A TRAINING PROCEDURE FOR VERIFYING STRING HYPOTHESES IN CONTINUOUS SPEECH RECOGNITION

R. C. Rose, B. H. Juang, and C. H. Lee
AT&T Bell Laboratories, Murray Hill, NJ 07974-0636

ABSTRACT

A procedure is proposed for verifying the occurrence of string hypotheses produced by a hidden Markov model (HMM) based continuous speech recognizer. Most existing procedures verify word hypotheses through likelihood ratio scoring procedures computed using ad hoc approximations for the density of the alternative hypothesis in the denominator of the likelihood ratio statistic. The discriminative training procedure described in this paper attempts to adjust the parameters of the null hypothesis and the alternate hypothesis models to increase the *power* of a hypothesis test for utterance verification. The training procedure was evaluated for its ability to detect a twenty word vocabulary in a subset of the Switchboard conversational speech corpus. Experimental results show that the use of this procedure results in significant improvement in the word verification operating characteristic, as well as an improvement in the overall system performance.

1 INTRODUCTION

This paper describes a procedure for verifying the occurrence of vocabulary words in continuous speech utterances. The work addresses a problem that is common to many speech recognition applications including telecommunications based speech recognition services. Since it is generally very difficult to design a speech recognition application so that all users' utterances are constrained to be within a well defined domain, it is necessary to have a mechanism for dealing with unexpected input from the user.

Unexpected input can appear for many tasks in many different forms. Techniques for verifying decoded vocabulary words and detecting out-of-vocabulary words have been proposed for dictation tasks [1], dialog tasks [2], command spotting [3], and keyword spotting in conversational speech [4, 5, 6]. The motivation for many of these techniques comes from a hypothesis testing procedure known as the likelihood ratio test (LRT). The application of the LRT is described in Section 2 as involving a decision rule which is based on the ratio between the likelihood of a decoded vocabulary word in context, referred to as the null-hypothesis model, and the likelihood of an alternative hypothesis model.

There have been many attempts to implement likelihood ratio scoring procedures using various different ad hoc approximations for the density of the alternative hypothesis in the denominator of the likelihood ratio statistic. One common approach to forming the alternative hypothesis has been to run a network of hidden Markov subword acoustic models in parallel with the word based search [7, 1, 4, 2]. However, except for [4] and [6], there is no mechanism in any of the existing systems to design the system according to a criterion which is directly related to the ability of the system to verify hypothesized keywords. The training procedure summarized in Section 3 attempts to adjust the parameters of the null hypothesis and the alternate hypothesis models to increase the *power* of the hypothesis test.

2 VERIFYING WORD OCCURRENCES IN CONTINUOUS SPEECH UTTERANCES

The proposed procedure, in its simplest form, provides a mechanism for hypothesis testing. The manner in which it is applied to testing whether or not a particular vocabulary word was uttered as part of a continuous speech utterance is described in Figure 1. This is a two stage procedure which both generates a hypothesized word occurrence and also verifies the word occurrence using a statistical hypothesis testing procedure. The hidden Markov model (HMM) based continuous speech recognition (CSR) "target network" in Figure 1 takes as input a sequence of T mel-frequency cepstrum feature vectors $Y = y_1, \ldots, y_T$ representing a speech utterance which may contain both within-vocabulary and out-of-vocabulary words. The output of the CSR target network includes the labels associated with both the hypothesized keyword and the non-keyword "filler" models.

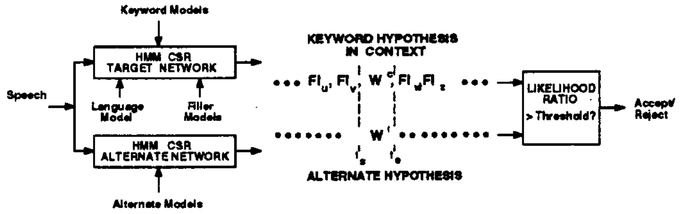

Figure 1: A speech recognition system with the capability for verifying the word hypotheses produced by a continuous speech recognizer. A word hypothesis W^C is shown decoded in the context of surrounding models FI_k. An "alternate network" exists only to generate the alternative hypotheses for the statistical likelihood ratio test.

The system in Figure 1 relies on a likelihood ratio test (LRT) to verify the hypothesized keyword. An LRT is a statistical hypothesis test which is designed to determine whether or not a sequence of feature vectors were generated by a given family of probability densities. The form of the density $p(Y \mid \lambda)$ is assumed to be known. For example, $p()$ might correspond to a hidden Markov density, and λ might correspond to the set of HMM parameters. The likelihood ratio test

$$\mathcal{L}(Y, \lambda^C, \lambda^I) = \frac{p(Y \mid \lambda^C)}{p(Y \mid \lambda^I)} \qquad (1)$$

tests the hypothesis that the sequence of observations Y was generated by the model λ^C corresponding to the hypothesized vocabulary word W^C versus Y having been generated by a model λ^I corresponding to the alternate hypothesis W^I. The "alternate network" shown in Figure 1 exists only

0-7803-2431-5/95 $4.00 © 1995 IEEE

to provide the probability $p(Y \mid \lambda^I)$ of the feature vectors for the alternative hypothesis W^I corresponding to the target hypothesis W^C. The alternate hypothesis can be parameterized in many ways. Section 4 describes a preliminary set of experiments for verifying target word hypotheses using simple alternate hypothesis models.

3 A TRAINING PROCEDURE FOR IMPROVING THE POWER OF HYPOTHESIS TEST

There are several problems with using the likelihood ratio in Equation 1 to test the hypothesis that a word was spoken in a given utterance. The first problem is that the model parameters for a particular word are not known, but must be estimated from training data. The second problem is that, for almost any modeling problem, the assumptions concerning the form of the density $p(Y \mid \lambda^C)$ are only approximate. It is well known in the hypothesis testing literature, that as the modeling assumptions become less accurate, the power of the test shown in Equation 1 rapidly degrades. Third, it is unclear as to what class of alternatives should be used for specifying the alternative model. The proposed procedure provides a mechanism for dealing with these problems. A training technique is described for adjusting the parameters λ^C and λ^I in the likelihood ratio test to maximize a criterion that is directly related to Equation 1. The notion behind the method is that adjusting the model parameters to increase this confidence measure on training data will provide a better measure of confidence for verifying word hypotheses during the normal course of operation for the service.

The training criterion adjusts model parameters to *minimize* a function of the logarithm of the inverse of the likelihood ratio test given in Equation 1:

$$S_C(Y) = \log P(Y \mid \lambda^I) - \log P(Y \mid \lambda^C)\,. \tag{2}$$

Keyword hypotheses corresponding to both actual keyword utterances ($Y \in C$: true hits) and to imposter utterances ($Y \in I$: false alarms) are presented to the training procedure. The goal is to decrease the weighted average value of $S_C(Y)$ for true hits ($Y \in C$) and to increase the average value of $S_C(Y)$ for false alarms ($Y \in I$).

In practice, the distance in Equation 2 is approximated for a hypothesized keyword by first obtaining the sequence of states that were most likely to have generated the decoded observation vectors for that word $Y = \vec{y}_1, \ldots, \vec{y}_T$. Using the Viterbi algorithm, one can obtain the sequence of states for both the correct and imposter models $\Theta^C = \theta_1^C, \ldots, \theta_T^C$ and $\Theta^I = \theta_1^I, \ldots, \theta_T^I$ respectively. The local log probability of an observation vector $\vec{y}_t$ for state θ_t is given as $\log b_{\theta_t}(\vec{y}_t) = \log p(\vec{y}_t \mid \theta_t, \lambda)$. A local log likelihood ratio distance can be defined as

$$d(\vec{y}_t) = \log b_{\theta_t^I}(\vec{y}_t) - \log b_{\theta_t^C}(\vec{y}_t)\,. \tag{3}$$

The word based log likelihood ratio distance $S_C(Y)$ in Equation 2 can be approximated using these local distances as

$$D(Y) = \sum_{t=1}^{T} d(\vec{y}_t)\,. \tag{4}$$

A gradient descent procedure is used to iteratively adjust the model parameters as new utterances are presented to the training procedure. An error measure corresponding to the error measure used for generalized probabilistic descent is used to provide a well-behaved model estimation procedure whose estimation error is well correlated with word verification performance [8]. A word level error,

$$L(Y,\Lambda) = \ell(\delta(Y,C,I)D(Y)), \tag{5}$$

is used here where $\ell(x) = \frac{1}{1+\exp(-\gamma x)}$, and

$$\delta(Y,C,I) = \begin{cases} 1 & Y \in C \\ -1 & Y \in I \end{cases}\,. \tag{6}$$

In Equation 5, $\Lambda = \{\lambda^C, \lambda^I\}$ represents both the target word model and alternate word model parameters. The indicator function $\delta(Y,C,I)$ in Equation 6 dictates that the average of the distance in Equation 3 be minimized for utterances Y corresponding to correct hypotheses and maximized for imposter utterances.

The observation probabilities, $b_i()$, are defined as mixtures of Gaussians of the form

$$b_i(\vec{y}_t) = \sum_{j=1}^{M} c_{i,j} f_{i,j}(\vec{y}_t)\,, \tag{7}$$

where $c_{i,j}$ is the mixture weight for the jth mixture Gaussian and $f_{i,j}(\vec{y}_t)$ are Gaussian densities with diagonal covariance matrices. This implies that the $f_{i,j}(\vec{y}_t)$ are completely defined by their means $\mu_{i,j}[k]$ and standard deviations $\sigma_{i,j}[k]$, where $i = 1, \ldots, M$, and $k = 1, \ldots, K$. Hence, the total set of parameters to be estimated are $\lambda_C = \{c_{i,j}^C, \mu_{i,j}^C[k], \sigma_{i,j}^C[k]\}$ corresponding to the correct class models, and $\lambda_I = \{c_{i,j}^I, \mu_{i,j}^I[k], \sigma_{i,j}^I[k]\}$ corresponding to the imposter or alternate class models.

For any parameter ϕ_i associated with a state i of either the correct or alternate hypothesis model the gradient $\nabla L(Y,\Lambda)$ of the error measure defined in Equation 5 can be written in terms of the partial derivative

$$\frac{\partial L(Y,\Lambda)}{\partial \phi_i} = \frac{\partial L(Y,\Lambda)}{\partial D(Y)} \frac{\partial D(Y)}{\partial \phi_i} \tag{8}$$

$$= \sum_{t=1}^{T} \gamma \ell(D(Y))(1 - \ell(D(Y)))n(\theta_t^C, \theta_t^I, i, Y) \frac{\partial \log b_i(\vec{y}_t)}{\partial \phi_i} \tag{9}$$

$$\tag{10}$$

where

$$n(\theta_t^C, \theta_t^I, i, Y) = \begin{cases} 1 & Y \in C, i = \theta_t^I \\ -1 & Y \in I, i = \theta_t^C \\ 1 & Y \in C, i = \theta_t^C \\ -1 & Y \in I, i = \theta_t^I \end{cases}\,. \tag{11}$$

The indicator function $n(\theta_t^C, \theta_t^I, i, Y)$ defined in Equation 11 defines the direction of the gradient depending on whether the utterance corresponds to a correctly hypothesized keyword or a false alarm and whether ϕ_i is a parameter from a correct or alternate hypothesis model. The expressions for the partial derivatives of $\log b_i(\vec{y}_t)$ when $b_i()$ is of the form given in Equation 7 can be found in many references including [8, 9, 10].

Figure 2 illustrates how the discriminative training procedure is applied. First, one or more word hypotheses are generated along with the associated word endpoints by the CSR target network. Second, the word hypothesis decoded for an utterance is labeled as corresponding to an actual occurrence of a vocabulary word (true hit) or a false alarm. Third, the distance given in Equation 2 is computed using the probabilities estimated from the target and alternate hypothesis models. Finally, the gradient update shown in Figure 2 is performed on the expected error $E\{L(Y,\Lambda)\}$ as

$$\Lambda_{n+1} = \Lambda_n - \epsilon \nabla E\{L(Y,\Lambda)\}\,, \tag{12}$$

where ϵ is a learning rate constant and the expectation in Equation 12 is computed by summing over all observation vectors in the training set labeled as keyword hypotheses.

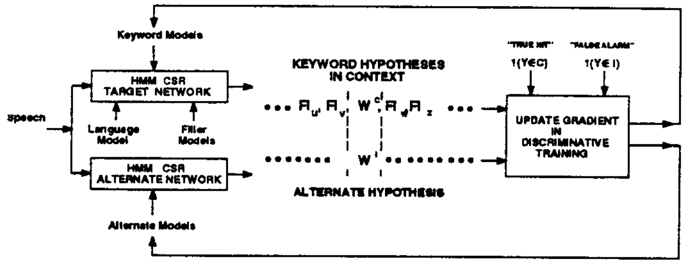
Keyword Models
"TRUE HIT"
"FALSE ALARM"
KEYWORD HYPOTHESES IN CONTEXT
1(Y∈C)
1(Y∈I)
HMM CSR TARGET NETWORK
speech
••• $R_{u'} R_{v'}$ $W^{c'}$ $R_{w'} R_{z}$ •••
UPDATE GRADIENT IN DISCRIMINATIVE TRAINING
Language Model
Filler Models
HMM CSR ALTERNATE NETWORK
••••••• $W^{I'}$ •••••••••
ALTERNATE HYPOTHESIS
Alternate Models

Figure 2: Block diagram illustrating a gradient descent based training procedure for verifying hypothesized vocabulary words decoded in continuous utterances.

4 EXPERIMENTS

The principal goal of this work was to demonstrate the effectiveness of discriminative training techniques in verifying word hypotheses generated during continuous speech recognition. The purpose of the experiments described in this section was to accomplish three things. First, they attempted to validate the basic notion that, using these techniques, it is possible to increase the average "separation" between correct and imposter keyword hypotheses according to a likelihood ratio distance of the type given in Equation 1. Second, the experimental results give insight into how the alternate word hypothesis models, or "anti-keywords", should be parameterized for best performance on the conversational speech task described below. Finally, we can measure the effect of these techniques on the performance of the entire system which includes both the CSR network and the word hypothesis verification procedure. It is hoped that verifying word hypotheses produced by the recognizer will allow us to simplify the structure of the CSR network without sacrificing speech recognition performance.

The experiments were performed according to the following procedure. First, maximum likelihood hidden Markov models were trained for a tied Gaussian mixture continuous speech recognizer from the "Credit Card" subset of the Switchboard conversational speech corpus [11]. A vocabulary of twenty keywords, selected from a total vocabulary of approximately 2200 words, was used in the word verification experiments. The keyword vocabulary was comprised of the words *account, american_express, balance, bank, card, cash, charge, credit, credit_card, discover, dollar, hundred, interest, limit, mastercard, money, month, percent, twenty, visa.*

Second, in order to generate the keyword hypotheses necessary for the model adjustment procedure described in Section 3, speech recognition was performed on 7283 utterances, also taken from the Switchboard corpus. A very simple null grammar CSR network was used including the keyword vocabulary and a network of 43 subword models to represent out–of–vocabulary utterances. Strings containing either true or false decodings of keywords were input to discriminative training along with strings containing alternative hypothesis models decoded from the same utterances. This resulted in a total of 1225 word hypotheses corresponding to correct keyword occurrences (ranging from 20 occurrences for the keyword *twenty* to 416 occurrences for the keyword *card*) and 2725 word hypotheses corresponding to "imposters" or incorrectly recognized keywords.

The observation densities in the continuous speech recognizer shown in Figure 2 were tied Gaussian mixtures defined over mel–frequency cepstrum and difference cepstrum observation vectors. This implies that a single set of M Gaussian densities in Equation 7 are tied to all HMM states in the network. Each vocabulary word was expanded according to tri–phone subword acoustic units. Triphones were represented by three state left–to–right HMM's. In the word hypothesis verification procedure, the keyword based tri–phone HMM parameters, λ_C, were reestimated according to the gradient defined in Equation 10.

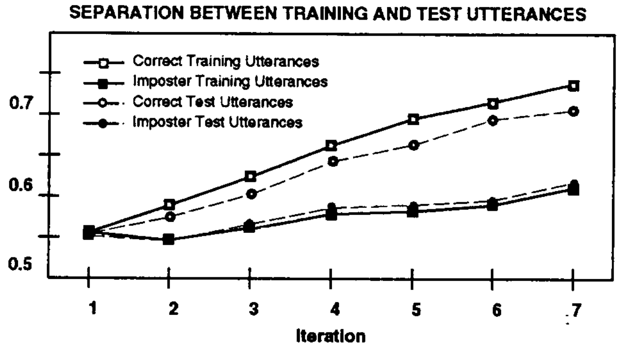

Figure 3: Average values of a a smooth function of $D(Y)$, as defined in Equation 4, computed separately for correct and imposter keyword hypotheses.

An alternate hypothesis model was defined for each keyword based tri-phone. Each alternate model was a single state mixture Gaussian. Because of the relatively small amount of training data available for most keywords, it was necessary to tie all alternate models in each keyword. As a result, an anti-keyword model in these experiments corresponds to a single mixture of Gaussians. Anywhere from 8 to 128 mixtures were used per keyword. Finally, an additional effort was made to reduce the total number of alternate hypothesis model parameters by tying all parameters so that λ_I corresponds to a single pool of mixtures.

The training procedure is meant to adjust model parameters to increase the average likelihood ratio for utterances where a keyword is present, and decrease the likelihood ratio for utterances where a keyword hypothesis was generated by the recognizer but in fact contained no keyword. To verify that the procedure does operate to accomplish this goal, plots of a smooth function of $D(Y)$, as defined in Equation 4, averaged over the training and test sets and computed after each iteration of the training procedure, are given in Figure 3. The curves shown in the figure represent the average $1 - \ell(D(Y))$ computed from utterances of the keyword *money*. The solid curves correspond to the training utterances and the dashed curves correspond to the test utterances. It is clear from Figure 3 that the relative separation between correct and imposter keyword hypotheses improves as a result of model adjustment over both the training data set and test data set.

Figure 4 displays operating characteristic curves which describe the performance of the second stage word hypothesis testing procedure. Both plots represent the performance as the probability of correct detection of a keyword hypothesis generated by the recognizer versus the probability of false acceptance of an imposter keyword hypothesis. Operating curves for each keyword are computed separately and averaged over all twenty keywords to give the curves shown in Figure 4. The three different curves display the operating characteristics for three separate scoring procedures. In the absolute scoring procedure, the duration normalized likelihood scores obtained from the speech recognizer were used for scoring. The maximum likelihood performance was computed as the log likelihood ratio between the vocabulary word model and an alternate model consisting of Gaussian mixtures trained using maximum likelihood estimation from imposter utterances in the training data. Finally, the last set of curves were computed as the log likelihood ratio between the target word models and the anti–keyword models obtained after eight iterations of the discriminative training procedure. The plots on the left and right of Figure 4 describe the operating characteristics using alternate models with 16 and 128 mixture components respectively. It is clear from the figure that the discriminative training proce-

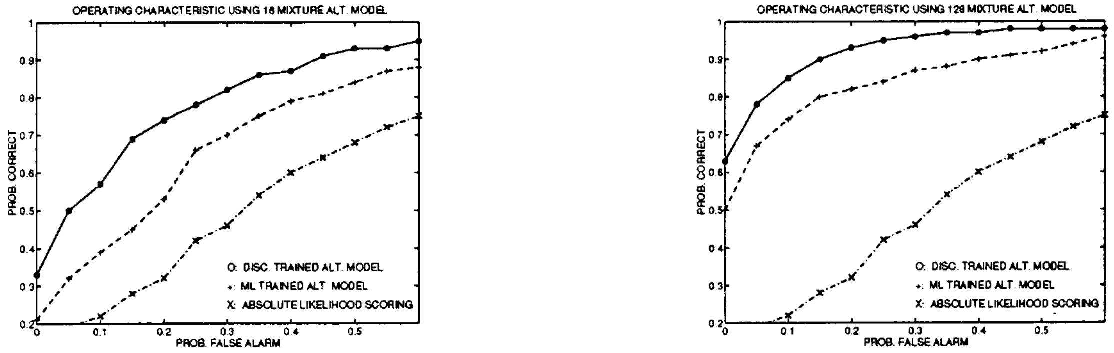

Figure 4: Word hypothesis testing operating characteristic curves computed for the test utterances taken from a subset of the Switchboard conversational speech corpus. The plot on the left corresponds to an alternate hypothesis model containing 16 component Gaussian mixtures and the curve on the right corresponds to a 128 component mixture alternate model.

dure results in significant improvement in word verification performance in both cases.

Additional experiments were performed to investigate the use of a single tied set of Gaussians to represent the alternate model for all 20 keywords. While a slight improvement in the average word verification operating characteristic was obtained, the effect was small in comparison to word dependent alternate models. Adjusting target model parameters with the discriminative procedure was also investigated. Here again, slight performance improvement was obtained, but the effects were minor when compared with the effects of alternate model adjustment.

It is important to consider the total performance of the entire system shown in Figure 1 which includes both the CSR network and the second stage word hypothesis verification procedure. The CSR network alone produced 2095 false keyword detections for 1.53 hours of speech in the test set amounting to a total of 68 false alarms per keyword per hour for the 20 keyword vocabulary. Furthermore, the average keyword detection rate was 88.7An operating point for the combined system can be obtained by weighting the figures from the CSR network with the operating characteristic of the discriminatively trained word verification system shown on the right in Figure 4. For example, at 10% probability of false alarm, a combined system operating point of 76.2% probability of detection at 6.8 false alarms per keyword per hour is obtained. This is important because the performance is roughly equivalent to the performance obtained from a much higher complexity CSR system with no word hypothesis verification [11].

5 SUMMARY

One interpretation of the proposed training procedure is that it modifies the definition of the target and alternate hypothesis parameter spaces in order to increase the power of a hypothesis test defined in 1. As such, this procedure should be applicable to any detection problem where representative exemplars of target and alternate hypothesis classes can be obtained from training. The procedure was presented here as part of a system that hypothesizes the occurrence of keywords in continuous speech and verifies the occurrence of the keywords through a statistical hypothesis test. There are two contributions related to the new training procedure: The first is a mechanism for adjusting model parameters to optimize a criterion which is directly related to the hypothesis test used for verifying a keyword occurrence. The second is the use of a set of word dependent alternate hypothesis models which are designed specifically for the purpose of representing the range of keyword false alarms that are expected to be generated by the HMM based CSR target network.

REFERENCES

[1] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large vocabulary speech recognition system," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. 305–308, April 1991.

[2] S. R. Young and W. H. Ward, "Recognition confidence measures for spontaneous spoken dialog," *Proc. European Conf. on Speech Communications*, pp. 1177–1179, September 1993.

[3] R. A. Sukkar and J. G. Wilpon, "A two pass classifier for utterance rejection in keyword spotting," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. II451–II454, April 1993.

[4] R. C. Rose, "Discriminant wordspotting techniques for rejecting non–vocabulary utterances in unconstrained speech," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, March 1992.

[5] R. P. Lippmann and E. Singer, "Hybrid neural-network/HMM approaches to wordspotting," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. I565–I568, April 1993.

[6] C. Torre and A. Acero, "Discriminative training of garbage model for non-vocabulary utterance rejection," *Proc. Int. Conf. on Spoken Lang. Processing*, June 1994.

[7] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, April 1990.

[8] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Proc.*, pp. 3043–3054, December 1992.

[9] W. Chou, B. H. Juang, and C. H. Lee, "Segmental gpd training of a hidden markov model based speech recognizer," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. 473–476, April 1992.

[10] J. K. Chen and F. K. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Trans on Speech and Audio*, vol. 2, pp. 206–216, January 1994.

[11] R. C. Rose, "Definition of acoustic subword units for word spotting," *Proc. European Conf. on Speech Communications*, pp. 1049–1052, Sept. 1993.