mixed distributions was used as the acoustic model covering 27 Japanese phonemes. These models are trained by Baum-Welch algorithm firstly, and are trained by MCE algorithm secondly.

For training data to Baum-Welch algorithm, we used a set of 5,240 common Japanese words and a set of 216 phonetically-balanced words uttered by 10 persons for each of male and female in the A set of the ATR database, and used 503 sentences uttered by 30 males and 34 females in the database of Acoustical Society of Japan.

For training data to MCE algorithm, we used a set of a set of 216 phonetically-balanced words uttered by 20 males and 20 females in the A and C set of the ATR database,

The proposed algorithm was evaluated by word recognition experiments by 1,202 words which contains 100 city name and the station name uttered by 5 men and 4 women.

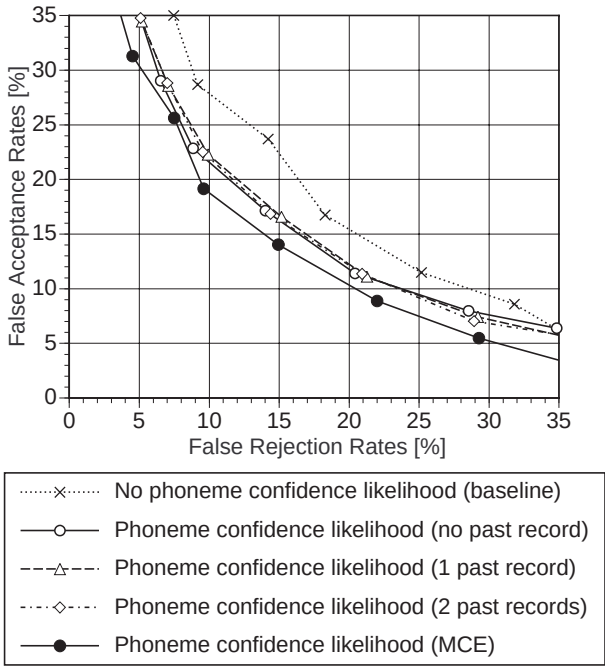

Figure 3: False acceptance rates vs. false rejection rates.

In the evaluation for unknown words, we used the 216 phonetically balanced words from the C set of ATR data base uttered by 10 males and 10 females, who were open speaker.

The center state of each phoneme model was used the calculation of $g_i(X_t)$.

4.2. Experimental Results

The rejection was decided according as threshold of normalized likelihood of a candidate at the end of utterance. The false acceptance rates versus the false rejection rates is shown in Figure 3 , and the word recognition rates versus the false rejection rates is shown in Figure 4 as an experimental result when the rejection threshold is changed.

"No phoneme confidence likelihood (baseline)" means conventional way that likelihood is normalized by the likelihood of the first candidate of vocabulary-free speech recognition without phoneme confidence likelihood. "Phoneme confidence likelihood (no past record)" means the proposed method using phoneme confidence likelihood without any past records, and "1 past record" or "2 past records" means the proposed method using phoneme confidence likelihood with one or two past records. "MCE" means the acoustic models trained by MCE algorithm were used.

In the figures, the coefficient of the sigmoid function is $a = 5.0 \times 10^{-5}$ for models not trained by MCE algorithm, and $a = 1.0 \times 10^{-4}$ for models trained by MCE algorithm. Here, the coefficient was $\alpha = 1.0$ when phoneme confidence likelihood was added.

First, in Figure 3, it is shown that the accuracy is better for the curve to approach the origin. This figure shows that the improvement of accuracy was obtained by using confidence likelihood. The equal error rate was improved

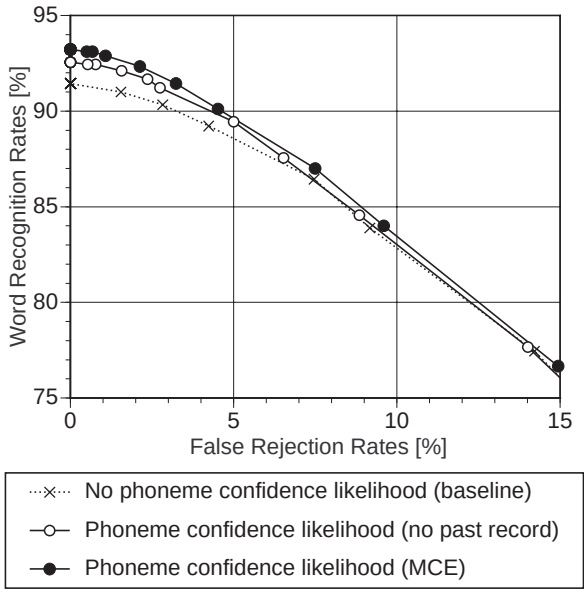

Figure 4: Word recognition rates vs. false rejection rates.

by 2%. (Table 1). Furthermore, MCE training improved by 4% compared to the baseline performance. The word recognition rate was improved by 5% as shown in Table 2. More 3% improvement was obtained by MCE training.

Second, Figure 4 shows the recognition rates of words in the vocabulary for the false rejection rates. The recognition rate is almost equal or is slightly improved even when the rejection performance is raised as shown. The performances of proposed method with past records of phoneme confidence likelihood were similar to the method without any past records, so they were not plotted. Table 3 shows that the word recognition rate without rejection was improved by 14.0 % in the error reduction rate. Moreover, 18.6% error reduction rate was obtained by MCE training compared to the baseline performance. It can be said that the phoneme confidence likelihood can improve accuracy at each phoneme, so some words are recognized correctly.

Any improvement was hardly obtained for the method using preceding phoneme confidence likelihood compared to the method without any preceding confidence likelihoods, though it is improved a little in the area where the false rejection rate was high.

Furthermore, The proposed method using acoustic models trained by MCE algorithm improved the rejection performance. The accuracy of phoneme confidence likelihood can be improved by MCE training.

5. CONCLUSION

We proposed a new method for task independent rejection of unknown words. The rejection accuracy improvement of unknown words was achieved by introducing the phoneme confidence likelihood at each phoneme during the search process. Furthermore, the recognition accuracy was improved or almost equal. Because the proposed method could impose penalty on segments of the candidates and could verify more accurate the entire utterance. The rejection performance was further improved by MCE training algorithm because the phoneme confidence likelihood was made more accurate by this algorithm.

Table 1: Equal error rates.

	Equal error rates [%]
No confidence likelihood (baseline)	18.0
Confidence likelihood (no past record)	16.0
Confidence likelihood (no past record, MCE training)	14.0

Table 2: Word recognition rates at the equal error rate.

	Recognition Rates [%]
No confidence likelihood (baseline)	71.0
Confidence likelihood (no past record)	75.0
Confidence likelihood (no past record, MCE training)	78.0

Table 3: Word recognition rates in case of no rejection.

	Recognition Rates [%]
No confidence likelihood (baseline)	91.4
Confidence likelihood (no past record)	92.6
Confidence likelihood (no past record, MCE training)	93.2

6. REFERENCES

[1] R. A. Sukkar and C. -H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition," IEEE Trans. Speech and Audio Processing, vol. 4, no. 6, pp. 420–429, 1996.

[2] M. G. Rahim, C. -H. Lee and B. -H. Juang, "Discriminative Utterance Verification for Connected Digits Recognition," IEEE Trans. Speech and Audio Processing, vol. 5, no. 3, pp. 266–277, 1997.

[3] E. Lleida, R. C. Rose, "Efficient Decoding and Training Procedures for Utterance Verification in Continuous Speech Recognition," Proc. ICASSP96, pp. 507–510, 1996.

[4] K. Kita, T. Ehara and T. Morimoto, "Processing Unknown Words in Continuous Speech Recognition," IEICE, Trans. vol. E74, no. 7, pp.1811–1816, 1991.

[5] K. Itou, S. Hayamizu and H. Tanaka, "Processing Unknown Words in Continuous Speech Recognition," IEICE, SP91-96, 1991. (in Japanese)

[6] C. -S. Liu, C. -H. Lee, W. Chou, B. -H Juang and A. E. Rosenberg, "A Study on Minimum Error Discriminative Training for Speaker Recognition," J. Acoust. Soc. Am., 97(1), pp.637–648, January 1995.