

Continuous Speech Recognition and Verification based on a Combination Score

Binfeng Yan, Rui Guo, Xiaoyan Zhu

State Key Laboratory of Intelligent Technology
and Systems, Tsinghua Univ., China
ybf@e1000e.cs.tsinghua.edu.cn

Abstract

In this paper we present a speech recognition and verification method based on the integration of likelihood and likelihood ratio. Speech recognition and verification is unified in one-phase framework. A modified agglomerative hierarchical clustering algorithm is adopted to train the alternative model used in speech verification. In the process of decoding likelihood ratio is combined with likelihood to get the combination score for searching the final results. Our experimental results showed that false-alarm rate get decreased a lot with only slight loss in accuracy rate.

1. Introduction

Keyword spotting systems are widely applied in human-machine communication, which find a good trade-off between the flexibility of conversation strategy and the performance of speech recognizer. When a keyword is detected, the utterance segment without keywords or with low confidence level should be rejected simultaneously. The rejection to out-of-vocabulary (OOV) words is implemented by speech verification.

R. A. Sukkar[1] proposed modularized two-phase verification method, which consisted of speech recognition and speech verification. After the first stage, speech recognition, the results of recognizer are sequentially verified in the second phase, where a confidence score is computed and those results with low score are rejected. The shortcoming of this method is that the inaccuracy of segmenting the utterance in the candidates of speech recognition will debase the feasibility of verification a lot[1][2][4]. Furthermore, the time cost is accordingly increased.

To overcome it, M. Rahim[3] and C. -H. Lee[5] used one-phase verification method based on statistical hypothesis test. The recognition phase and verification phase are combined into one process. Likelihood ratio, which represents the confidence level, is used to score searching for the candidates. So the final result given by the recognizer is the one with the highest confidence level. It is a compromise between the performance of keyword detection and that of rejection to OOV words. As a result, one-phase method is not as good as two-phase method in the performance of recognition, but does better in the verification.

Here we propose a novel one-phase method for recognition and verification in attempt to get a better trade-off in both performances. In the searching process, likelihood and likelihood ration are computed for each frame of speech feature vector and the weighted sum of both is calculated as the combination score. Thus the final results are those candidates

with high likelihood and high confidence level. So its performance of keyword recognition excels the traditional one-phase a lot. At the same time, there is only a little impact on the verification process.

The key problem of likelihood ratio is the selection and training of the alternative models. In this paper a modified agglomerative hierarchical clustering algorithm is used to train these models, which is based on the divergence between HMM models.

This paper is organized as follows. In Section 2 we discuss the training of Alternative models. In section 3 we then describe the one-phase keyword recognition and verification. Detailed experimental results are showed in Section 4. Finally we summarize the major features of our one-phase recognition and verification method.

2. MODEL TRAINING

2.1. Alternative Model

The theoretic base of speech verification in statistics is statistical hypothesis test. Generally there are two hypotheses:

- H_0 : Keyword KW_k occurs in the segment of speech and is recognized correctly.
- H_1 : Keyword KW_k didn't occur in the segment, or it occurred but was incorrectly recognized.

The ratio of $P_k(O|H_0)$ and $P_k(O|H_1)$ was used to judge whether the result should be accepted or rejected. Here the following is the decision-rule in speech verification where the logarithm of ratio is used:

$$PR(k) = \frac{P_k(O|H_0)}{P_k(O|H_1)} \begin{cases} > \tau_k, & H_0 \text{ was accepted} \\ \leq \tau_k, & H_1 \text{ was accepted} \end{cases} \quad (1)$$

λ_k is the acoustic model of keyword KW_k and its corresponding alternative model is λ_k^a . τ_k is the threshold for decision.

The selection and training of alternative models are essential to this verification method. In some systems, the average score of N-Best results was used to represent the score of alternative model [6]. Another method like [5], the data of those keywords most similar to the keyword were used to train its alternative model. In this paper, we proposed a data-driven method for training alternative model. The first step is to classify all the keyword models into a reasonable number of categories using the modified agglomerative hierarchical clustering algorithm; in the second step, for each keyword model we used all the

sample data of other models that are in the same category with this keyword to train its alternative model.

2.2. Clustering Algorithm

As we told before, the clustering algorithm determined the performance of alternative models. To date agglomerative hierarchical clustering algorithm is used for training acoustic models in many systems.

First of all, the similarity between two HMMs should be quantified by finding a distance measure between them. In our experiments a formula based on divergence was adopted:

$$d(i, j) = \frac{1}{N_i + N_j} \left(\sum_k \log \left[\frac{P(O_{ik} | \lambda_i)}{P(O_{jk} | \lambda_j)} \right] + \sum_k \log \left[\frac{P(O_{jk} | \lambda_j)}{P(O_{ik} | \lambda_i)} \right] \right) \quad (2)$$

where O_{ik} is the k_{th} training example of Model λ_i , $P(O_{ik} | \lambda_i)$ is the isolated emission probability of O_{ik} by model λ_i , and N_i is the number of training examples of model λ_i .

Though agglomerative hierarchical clustering algorithm is really an efficient way to classify samples, it can not control the sample number of each class effectively because of its greedy character, so there may exist a wide diversity among classes. To overcome this shortcoming we modified the standard algorithm in two points. The basic idea of the modification is to limit the sample number of each class in a proper range during the clustering procedure.

Firstly, the upper threshold N_{\max} was used to control the maximum cluster number in one class. After finding the most similar pair of classes, a criteria of maximum bound is proposed to determine whether these two classes can be merged. That is, only two classes which satisfy the following conditions

$$d(i, j) = \min_{\forall p, q \in I} \left\{ d(p, q) \mid N_p + N_q \leq N_{\max}, i < j \right\} \quad (3)$$

can be merged into one class. In the above equation, I is the set of the labels of classes, $d(i, j)$ is the distance between Class i and Class j , and N_p is the sample number of Class p .

Secondly, when the automatic clustering procedure completes, adjustment is made to those classes whose sample numbers are below the lower threshold N_{\min} . Such a class is merged into another class with the smallest distance to it and at the same time the sum of their sample numbers should not be greater than N_{\max} . The adjustment is iteratively done to each class until the final number of samples per class is constrained between N_{\min} and N_{\max} . After these two modifications, the clustering procedure can be supervised so as to avoid having too few/too many samples per class.

3. RECOGNITION AND VERIFICATION

3.1. The Combination Score in Frame Layer

In the search process, the likelihood score of the t_{th} frame of feature o_t , $\gamma_i(o_t)$ is the emitting probability density, $l_i(o_t)$.

$$\gamma_i(o_t) = l_i(o_t) = \sum_{k=1}^M \zeta_{ik} p_{ik}[o_t], \quad \sum_{k=1}^M \zeta_{ik} = 1 \quad (4)$$

ζ_{ik} is the weight of k_{th} gaussian distribution, $p_{ik}[o_t]$. At the same time, by using the alternative model, the likelihood ratio LR , can be computed,

$$LR_i(o_t) = \log l_i(o_t) - \log l_i^a(o_t) = \log p(o_t | \lambda_i) - \log p(o_t | \lambda_i^a) \quad (5)$$

λ_i is the parameter of the current model, while λ_i^a is the parameter of the alternative model. But LR can not be used directly with $LR_i(o_t)$, otherwise the pruning would be impacted severely. So it is normalized using the sigmoid function of LR :

$$\log CM = \Xi(LR) = \log \frac{1}{1 + \exp(-\alpha \cdot (LR - \beta))} \quad (6)$$

α and β are respectively constants which control the slope and the position of sigmoid function. For the convenience of computation, it can be simplified like this:

$$\Xi(LR) = \begin{cases} \alpha \cdot (LR - \beta), & \text{if } LR < \beta - 6/\alpha \\ 0, & \text{if } LR > \beta + 6/\alpha \\ \Xi(LR), & \text{otherwise} \end{cases} \quad (7)$$

Thus the value of $\log CM$ is between 0 and 1.

By setting the proper weight, the combination score of likelihood and confidence level could be computed:

$$\log \hat{\gamma}_i(o_t) = w_1 \log l_i(o_t) + w_2 \log CM_i(o_t) \quad (8)$$

w_1 and w_2 are the weights set with experience. Thus the optimal path found would be the word sequence with the highest combination score $\hat{\gamma}_i(o_t)$.

3.2. The Combination Score in Other Layers

The above is the combination score in the frame-level. It should be noticed that there is wide variety among the confidence level of frames. So during searching procedure, local frame-layer confidence score may influence the path to an unexpected degree. To avoid it, we take the likelihood ratio score into consideration only when the syllabic boundary is met. If syllable n ends at the k_{th} , the likelihood ratio at the syllable layer can be computed as follows:

$$SLR_n(o_t) = 1/\tau \cdot \sum_{t-\tau < t \leq t} LR_i(o_t) \quad (9)$$

τ is the duration of syllable n and $LR_i(o_t)$ is the likelihood ratio of the i_{th} state of the l_{th} frame of feature. Further, the combination score can be computed at the keyword level if more smooth of the usage of likelihood ratio is necessary. If keyword k ends at the t_{th} frame, the likelihood ratio score at the word level can be defined:

$$KLR_k(o_t) = 1/N \cdot \sum_n SLR_n(o_t) \quad (10)$$

4. EXPERIMENTAL RESULTS

4.1. Experimental Platform

The platform is a continuous voice control system whose speech recognition engine is a keyword spotter. The whole system consists of three components: speech recognition module, speech verification module and conversation control module.

The speech data comprise CIDS (Chinese Isolate Words, Digits and Syllables) corpus and place name corpus which was recorded from spontaneous utterances of 10 people. There are totally 100 names in the corpus, 50 within the task and 50 outside the vocabulary. Speech signals digitized at 11.025 kHz are 16 bits and mono-track. The speech feature vectors used in all of the experiments are 39 dimensions, including 12 MEL-Frequency Cepstrum coefficient (MFCC), plus its 1st and 2nd order differences, as well as 1 normalized energy plus its 1st and 2nd order differences. For each keyword model and corresponding alternative model a semi-continuous left-to-right HMM is trained. The size of coding book is 64 and the number of states in HMM is data-dependent.

4.2. Experimental Results and Analyses

We used 40 people's utterances as training data and 20 people's utterance as testing data. The testing set included 100 words, in which 50 words are keywords and other 50 words are OOV words.

4.2.1. The performance of Alternative Models

The DET(Detection Error Tradeoff) figure can be used to evaluate the performance of alternative models by comparing the overlapped area of the keywords and the OOV words. Here we compared alternative models trained with two different methods. One is to use modified agglomerative hierarchical clustering algorithm (Method 1), and the other, a typical traditional method, is to use the samples of 5 models that are most similar to the keyword model (Method 2).

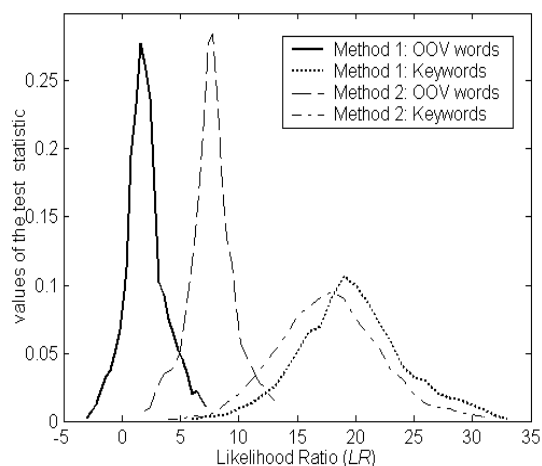


Figure 1: The DET of alternative models

From the figure 1, we can clearly see that the overlapped area of our method is much smaller than the traditional method. It is an intuitional revelation of the strong ability of our method in discriminating keywords and OOV words.

4.2.2. The Performance Of Speech Verification

First we compared the performance in speech recognition of two different methods. One was our method using the combination score, and the other was traditional method using likelihood score only. Here we set the weight $w_1=1$, and got the results corresponding to different values of w_2 in Table 1. When $w_2=0$, it was the traditional method.

Table 1: The detection rate of the combination score method

w_2	0	0.2	0.4	0.6	0.8	1.0
Detection Rate (%)	97.2	96.3	96.0	91.6	87.7	82.3

Because the models were trained with maximum likelihood method, with the combination of likelihood ratio, the detection rate of keywords decreased necessarily. To get the tradeoff, we set $w_2=0.4$ in the following experiments.

Afterwards we compared the performance of speech verification. The verification method was to firstly compute the average score of all frames. Then it was compared with the preset threshold to decide whether to reject the candidate or not. By adjusting the threshold ratio in the keyword verification stage, we got a series of detection rate at different false alarm rates, so a Receiver Operating Characteristics (ROC) curve was made and showed in Figure 2. In the above figure, the combination score method shows its superiority after the verification stage.

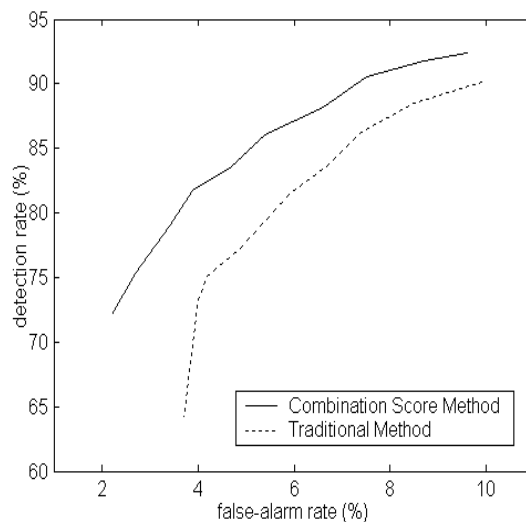


Figure 2: The verification performance of two methods

5. SUMMARY

A modified one-phrase speech recognition and verification method was described in this paper. The likelihood ratio was used in decoding process to compute a combination score for searching the candidates. We also proposed modified agglomerative hierarchical clustering algorithm to train the alternative models. Experimental results showed that the discriminative ability of alternative models had improved markedly and verification performance of our method excelled the traditional method.

6. REFERENCES

- [1] R. A. Sukkar and J. G. Wilpon, "A two pass classifier for utterance rejection in keyword spotting", *Proc. IEEE ICASSP 93, 1993*, pp. 451-454
- [2] Binfeng Yan, Rui Guo, Xiaoyan Zhu, Bo Zhang, "An approach of keyword spotting based on HMM", *IEEE Proceedings of the 3th World Congress on Intelligent Control and Automatic, P.R.China. PP. 2757-2759, 2000*
- [3] M. Rahim, C. -H. Juang, "Discriminative utterance verification for connected digits recognition", *IEEE Trans. Speech Audio Processing, Vol. 5, pp. 266-277, May 1997*
- [4] Tsiporkova, T. Vanpoucke, F. and Van hamme, H., "Evaluation of Various Confidence-Based for Isolated Word Rejection", *Proc. ICASSP, 2000*
- [5] C. H. Lee, "A tutorial on speaker and speech verification", *Proc. NORSIG-98, Vigso, denmark, June 1998, pp. 9-16*
- [6] H. Bourlard, B. D'hoore, and J.-M. Boite, "Optimizing recognition and rejection performance in word spotting systems," *ICASSP-1994, Vol. I, p: 373-376*