Rejection in Speech Recognition Based on CDCPMs

Mingxing Xu, Fang Zheng, Wenhu Wu
Speech Lab, Dept. of Comp. Sci. & Tech, Tsinghua Univ., Beijing 100084, China
[xumx, fzheng]@sp.cs.tsinghua.edu.cn, fzheng@cenpok.net

Abstract

Rejection is important for two-stage speech recognition. In this paper a new rejection method based on Center-Distance Continues Probability Model (CDCPM) is proposed, named CAP, which is the feature percentage in critical area (CAP) according to the probability theories. Also another rejection method named recognition score gap (RSG) is proposed to cooperate with CAP. Experiments are done across a large real-world database with 20,000 test samples. The average recognition accuracy is 86.33% with 3.46 candidates number on an average.

1. Introduction

In two-stage speech recognition or keyword spotting systems, rejection is a very important stage. In the first-stage, often as many candidates as possible are selected so that the correct candidate is contained and the accuracy is guaranteed. In the acceptation/rejection stage, efficient methods should be adopted to reject false hits in order to lower down the false alarm rate.

In many speech recognition applications, such as keyword spotting, the acceptation/rejection is performed using statistical hypothesis testing [Rahim 1995, Rose 1995, Sukkar 1995]. Moreover, many of the proposed methods use the recognition models themselves in formulating the verification likelihood ratio. In such a case, the recognition models are used for both recognition and rejection, and recognition/rejection performance tradeoff have to be considered. And there are some applications formulate the rejection test by constructing and disciminatively training verification-specific models to estimate the distributions of the null and alternate hypotheses. All of these methods need extra modeling and training.

In this paper a new rejection method based on Center-Distance Continues Probability Model (CDCPM) [Zheng 1996] is proposed, named the feature percentage in the critical area (CAP). The parameters used in CAP are different from those in the first recognition stage. According to this method the acceptation/rejection stage evaluate each recognition candidate individually. In experiments, we find the correct candidate's position has relation with the distribution of the recognition scores of candidates, which introduces another rejection method named recognition score gap (RSG) to cooperate with CAP. To evaluate the rejection efficiency, we give four definitions, probability of correctness (PC), probability of occurrence (PO), average recognition accuracy (ARA) and average candidates number (ACN).

This paper is organized as follows. In section 2, the theories of CAPs are discussed. In section 3, the method RSG is discussed. In section 4, we discuss how to use CAP and RSG cooperatively. In section 5, the experimental results are analyzed . In section 6, the conclusion is presented.

2. The Feature Percentage in Critical Area (CAP)

The CDCPM is a modified version of HMM with left-to-right architecture [Yang 1995], which eliminates the initial probability distribution and the probability transition matrix. The feature space of each state is divided into several sub-spaces described by one Center-Distance Normal (CDN) distribution [Zheng 1997a]. These sub-spaces can be estimated by some clustering method according to some kind of criterion [Zheng 1997b].

For the Normal distribution

$$N(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2},\quad x\in(-\infty,\infty) \tag{1}$$

there are 95% samples fall into the critical area $[\mu-2\sigma,\mu+2\sigma]$. Similarly, for the normal-derived Center-Distance Normal (CDN) distribution

$$N_{CD}(x;\mu,\sigma)=\frac{2}{\pi D}e^{-(x-\mu)^2/\pi D^2},\quad x\in[0,\infty) \tag{2}$$

there will be about 95% samples' center-distances fall into the critical interval $[0, 2.5D]$ because $\sigma=\frac{\sqrt{2\pi}}{2}D\approx 1.25D$.

Hereafter, denote the model by $\Lambda=\{\mu_{nm},D_{nm}|1\le n\le N,1\le m\le M\}$ and the observation feature vector sequence by $O=\{o_1,o_2,\cdots,o_T\}$, where N is the number of states

in the model, M is the number of densities in a model state and T is the utterance length in frame, three CAP rejection methods are discussed in details [Zheng 1997c].

2.1 CAP1

Define the acceptation/rejection score (ARS) for any feature vector o_t as

$$Score(o_t|\Lambda)=\begin{cases}1, & \max\limits_{1\le m\le M}\left\{d(o_t,\mu_{nm})|o_t \in Density(n,m)\right\}\in[0,kD_{nm}]\\ 0, & otherwise\end{cases} \tag{3}$$

In Equ. (3), $Density(n,m)$ denotes the m-th density in n-th state for the model Λ and $d(\cdot,\cdot)$ is the distance measure for feature vectors. The critical area controlling parameter k=2.5 (or other value). Based on Equ. (3), the ARS of feature sequence O with model Λ is defined as

$$Score(O|\Lambda)=\frac{1}{T}\sum_{t=1}^{T}Score(o_t|\Lambda) \tag{4}$$

Obviously, this definition satisfies the limitation $Score(O|\Lambda)\in[0,1]$. And if we set the acceptation/rejection thresholds such as $TSH_h \ge 0.5$ and $TSH_l \le 0.5$, the category of O will be determined by

$$O\begin{cases}\in\Lambda, & \text{if } Score(O|\Lambda) > TSH_h\\ \notin\Lambda, & \text{if } Score(O|\Lambda) < TSH_l\end{cases} \tag{5}$$

The thresholds can be chosen empirically or by the analysis of the training data.

2.2 CAP2

In Equ. (2), the ARS is related only to the maximally matched density, if all densities are considered, the vector ARS can be defined as

$$Score(o_t|\Lambda)=\begin{cases}1, & \sum\limits_{m=1}^{M}Score\left(o_t|Density(n,m)\right) > TSH_{NUM}\\ 0, & otherwise\end{cases} \tag{6}$$

where

$$Score(o_t|Density(n,m))=\begin{cases}1, & d(o_t,\mu_{nm})|o_t \in Density(n,m))\in[0,kD_{nm}]\\ 0, & otherwise\end{cases} \tag{7}$$

and TSH_{NUM} is a number ranging from 1 to M-1.

Equ.s (6), (7) with (4), (5) gives another CAP acceptation/rejection quantity which is different from CAP1 only in that CAP1 considers the nearest density while CAP2 considers all densities inside one state.

2.3 CAP3

In CAP1 and CAP2, the ARS of each feature vector is a two-value function. Our experiments show that different number of densities for different states performs better [Zheng 1997b]. In that situation, CAP1 and CAP2 can not reflect the differences. We think it better to consider all the 2-value scores for the feature vector with every density. Thus the ARS for the sequence O with the given model is defined as

$$Score(O|\Lambda) = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M(n(t))} Score(o_t | Density(n(t), m))}{\sum_{n=1}^{N} M(n)} \tag{8}$$

where $n(t)$ denotes the state that the t -th feature vector belongs to and $M(n)$ is the density number in state n.

Equ. (8) leads to CAP3, where there is no need to set up TSH_{NUM} thresholds for all states as in CAP2.

3. The Recognition Score Gap (RSG)

The matching score provided by recognition module indicates how the utterance matches the model. In order to include the correct result, the first-stage recognition module often outputs the K best candidates. The scores of Top K candidates contain the information of the position of the correct answer. In our experiments, we find the score differences between adjacent candidates are useful for the acceptation/rejection stage.

At first we calculate the mean and variance of the candidates' scores, then we look for the relation between these values and the position of correct candidate. The experimental results show that a large score gap appears between the right candidates and wrong ones, as shown in figure 1.

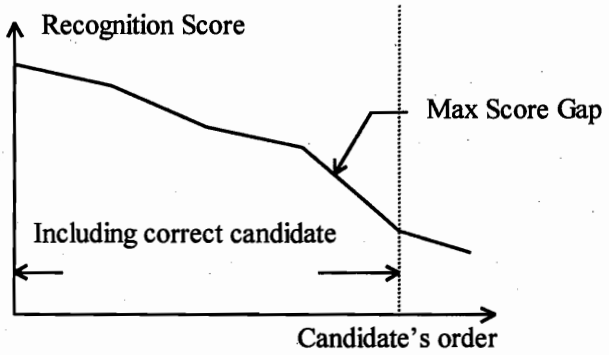


Figure 1 The Curve of Recognition Scores

A specified threshold is used to determine the number of reserved candidates, only those with score higher than the threshold are reserved.

Let $RS(k)$ denotes the recognition score of the k-th candidate, $k = 0,1,\cdots,K-1$, where K is the number of candidates that the first-stage module outputs. Define the Recognition Score Gap (RSG) for the k-th candidate as

$$RSG(k) = RS(k) / RS(0) \tag{9}$$

Setting up a threshold TSH for RSG, we can throw off those candidates whose RSG's are smaller than the given threshold as

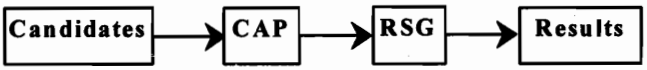
$$k\begin{cases}\in \Lambda, & \text{if } RSG(k) \geq TSH \\ \notin \Lambda, & \text{if } RSG(k) < TSH\end{cases} \tag{10}$$

4. Using Several Rejection Methods Jointly

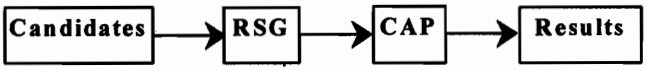
In above rejection methods, CAPs calculate rejection score for each candidate independently. In other word the result score is only dependent on the utterance's feature vector sequence. Clearly this kind of rejection operation is a filtration procedure that will perhaps output no candidates. Whereas the RSG method computes the rejection score according to the relation among candidates provided by recognition module at the first stage. Usually these candidates are sorted by recognition scores. We use RSG method to delete candidates from the rejection point in the candidates list. So the result of this kind of rejection operation is just like a bobtail with the first candidate at least.

Because CAP and RSG are based on different theories, there is no correlation between these two rejection methods. We can use them jointly. For example, we can:

(1) use CAP at first, then RSG, as



(2) or use RSG at first, then CAP as



5. Experimental Results

Some experiments have been done across a real-world spontaneous database. The speech data are taken from telephone network and sampled at 8KHz. The samples are 13-bit linear PCMs expanded from A-law codes. The database consists of speech data uttered by 200 people, and the amount is about 4GB. 10^{th} order mel-frequency cepstral (MFCC) analysis is performed on 32 ms speech window every 16 ms. Auto-regressive analysis is also performed on 5 adjacent frames of MFCC vectors. The MFCCs and their corresponding auto-regressive coefficients are the features used for the CDCPM [Zheng 1996, 1997a] in this paper. The SRUs are 419 Chinese syllables. The first stage outputs 10 candidates sorted according to recognition scores to the next stage. We test 20,000 utterances. The experimental result is shown in figure 2.

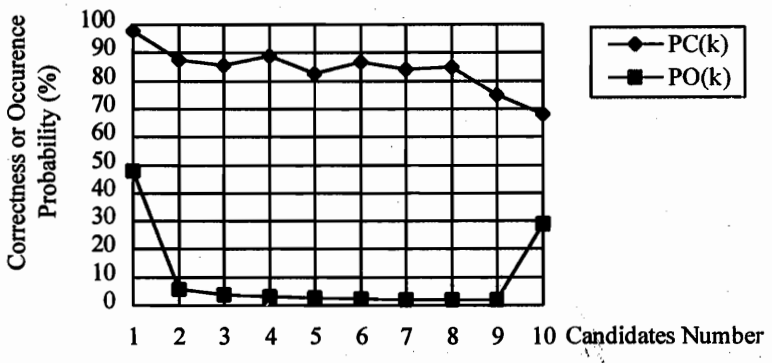


Figure 2 Rejection Results

The number of rejected candidates is 110,837. There are 302 correct candidates in them. The error rate is 0.27%.

In order to evaluate the performance of the rejection methods, we will give some definitions. Denote the total number of testing samples (e.g. 20,000) by TN, the total number of samples where k candidates are outputted in the acceptation/rejection stage by $C(k)$, and the total number of samples where k candidates including the correct one are outputted in the acceptation/rejection stage $R(k)$. Obviously $R(k)$ samples must be included in those $C(k)$ samples. Definitions are given as follows.

(1) Probability of Correctness is defined as

$$PC(k) = \frac{R(k)}{C(k)} \tag{11}$$

which specifies the correctly detection probability conditioned on k candidates are outputted.

(2) Probability of Occurrence is defined as

$$PO(k) = \frac{C(k)}{TN} \tag{12}$$

which indicates the probability of outputting k candidates, where

$$\sum_{k} PO(k) = 1 \tag{13}$$

(3) Average Recognition Accuracy (ARA) can be calculated by

$$\hat{ARA} = \sum_{k=1}^{10} PC(k) * PO(k) \tag{14}$$

which describes the total performance of rejection method.

(4) Average Candidates Number (ACN) can be calculated by

$$ACN = \sum_{k=1}^{10} k * PO(k) \tag{15}$$

which also describes the total performance of rejection method.

From the data in Figure 2, we can calculate ARA and ACN. The results are ARA = 86.33% and ACN = 3.46 < 4.

6. Conclusion

We introduce two new rejection methods, named CAP and RSG, to decrease the number of candidates given in the first recognition stage. We also define some parameters to evaluate the total performance of rejection method. The experimental results show that CAP and RSG can provide significantly better performance. They have following four features:

(1) CAP and RSG are easy to calculate without extra training and modeling as in some other rejection methods.

(2) The *ACN* has been decreased to 4 after rejection procedure, which indicates the rejection method proposed is an efficient method.

(3) The *PO* is a U-type curve. This typical polarization feature fits the distribution of correct candidates in recognition results.

(4) The incorrectly rejection rate is as low as 0.27%, which shows the rejection methods we used can work well with CDCPM. This good performance also shows that the CDCPM can model Chinese speech well.

References

[1] **Rahim, M.G., Lee, C.-H., Juang, B.H.,** "Discriminative utterance verification for connected digits recognition," Proc. *Eurospeech'95*, pp. 529-532, Sept.1995

[2] **Rose, R.C., Juang, B.H., Lee, C.H.,** "A training procedure for verifying string hypotheses in continuous speech recognition", Proc. *ICASSP'95*, Vol. I, pp. 281-284, May 1995

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

[4] **Zheng, F., Wu, W.-H., Fang, D.-T.,** "CDCPM with its application in speech recognition," *(Chinese) J. of Software*, 7: 69-75, Oct. 1996

[5] **Yang, X.-J., et al** Speech Digital Signal Processing. Beijing: Electronic Industry Publishing House, 1995 (in Chinese)

[6] **Zheng, F., Chai, H.-X., Shi, Z.-J., Wu, W.-H., Fang, D.-T., (1997a)** "A real-world speech recognition system based on CDCPMs," *Int'l Conf. on Computer Processing of Oriental Languages (ICCPOL'97)*, 1: 204-207, Apr. 2-4, 1997, Hong Kong

[7] **Zheng, F., Xu, M.-X., Wu, W.-H., (1997b)** "The description of the intra-state feature space". Somewhere in this ROCLing X proceeding.

[8] **Zheng, F., (1997c)** "Studies on approaches of keyword spotting in unconstrained continuous speech," Ph.D. Dissertation. Beijing: Dept. of Comp. Sci. & Tech., Tsinghua Univ., 1997