EXPLOITING NON-TARGET REGION INFORMATION FOR CONFIDENCE MEASURE BASED ON BAYESIAN INFORMATION CRITERION

Cong Liu, Yu Hu, Xiong-Guo Lei, Zhi-Guo Wang, Li-Rong Dai, Ren-Hua Wang

iFly Speech Lab, Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei
yylhbt@mail.ustc.edu.cn yuhu@iflytek.com xglei@iflytek.com
zgwang@iflytek.com lrdai@ustc.edu.cn rhw@ustc.edu.cn

ABSTRACT

In this paper appropriate confidence measures (CMs) are investigated for Mandarin command word recognition, both in the so-called *target region* and *non-target region*, respectively. Here the *target region* refers to the recognized speech part of command word while the *non-target region* refers to the recognized silence part. It shows that exploiting extra information in the *non-target region* can effectively complement the traditional CM which usually focus on the *target region*. Furthermore, when analyzing the *non-target region* in a more theoretical way, where Bayesian information criterion (BIC) is employed to locate more precise boundary in the *non-target region*, even more improvement is achieved. In two different Mandarin telephone command word tasks, more than 20% relative reduction of equal error rate (EER) is obtained.

Index Terms— Speech Recognition, Confidence measure, Bayesian information criterion

1. INTRODUCTION

Automatic Speech Recognition (ASR) has achieved great progress in past few decades. But it is still hard for any real-world application to get robust performance because of various tough conditions such as speaker variations, background noise and accents, etc. So the ability of automatically assessing for every recognized word is very important for an ASR system. So far, researchers in this field have proposed to compute one score (preferably between 0 and 1) which is called Confidence Measure (CM), to indicate the reliability of any recognition result generated by the ASR systems [1, 2]. CM technology plays a momentous role in the stage of speech recognition and it is, nowadays, being applied in many aspects such as detection of Out-Of-Vocabulary (OOV) words, rejection of non-speech noises, instruction of performing unsupervised training, cleanup of human transcription errors in a large speech database, and so on [1, 2, 3, 4].

Generally speaking, all the methods proposed by now for computing CM score can be classified into three categories [1]. The first one is to calculate CM score based on the so-called *predictor features* which are generated during the decoding procedure, or some combination of them. The second method is the well known *word posterior probability* (WPP), as is indicated in the standard maximum *a posteriori* (MAP) decision rule. However, it is hard to estimate its normalization term in the denominator precisely. And different methods, ranging from simple *filler*-based method [5] to complex word-graph-based approaches [6], have been proposed to approximate it. The third CM can be framed in Utterance Verification (UV),

and partially be calculated based on the Likelihood Ratio Testing (LRT), which belongs to the non-Bayesian approach.

In this paper, we focus on CMs for Mandarin command word application. As is known to all, CM is the postprocessing of ASR system and bases on the output information of the decoder, such as boundary, word sequences, likelihood and so on. In this work we define the so-called *target region* and *non-target region* based on the output boundary and recognized words. The *target region* covers the speech part of recognized command word, i.e., one of all the command words in the recognition grammar. The margins surrounding the middle *target region* are called *non-target region*, actually composed of those silence part. Usually people pay more attention to CM calculated based on only the *target region*, and we try to exploit additional information from the *non-target region*. For *target region* in command word recognition, we employ the WPP-based CM in our study. However, the word graph is usually too sparse to warrant reliable WPP estimation, due to the characteristic of strict grammar for command word recognition. So simple method, such as *filler*-based approximation is preferred. While for *non-target region*, to verify the existence of speech or real silence, it is more appropriate to introduce the hypothesis testing to perform likelihood ratio testing. However, it seems heuristic and the assumed hypotheses are not precise enough when regarding the whole *non-target region*. So the Bayesian information criterion is further employed to deal with the *non-target region* more precisely, which leads to a more theoretical and robust way to exploit extra information in the *non-target region*.

The structure of this paper is organized as follows: In section 2, we present the detail of why and how to exploit extra information in *non-target region*. In section 3, we discuss about choosing appropriate CM methods for different regions. In section 4, Bayesian information criterion (BIC) is employed for more precise locating of *non-target region*. The experiments are given in section 5. In section 6, we present our conclusions.

2. EXPLOITING EXTRA INFORMATION IN NON-TARGET REGION

As referred above, based on the output of recognition, the whole speech can be segmented into the so-called *target region* and *non-target region*. If the recognized boundary is precise enough that the *target region* contains all real speech, then calculating CM score for only *target region* is enough. In other words, there may be little extra information when exploiting *non-target region*. However, in practical command word application, sometimes we will encounter such case which is showed in Fig. 1. Note that for *non-target region* in this study, it is actually a special silence model which is trained

using silence data plus some other types of noise (also to improve the robustness). So speech may be absorbed into silence part to form the result in Fig. 1.

Fig. 1. Illustration of Target Region and Non-target Region

Because of confusion between words, usually an OOV error or sometimes an in-vocabulary error occurs, to make part of the initial speech to be absorbed into the *non-target region*, such as "mei" in above figure. In this case, if confidence score is calculated only for the *target region* to represent the whole sentence, e.g., "wo", it is usually high enough to be accepted, which leads to a false accept.

So we believe that the recognized *non-target region* has complementary information for confidence measure, and we will also introduce proper CM for those parts. Based on CMs of both regions, we can perform combination between them as follows:

$$S_{all} = f(S_{tr}, S_{nr}) \quad (1)$$

where S_{tr} and S_{nr} mean CM score for *target region* and *non-target region* respectively, S_{all} is the combined score of the whole sentence, and f is the function of combination, which is a simple linear interpolation with linear weighting in this paper.

3. DISCUSSION ON CM METHODS FOR DIFFERENT REGIONS

To conduct combination between *target region* and *non-target region* based CMs referred above, we consider to choose appropriate CM methods for the two regions, respectively.

For the core *target region*, the well-known WPP-based CM is introduced here:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \quad (2)$$

where $P(X|W)$ is the probability of the observation X given corresponding word hypothesis W and $P(W)$ is the prior probability of W (for command word recognition it is usually fixed). However, the normalization term $P(X)$ is hard to approximate. Because the word graph does not have sufficient information for command word application, we use the simple *filler* model for approximating $P(X)$. In our experiments, the *filler* model is a Gaussian mixture model (GMM) trained using a wide range of speech and also some types of noise and side-speech, to make the approximation more robust for speech from different environments. So the final CM score of the *target region* is as (employ phone as basic unit):

$$S_{tr} = \frac{1}{N} \sum_{i=1}^N \frac{1}{T_{tr}^i} [\log p(X_{tr}^i | \lambda_{tr}^i) - \log p(X_{tr}^i | \lambda_{tr}^{filler})] \quad (3)$$

where X_{tr}^i is the speech segment recognized as the i th phone unit in *target region*, λ_{tr}^i is the acoustic model for the i th phone, λ_{tr}^{filler} is the *filler* model for all units, T_{tr}^i is the duration of the i th phone, and N is the number of phones in current *target region* segment.

However, for the *non-target region*, more attention is paid to judging whether any real speech is absorbed into the silence part. For instance, if there is any real speech existing in the *non-target region*, there tends to be an recognition error (illustrated in Fig. 1). So hypotheses testing is a proper choice here. In a hypothesis testing problem, first we should bring in two complementary hypotheses, namely the *null hypothesis* H_0 and the *alternative hypothesis* H_1 , respectively. If given a segment from *non-target region* X_{nr} and assume that an ASR system recognizes it as silence which is represented by an HMM λ_{sil} . H_0 hypothesis means that X_{nr} is correctly recognized and generated from model λ_{sil} , i.e., X_{nr} is a segment of pure silence, without any speech in it. While H_1 represents that X_{nr} is wrongly recognized and does not belong to model λ_{sil} , but come from some other models. Thus, we can test H_0 hypothesis against H_1 hypothesis to make a decision that whether we should accept the fact that the segment is pure silence or not. According to the rule of Neyman-Pearson Lemma, if the probability distribution $p(X_{nr}|H_0)$ and $p(X_{nr}|H_1)$ are given, we can obtain the optimal solution through the Likelihood Ratio Testing (LRT) as follows:

$$\frac{p(X_{nr}|H_0)}{p(X_{nr}|H_1)} \begin{cases} \geq \tau & \text{judge it as } H_0 \\ < \tau & \text{judge it as } H_1 \end{cases} \quad (4)$$

where τ is the threshold with which we can make a decision for measuring confidence and we can also call it *operating point*.

Based on the LRT formulation above, we can also transform the LRT to a score of confidence measure using a monotonic one-to-one mapping function [1]. The major problem with LRT is how to model the *null* and *alternative hypothesis* appropriately. Here we use the acoustic model (the special silence model referred above) for the *null hypothesis*. And for H_1 we use an anti-model λ_{nr}^{anti} , which is trained using pure speech data. So the confidence score for the whole *non-target region* can be defined as:

$$S_{nr} = \frac{1}{T_{nr}} [\log p(X_{nr} | \lambda_{nr}^{sil}) - \log p(X_{nr} | \lambda_{nr}^{anti})] \quad (5)$$

where $p(X_{nr} | \lambda_{nr}^{sil})$ and $p(X_{nr} | \lambda_{nr}^{anti})$ stand for the probability of X_{nr} given H_0 model and H_1 model, respectively, and T_{nr} is the duration of the *non-target region*.

4. ANALYZING OF NON-TARGET REGION USING BAYESIAN INFORMATION CRITERION

During implementation, we find that sometimes the duration of the real silence part in *non-target region* is much longer than that of the speech absorbed into it, which leads to very high CM score for that silence part. In other words, the hypothesis referred in last section is not precise enough for this case, neither H_0 hypothesis nor H_1 hypothesis is right. To alleviate the problem, one way is to set a floor for score of the *non-target region* to avoid that too high *non-target* CM score will influence the total score. However, it is heuristic and needs precise tuning of the floor, so in this paper we present another

way to locate the position of the speech part which is absorbed into the silence part, and only score of the speech part located is calculated to contribute to global CM score, i.e., the hypothesis testing is only performed on the refined *non-target region* which is located from the initial silence part. In this paper we employ the famous Bayesian information criterion (BIC) for locating position of speech.

BIC is proposed to select appropriate dimension of a statistical model, e.g., polynomial regression, multi-step Markov chains, etc, by introducing a penalty to refine the maximum likelihood principle. In [7] it is proposed to use BIC for segmentation and clustering of speech. It can still be treated as a hypothesis testing problem, where the *null* hypothesis H_0 means there is no change among one speech segment (all the samples in this segment meets the same distribution), and the alternative hypothesis H_1 means some change happens among this segment (we need to use two different distribution to model these samples). If Gaussian distribution is used here, we can represent H_0 and H_1 as follows:

$$H_0 : X_1^N \sim \mathbf{N}(\mu_0, \Sigma_0) \quad (6)$$

$$H_1 : X_1^i \sim \mathbf{N}(\mu_1, \Sigma_1); X_{i+1}^N \sim \mathbf{N}(\mu_2, \Sigma_2) \quad (7)$$

where X_1^N is one speech segment with N frames, μ_0 and Σ_0 is the mean and covariance matrix of the Gaussian when data of the segment are hypothesized to meet the same distribution. While μ_1 and μ_2 are the means of the two segmented distribution, Σ_1 and Σ_2 are corresponding covariance matrices, when the alternative hypothesis is given.

So the difference of BIC value between H_1 and H_0 is:

$$\text{BIC}(i) = [L(X_1^i | \mu_1, \Sigma_1) + L(X_{i+1}^N | \mu_2, \Sigma_2)] - L(X_1^N | \mu_0, \Sigma_0) + \lambda d * \log N \quad (8)$$

where $L(\cdot)$ is the log-likelihood function of the observation given corresponding Gaussian distribution, λ is the factor of penalty, and d is the dimension of the distribution (here a diagonal covariance matrix is used for Gaussian distribution).

So we can choose the proper i as the segment point of the initial *non-target region*, or still treat it as a whole silence (there is no real speech absorbed into the *non-target region*), when the BIC value is under a preset threshold. However, BIC is not accurate when the speech segment is too short to estimate a Gaussian distribution, so during implementation we add a constraint that both sub-part after segmentation need to be no less than 10 frames.

5. EXPERIMENTS

5.1. Experimental Setup

In our experiments, 39-dimensional MFCCs (12-d static MFCC, log-energy, delta and acceleration coefficients) plus 4-dimensional pitches are extracted to train a tonal initial final model, with 3 state for all initials and 5 states for all finals, and 12 mixtures per state. There are totally about 170,000 utterances in the training set.

We use two different telephone command word databases for the testing: the first database, which we called DB1, is constructed by 2,159 IV words and 3,000 OOV words, with an IV grammar of 1,000 different words. DB1 is recorded artificially, and its quality is relatively good. The other testing set, the so-called DB2 database, has 3,467 IV words and 2,372 OOV words and all the data come from the practical environment, with worse speech quality. The size

of DB2's IV grammar is 600. All experiments are performed with HTK package.

To evaluate the performance of CM score, we usually use the DET Curve to show the false-rejection-rate (FRR) relating to the corresponding false-alarm-rate (FAR). And when the FRR equals the FAR we also give the equal-error-rate (EER) performance. All the parameters we referred can be defined as follows:

$$\text{FAR} = \frac{\text{\#false acceptances of IV} + \text{\#false acceptances of OOV}}{\text{\#mis-recognized IV words} + \text{\#OOV words}} \quad (9)$$

$$\text{FRR} = \frac{\text{\#false rejections of IV}}{\text{\#correctly recognized IV words}} \quad (10)$$

$$\text{EER} = \text{FAR} = \text{FRR} \quad \text{when FAR equals FRR} \quad (11)$$

The definition can also be clarified as: Given a rejection threshold and the calculated CM score, the recognized output can be accepted as correct or rejected as wrong. If a wrongly recognized word is accepted since the CM score exceeds the threshold, it is a false acceptance. In the other way, if a correctly recognized word is rejected because the CM score is lower than the rejection threshold, a false rejection occurs.

5.2. Experimental Results

The recognition results of the two databases are listed in Table 1. In this experiment we use the corresponding IV grammar to get the results of Word Error Rate of IV (WER_{IV}). The proportion of IV utterances in both databases are also showed in the same table.

Table 1. Recognition Results of DB1 and DB2

database	size of grammar	$\text{WER}_{\text{IV}}(\%)$	$\text{P}_{\text{IV}}(\%)$
DB1	1000	7.78	41.8
DB2	600	13.67	59.4

From Table. 1 we can see although DB2 has a smaller command word grammar, which generally means less competition among the IV grammar, it achieves a worse recognition performance than DB1. The reason is that data of DB2 come from more practical and complex environment, which leads to worse quality of speech.

Then we will compare the performances of several CMs with different strategies, by showing the corresponding EER value of each configuration, both on DB1 and DB2.

First, we setup the baseline by using the CM score calculated only based on the *target region*, employing the *filler*-based CM method referred above. Then we exploit the information included in the *non-target region* and use only these information for confidence judgement of the whole sentences, i.e., no *target region* information is used here. The performance comparison of baseline and the *non-target region* only score with initial recognized boundary (heuristic assumption) and the BIC theory are listed in Table. 2.

Table 2. Comparison of Equal Error Rate (EER) Performance on databases DB1 and DB2, with CM Score of *Non-target Region* Only

EER (%)	DB1	DB2
baseline	16.84	27.84
non-BIC	35.22	36.52
BIC	25.75	31.77

Fig. 2. DET Curves of Different systems on Databases DB1 and DB2

The above table shows that although the performance of only *non-target region* based CM score is worse than that of only *target region* based, it can achieve significant improvement when introducing the BIC theory to locate the speech part absorbed into the *non-target region*.

Furthermore we build systems to combine CM scores of *target region* and *non-target region*. For the so-called system1, we exploit the extra information by calculating CM score of the *non-target region*, using the initial whole boundary of recognized silence part. Then we employ the BIC theory for the *non-target region* to build another system2. We list the EER performances of above three systems' CM in Table. 3 on both testing sets.

Table 3. Comparison of Equal Error Rate (EER) Performance on databases DB1 and DB2

system/ database	DB1		DB2	
	EER (%)	Rel. Red.	EER (%)	Rel. Red.
baseline	16.84	—	27.84	—
system1	15.80	6.18%	25.15	9.66%
system2	13.37	20.61%	21.83	21.59%

In Table. 3 we observe that when comparing with the baseline, roughly exploiting extra information of *non-target region* for combination can achieve 6.18% and 9.66% relative reduction of EER on DB1 and DB2, respectively. Furthermore, when introducing BIC to locate the position of speech absorbed into the silence part, which means exploiting more precise information in *non-target region*, significant performance is obtained, with 20.61% and 21.59% relative reduction on DB1 and DB2, respectively.

In Fig. 2, we also show the DET curves of the baseline system and the other two combination systems which exploit extra information of the *non-target region*, both on DB1 and DB2. It also shows consistent improvement from the DET curves.

6. CONCLUSION

In this paper some famous confidence measures (CM) in Mandarin command word recognition are investigated, both in the so-called

target region and *non-target region*, respectively. It shows that exploiting extra information in the *non-target region* can effectively complement the traditional CM which usually focus on the *target region*. Furthermore, when analyzing the *non-target region* in a more theoretical way, where Bayesian information criterion (BIC) is employed to locate more precise boundary in the *non-target region*, even more improvement is achieved. In two different Mandarin telephone command word tasks, more than 20% relative reduction of equal error rate (EER) is obtained. Exploiting extra information in *non-target region* can complement not only the simple *filler*-based CM in this study, but also for other more complex CM methods for only *target region*.

7. REFERENCES

- [1] H. Jiang, "Confidence measures for speech recognition: a survey", *Speech Communication*, 45(4): 455-470, 2005.
- [2] F. Wessel, R. Schluter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", *IEEE Trans. Speech and Audio Proc.*, 9(3): 288-298, 2001.
- [3] P. Liu, Y. Tian, J. L. Zhou and F. K. Soong, "Background model based posterior probability for measuring confidence", in *Proc. Eurospeech 2005*, pp. 1465-1468, 2005.
- [4] G. Guo, C. Huang, H. Jiang, R. H. Wang, "A comparative study on various confidence measures in large vocabulary speech recognition", in *Proc. ISCSLP 2004*.
- [5] Young, S.R., "Detecting misrecognitions and out-of-vocabulary words." *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. II-21CII-24, 1994
- [6] F. K. Soong, W. K. Lo, and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words", in *Proc. SWIM-2004*.
- [7] Guo, Xue-feng, Zhu, Wei-bin, Shi, Qiu, "The IBM LVCSR system used for 1998 Mandarin broadcast news transcription evaluation", in *Proceedings of the 1999 DARPA Broadcast News Workshop*, 1999.