SUBWORD-BASED MINIMUM VERIFICATION ERROR (SB-MVE) TRAINING FOR TASK INDEPENDENT UTTERANCE VERIFICATION

Rafid A. Sukkar

Lucent Technologies Bell Laboratories
2000 N. Naperville Rd., Naperville, IL 60566, USA
sukkar@lucent.com

ABSTRACT

In this paper we formulate a training framework and present a method for task independent utterance verification. Verification-specific HMMs are defined and discriminatively trained using minimum verification error training. Task independence is accomplished by performing the verification on the subword level and training the verification models using a general phonetically balanced database that is independent of the application tasks. Experimental results show that the proposed method significantly outperforms two other commonly used task independent utterance verification techniques. It is shown that the equal error rate of false alarms and false keyword rejection is reduced by more than 22% compared to the other two methods on a large vocabulary recognition task.

1. INTRODUCTION

One of the main features of subword-based speech recognition is that, if the acoustic subword models are trained in a task independent fashion, then the recognizer can reliably be applied to many different tasks without the need for retraining. In such a case, only the language model needs to be updated. Recent advances in speech recognition technology have enabled the development of very large vocabulary systems, where it is almost impossible to use anything but subword-based acoustic modeling. With any deployable speech recognition system comes the need for utterance verification to reliably identify and reject out-of-vocabulary speech and extraneous sounds. Task independent utterance verification is therefore very desirable to complement task independent subword-based recognition.

Certain methods for task independent utterance verification have been proposed. For example in [1] an "on-line garbage" likelihood is computed and a likelihood ratio is then formed between the "on-line garbage" likelihood and the likelihood of the recognized word, phrase, or sentence. In [2] a linear discriminator is defined and trained to construct a subword level verification score that is incorporated into a string (sentence) level verification score. Another method that has been used is based on forming a likelihood ratio test between the likelihood of a free subword decoder and the likelihood of the recognized sentence [2,3].

In this paper we present a new method for task independent utterance verification. This method is a generalization of the method presented in [2]. While in [2], linear discrimination is employed for the verification task, in this work we define and discriminatively train verification-specific HMMs, separate from the recognition HMMs, to perform subword level verification. We formulate a subword-based minimum verification error (SB-MVE) training and use it to train these HMMs. Verification is first performed on the subword level and then, in a second stage, on the phrase or sentence level. In this fashion, we can accomplish task independence since the verification models are subword-based and trained in a task independent mode. This SB-MVE formulation extends the word-based minimum verification error (WB-MVE) training introduced in [4].

The organization of this paper is as follows: In the next section we formulate the subword-based verification problem, and in Section 3 we describe the SB-MVE training procedure. Experimental results are given in Section 4 followed by conclusions in Section 5.

2. FORMULATION

Given input speech to an HMM recognizer, let W_k be the most likely word, or string of words obtained by Viterbi decoding. In the context of subword recognition, W_k is a concatenation of subword units which can be written as

$$W_k = s_1^{(k)}\ s_2^{(k)}\ \cdots\ s_{N_k}^{(k)} \tag{1}$$

where the subword string $s_1^{(k)}\ s_2^{(k)}\ \cdots\ s_{N_k}^{(k)}$ is the subword lexical representation of W_k, and N_k is the number of subword units comprising W_k. Assuming independence among subword units, maximum likelihood Viterbi decoding implies that we can write the likelihood of the observation sequence, $\mathbf{O}$, given W_k as,

$$L(\mathbf{O}|W_k) = \max_{t_1, t_2, \cdots, t_{N-1}} L(\mathbf{O}_{t_0}^{t_1}|s_1^{(k)})\ L(\mathbf{O}_{t_1}^{t_2}|s_2^{(k)})\ \cdots\ L(\mathbf{O}_{t_{N-1}}^{t_N}|s_{N_k}^{(k)}). \tag{2}$$

where $\mathbf{O}$ is the total observation sequence, $\mathbf{O}_{t_{j-1}}^{t_j}$ is the observation sequence between time t_{j-1} and t_j corresponding to

the speech segment for subword unit $s_j^{(k)}$, and $L(\mathbf{O}_{t_{j-1}}^{t_j} | s_j^{(k)})$ is the likelihood of the segment $\mathbf{O}_{t_{j-1}}^{t_j}$ given $s_j^{(k)}$.

Given the most likely subword string corresponding to the recognition output, W_k, we now would like to test the hypothesis that the input speech does indeed consist of W_k. To perform this utterance verification task, we employ statistical hypothesis testing by formulating a likelihood ratio test as follows:

$$T(\mathbf{O};W_k) = \frac{L(\mathbf{O}|H_0(W_k))}{L(\mathbf{O}|H_1(W_k))}, \quad (3)$$

where $L(\mathbf{O}|H_0(W_k))$ is the likelihood of the observation sequence given the null hypothesis that W_k was spoken, and $L(\mathbf{O}|H_1(W_k))$ is the likelihood of the observation sequence given the alternate hypothesis that W_k was not spoken. The hypothesis test is performed by comparing the likelihood ratio, $T(\mathbf{O};W_k)$, to a predefined critical threshold, r_k. The region $T(\mathbf{O};W_k) \geq r_k$ is called the acceptance region, and the region $T(\mathbf{O};W_k) < r_k$ is called the critical rejection region. As a result, two types of errors can occur: false rejection (Type I) errors, and false acceptance or false alarm (Type II) errors. A given critical threshold value implies certain false rejection and false alarm rates. Tradeoff between the two types of errors can be controlled by varying r_k.

Rather than dealing with the likelihood ratio directly, it is more convenient to use the log likelihood ratio which can be written as

$$\begin{aligned} G(\mathbf{O},W_k) &= \log T(\mathbf{O};W_k) \\ &= \log L(\mathbf{O}|H_0(W_k)) - \log L(\mathbf{O}|H_1(W_k)). \end{aligned} \quad (4)$$

Since W_k consists of a string of N_k subwords according to equation (1), we will represent $G(\mathbf{O};W_k)$ as an average of N_k log likelihood ratios corresponding to the individual subwords in W_k, as follows:

$$G(\mathbf{O};W_k) = \frac{1}{N}\sum_{j=1}^{N_k} \log T(\mathbf{O}_{t_{j-1}}^{t_j};\, s_j^{(k)}), \quad (5)$$

where

$$T(\mathbf{O}_{t_{j-1}}^{t_j};\, s_j^{(k)}) = \frac{L(\mathbf{O}_{t_{j-1}}^{t_j}|H_0(s_j^{(k)}))}{L(\mathbf{O}_{t_{j-1}}^{t_j}|H_1(s_j^{(k)}))}, \qquad 1 \leq j \leq N_k. \quad (6)$$

Here $H_0(s_j^{(k)})$ is the hypothesis that the segment $\mathbf{O}_{t_{j-1}}^{t_j}$ consists of the correct sound for subword $s_j^{(k)}$, and $H_1(s_j^{(k)})$ is the hypothesis that the segment $\mathbf{O}_{t_{j-1}}^{t_j}$ consists of a different sound. To simplify the notation and without loss of generality, we will drop the superscript (k) from $s_j^{(k)}$ and represent $\mathbf{O}_{t_{j-1}}^{t_j}$ as $\mathbf{O}_j$.

Since the probability densities corresponding to the likelihoods of equation (6) are not known, we will approximate them by defining and discriminatively training verification-specific HMMs for each subword in the recognizer subword set. Therefore, using the simplified notation, we can write equation (6) as

$$T(\mathbf{O}_j;s_j) = \frac{L(\mathbf{O}_j|\lambda_j)}{L(\mathbf{O}_j|\psi_j)}, \quad (7)$$

where λ_j and ψ_j are the HMM models corresponding to the null and alternate hypotheses for word s_j, respectively. Note that λ_j and ψ_j are HMMs that are different than the HMMs used during the recognition process. Considering the likelihood ratio of equation (7), we can view λ_j as a verification-specific subword model for subword s_j and ψ_j as a verification-specific anti-subword model for subword s_j. This viewpoint is underscored by the fact that we use MVE training to determine the parameters of λ_j and ψ_j. We denote the verification-specific model set for a given subword, s_j, as $V_j = \{\lambda_j, \psi_j\}$.

3. SUBWORD BASED MINIMUM VERIFICATION ERROR (SB-MVE) TRAINING

Discriminative training is employed to determine the parameters of the verification model set, V_j, for each of the subwords in the recognizer subword set. Based on the definition of the subword likelihood ratio given in equation (7), the goal of the discriminative training is to make $L(\mathbf{O}_j|\lambda_j)$ large compared to $L(\mathbf{O}_j|\psi_j)$ when there is a correct recognition, and to make $L(\mathbf{O}_j|\psi_j)$ large compared to $L(\mathbf{O}_j|\lambda_j)$ when there is a misrecognition.

We define a distance function by taking the log of the inverse subword likelihood ratio of equation (7) as follows:

$$d(\mathbf{O}_j;s_j) = -\log L(\mathbf{O}_j|\lambda_j) + \log L(\mathbf{O}_j|\psi_j). \quad (8)$$

The training procedure iteratively adjusts the parameters of V_j by minimizing $d(\mathbf{O}_j;s_j)$ in the case of a correct recognition and maximizing it in the case of a misrecognition.

The function, $d(\mathbf{O}_j;s_j)$, is optimized using the generalized probabilistic descent framework [5]. In such a framework, $d(\mathbf{O}_j;s_j)$ is incorporated into a smooth loss function that is conducive to applying a gradient descent procedure to iteratively adjust the parameters of V_j. Specifically, the loss function gives a measure of the verification error rate for a given s_j and takes the form of a sigmoid function which is written as

$$Q(\mathbf{O}_j;s_j) = \frac{1}{1 + \exp[-b\,\mu\, d(\mathbf{O}_j;s_j)]}, \quad (9)$$

where μ is a positive constant controlling the smoothness of the sigmoid function, and b is set to 1 in the case of a correct recognition and to -1 in the case of a misrecognition. The value for μ is set to 1.0 in our experiments. The loss function in equation (9) is iteratively minimized with respect to the parameters of V_j using gradient descent. In our experiments, correct recognitions are obtained during SB-MVE training by force segmenting a given sentence using its correct lexical transcription. Misrecognitions are obtained by force

segmenting a given sentence using a random lexical transcription. We set b to -1 for all subwords corresponding to a misrecognition and to 1 for all subwords corresponding to a correct recognition.

It is important to note here that task independence is accomplished by training V_j using a general phonetically balanced subword database. Given that the set of subwords remains the same, the resulting V_j can be used to perform utterance verification for any recognition task without the need for retraining.

4. EXPERIMENTAL RESULTS

This vocabulary independent utterance verification method was evaluated on a *company name* recognition task, where the goal is to recognize the name of a company out of 6963 possible names. The average number of words per company name is 3.7 words and the average number of subword units per company name is 18.9 units. The lexical transcription of the company names were obtained using a text-to-speech front end. A total of 40 context independent subword models and one silence model were used in the recognition phase. Each subword model was represented by a 3-state continuous density HMM, where the maximum number of Gaussian mixtures was set to 16. The silence model was represented by a single state HMM with 32 mixtures. The recognizer feature vector consisted of the following 39 parameters: 12 LPC derived cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, normalized log energy, and the delta and delta-delta of the energy parameter. The database used to train these recognition models consisted of 9865 phonetically balanced phrases and sentences collected over the public telephone network. Minimum classification error (MCE) training was employed to train the recognition subword models [5]. The above phonetically balanced database was also used to train the verification models using the SB-MVE training procedure described in Section 3. The verification model set, V_j, for a given subword, s_j, consists of two continuous density HMMs, λ_j and ψ_j, having a topology of 3 states with 8 Gaussian mixtures in each state. Therefore, there were a total of 80 verification HMMs corresponding to the 40 recognition subwords.

The company name database used for performance evaluation is independent of the phonetically balanced database used in the training phase. This testing database was collected over the public telephone network and consists of 11552 utterances spoken by 2500 different speakers covering the 6963 company names. Since we are evaluating the performance of an utterance verification method, we also need to define a separate database consisting of out-of-vocabulary speech. Towards this end, we used a database consisting of 10511 utterances of speakers saying their first and last names. This out-of-vocabulary database was also collected over the public telephone network.

Prior to the utterance verification stage, the recognition rate on the company name database was 93.1%. The utterance verification performance is shown in Figure 1. The top plot in this figure shows the false acceptance rate (false alarms) of the out-of-vocabulary utterances as a function of the false rejection rate of the company name utterances. Since successful recognition requires not only correct verification but also correct classification, the bottom plot shows the recognizer substitution error rate on non-rejected company names versus false rejection rate. By fixing the false rejection rate, the recognizer operating point can be determined by identifying the corresponding false alarm and substitution error rates. The SB-MVE performance at any given operating point is obtained by evaluating the verification score, $G(\mathbf{O};W_k)$, given in equation (5), and comparing the results to a predefined threshold.

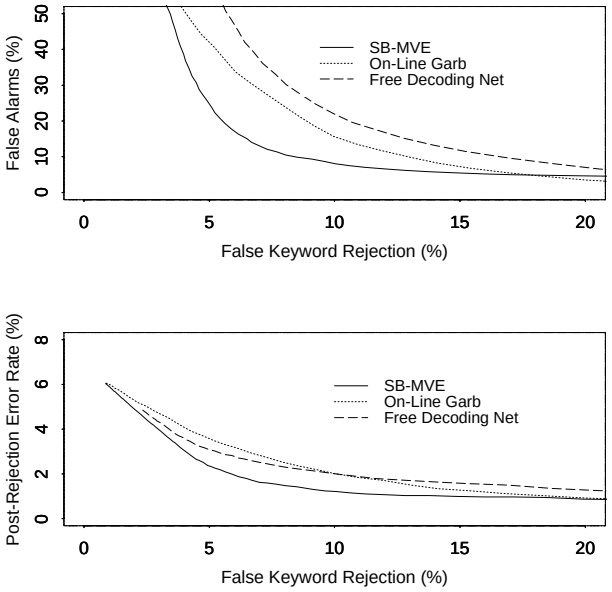

Figure 1. Utterance verification performance comparison.

Figure 1 also compares the performance of the SB-MVE method to two other utterance verification methods. The first is very similar to the on-line garbage method, proposed in [1] and also evaluated in [3]. In our experiments the on-line garbage verification score for a given recognized company name is computed by averaging the on-line garbage scores of the subwords constituting the recognized company name excluding any silence segments. It is useful to note that the SB-MVE method also excludes all silence segments when computing the verification score. Mathematically, the on-line verification score that we used to obtain the on-line garbage performance shown in Figure 1 is given by

$$R(\mathbf{O};W_k) = \frac{1}{N_k}\sum_{j=1}^{N_k}\log\frac{L(\mathbf{O}_{t_{j-1}}^{t_j}\,|\,s_j^{(k)})}{L_{on-line}(\mathbf{O}_{t_{j-1}}^{t_j})}, \qquad (10)$$

where $L_{on-line}(\mathbf{O}_{t_{j-1}}^{t_j})$ is the on-line garbage likelihood obtained by computing for every frame the average likelihood score of the M top scoring states and summing over the segment $\mathbf{O}_{t_{j-1}}^{t_j}$. In our experiments M was set to 16.

The second method with which we compared employs a verification HMM network parallel to the HMM network defined by the company name lexicon. The verification network acts as the out-of-vocabulary network and consists of a self loop of all the subword and silence models in the recognizer model set. In effect, this out-of-vocabulary network results in a "free-subword" maximum likelihood HMM decoding of the input speech utterance. The verification score is defined as a log likelihood difference between the likelihood of the recognized company name and the likelihood of the non-keyword network.

It is clear from Figure 1 that the SB-MVE method significantly outperforms the other methods on two fronts. First, the SB-MVE method results in a false alarms rate that is consistently lower than the other two methods. Second, the post-rejection substitution error rate is also lower, implying that the SB-MVE method is more likely than the other two methods to reject substitution errors, a very desirable property for many applications. Fixing the false rejection rate at 7.0% and 10.0%, Table 1 shows a comparison of the utterance verification performance of the three methods. These results were obtained from the plots of Figure 1. Another point of interest is the equal error rate (EER) of false alarms and false rejections. Table 2 compares the equal error rates of the three methods and shows that the SB-MVE method results in an EER that is 22.0% lower than the on-line garbage method and 32.8% lower than the free decoding network method.

Table 1. Utterance verification performance

Method	False Rej. (%)	False Alarms (%)	Post-Rej. Error (%)
SB-MVE	7.0	13.0	1.6
On-Line Garb.	7.0	29.1	2.8
Free Decoding	7.0	36.8	2.5
SB-MVE	10.0	8.0	1.2
On-Line Garb.	10.0	15.7	2.0
Free Decoding	10.0	22.1	2.0

It is important to note that the free decoding method is much more computationally intensive that either the SB-MVE or the on-line garbage method. On the other hand the difference in computational complexity between the SB-MVE and on-line garbage method is relatively small. The SB-MVE method does, however, require additional model storage capacity compared to the other two methods for storing the verification-specific models.

Table 2. Equal error rate comparison

Method	EER (%)
SB-MVE	9.2
On-Line Garb.	11.8
Free Decoding	13.7

5. CONCLUSIONS

In this paper we formulated a framework and presented a method for task independent utterance verification. Verification-specific models were defined and trained using minimum verification error training. To accomplish task independence, the verification was performed on the subword level and the verification models were trained using a general task independent database that consisted of phonetically balanced phrases and sentences. Comparing this proposed method to two other commonly used utterance verification methods showed that the SB-MVE method reduces the equal error rate of false rejections and false alarms by at least 22%. In addition, the SB-MVE method consistently resulted in lower post-rejection substitution error rate, implying that the SB-MVE method was more likely to reject substitution errors compared to the other two methods.

6. ACKNOWLEDGEMENTS

The author acknowledges helpful discussions with Chin H. Lee regarding minimum verification error training. The author also acknowledges the valuable software support provided by Carl Mitchell.

7. REFERENCES

[1] H. Bourlard, B. D'hoore, and J.-M. Boite "Optimizing recognition and rejection performance in wordspotting systems," *Proc. ICASSP '94*, pp. 373-376, Vol. 1, April 1994.

[2] R. A. Sukkar, C. H. Lee, and B. H. Juang, "A vocabulary independent discriminatively trained method for rejection of non-keywords in subword-based speech recognition," *Proc. Eurospeech '95*, pp. 1629-1632, Sept. 1995.

[3] R. C. Rose and E. Lleida, "Speech recognition using automatically derived acoustic baseforms," *Proc. ICASSP '97*, pp. 1271-1274, April 1997.

[4] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C. H. Lee, "Utterance verification of keyword strings using Word-Based Minimum Verification Error (WB-MVE) training," *Proc. ICASSP '96*, Vol. I, pp. 518-521, May 1996.

[5] W. Chou, B. H. Juang, and C. H. Lee, "Segmental GPD training of HMM-based speech recognizer," *Proc. ICASSP '92*, Vol. I, pp. 473-476, April 1992.