Score Distribution Based Term Specific Thresholding for Spoken Term Detection

Doğan Can and **Murat Saraçlar**
Electrical & Electronics Engineering Department
Boğaziçi University
İstanbul, Turkey
{dogan.can, murat.saraclar}@boun.edu.tr

Abstract

The spoken term detection (STD) task aims to return relevant segments from a spoken archive that contain the query terms. This paper focuses on the decision stage of an STD system. We propose a term specific thresholding (TST) method that uses per query posterior score distributions. The STD system described in this paper indexes word-level lattices produced by an LVCSR system using Weighted Finite State Transducers (WFSTs). The target application is a sign dictionary where precision is more important than recall. Experiments compare the performance of different thresholding techniques. The proposed approach increases the maximum precision attainable by the system.

1 Introduction

The availability of vast multimedia archives calls for solutions to efficiently search this data. Multimedia content also enables interesting applications which utilize multiple modalities, such as speech and video. Spoken term detection (STD) is a subfield of speech retrieval, which locates occurrences of a query in a spoken archive. In this work, STD is used as a tool to segment and retrieve the signs in news videos for the hearing impaired based on speech information. After the location of the query is extracted with STD, the sign video corresponding to that time interval is displayed to the user. In addition to being used as a sign language dictionary this approach can also be used to automatically create annotated sign databases that can be utilized for training sign recognizers (Aran et al., 2008). For these applications the precision of the system is more important than its recall.

The classical STD approach consists of converting the speech to word transcripts using large vocabulary continuous speech recognition (LVCSR) tools and extending classical information retrieval techniques to word transcripts. However, retrieval performance is highly dependent on the recognition errors. In this context, lattice indexing provides a means of reducing the effect of recognition errors by incorporating alternative transcriptions in a probabilistic framework. A system using lattices can also return the posterior probability of a query as a detection score. Various operating points can be obtained by comparing the detection scores to a threshold. In addition to using a global detection threshold, choosing term specific thresholds that optimize the STD evaluation metric known as Term-Weighted Value (TWV) was recently proposed (Miller et al., 2007). A similar approach which trains a neural network mapping various features to the target classes was used in (Vergyri et al., 2007).

The rest of the paper is organized as follows. In Section 2 we explain the methods used for spoken term detection. These include the indexing and search framework based on WFSTs and the detection framework based on posterior score distributions. In Section 3 we describe our experimental setup and present the results. Finally, in Section 4 we summarize our contributions and discuss possible future directions.

Proceedings of NAACL HLT 2009: Short Papers, pages 269–272,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

2 Methods

The STD system used in this study consists of four stages. In the first stage, an LVCSR system is used to generate lattices from speech. In the second stage the lattices are indexed for efficient retrieval. When a query is presented to the system a set of candidates ranked by posterior probabilities are obtained from the index. In the final stage, the posterior probabilities are compared to a threshold to decide which candidates should be returned.

2.1 Indexing and Retrieval using Finite-State Automata

General indexation of weighted automata (Allauzen et al., 2004) provides an efficient means of indexing for STD (Parlak and Saraçlar, 2008; Can et al., 2009), where retrieval is based on the posterior probability of a term in a given time interval. In this work, the weighted automata to be indexed are the preprocessed lattice outputs of the ASR system. The input labels are phones, the output labels are quantized time-intervals and the weights are normalized negative log probabilities. The index is represented as a WFST where each substring (factor) leads to a successful path over the input labels whenever that particular substring was observed. Output labels of these paths carry the time interval information followed by the utterance IDs. The path weights give the probability of each factor occurring in the specific time interval of that utterance. The index is optimized by WFST determinization and minimization so that the search complexity is linear in the sum of the query length and the number of times the query appears in the index.

2.2 Decision Mechanism

Once a list of candidates ranked with respect to their posterior probabilities are determined using the index, the candidates exceeding a threshold are returned by the system. The threshold is computed to minimize the Bayes risk. In this framework, we need to specify a cost function, prior probabilities and likelihood functions for each class. We choose the cost of a miss to be 1 and the cost of a false alarm to be a free parameter, α. The prior probabilities and the likelihood functions are estimated from the posterior scores of the candidate results for each query.

The likelihood functions are found by fitting parametric models to the score distributions (Manmatha et al., 2001). In this study, the score distributions are modeled by exponential distributions. When the system returns a score, we do not know whether it belongs to the correct or incorrect group, so we use a mixture of two exponential distributions to model the posterior scores returned by the system. The exponential mixture model (EMM) parameters are determined via unsupervised estimation using the Expectation-Maximization (EM) algorithm. Figure 1 shows the normalized histogram of posterior scores and the EM estimate given by our method for an example query.

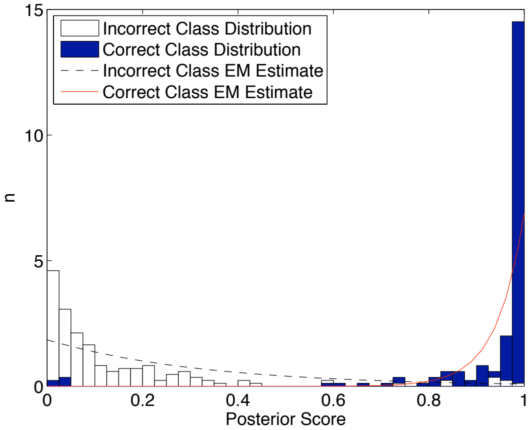

Figure 1: The normalized histogram of posterior scores and the EM estimates for correct and incorrect detections given an example query.

If we denote the posterior score of each candidate by x, incorrect class by c_0 and correct class by c_1, we have

$$p(x) = P(c_0)p(x|c_0) + P(c_1)p(x|c_1)$$

where the incorrect class likelihood $p(x|c_0) = \lambda_0 e^{-\lambda_0 x}$ and correct class likelihood $p(x|c_1) = \lambda_1 e^{-\lambda_1(1-x)}$. The model parameters $\lambda_0, \lambda_1, P(c_0), P(c_1)$ are estimated using the EM algorithm given the scores x_i for $i = 1, \ldots, N$. Each iteration consists of first computing $P(c_j|x_i) = P(c_j)p(x_i|c_j)/p(x_i)$ for $j = 1, 2$ and then updating

$$P(c_j) = \frac{1}{N}\sum_i P(c_j|x_i),$$

$$\lambda_0 = \frac{\sum_i P(c_0|x_i)}{\sum_i P(c_0|x_i)x_i},$$

$$\lambda_1 = \frac{\sum_i P(c_1|x_i)}{\sum_i P(c_1|x_i)(1-x_i)}.$$

After the mixture parameters are estimated, we assume that each mixture represents a class and mixture weights correspond to class priors. Then, the Minimum Bayes Risk (MBR) detection threshold for x is given as:

$$\frac{\lambda_1 + \log(\lambda_0/\lambda_1) + \log(P(c_0)/P(c_1)) + \log\alpha}{\lambda_0 + \lambda_1}.$$

3 Experiments

3.1 Data and Application

Turkish Radio and Television Channel 2 (TRT2) broadcasts a news program for the hearing impaired which contains speech as well as signs. We have collected 11 hours (total speech time) of test material from this broadcast and performed our experiments on this data with a total of 10229 single word queries extracted from the reference transcriptions. We used IBM Attila speech recognition toolkit (Soltau et al., 2007) at the back-end of our system to produce recognition lattices. The ASR system is trained on 100 hours of speech and transcription data collected from various TV and radio broadcasts including TRT2 hearing impaired news, and a general text corpus of size 100 million words.

Our application uses the speech modality to retrieve the signs corresponding to a text query. Retrieved results are displayed as video demonstrations to support the learning of sign language. Since the application acts like an interactive dictionary of sign language, primary concern is to return correct results no matter how few they are. Thus high precision is appreciated much more than high recall rates.

3.2 Evaluation Measures

In our experiments, we use precision and recall as the primary evaluation metrics. For a set of queries $q_k, k = 1, \ldots, Q$,

$$\text{Precision} = \frac{1}{Q}\sum_k \frac{C(q_k)}{A(q_k)} \quad \text{Recall} = \frac{1}{Q}\sum_k \frac{C(q_k)}{R(q_k)}$$

where:
$R(q_k)$: Number of occurences of query q_k,
$A(q_k)$: Total no. of retrieved documents for q_k,
$C(q_k)$: No. of correctly retrieved documents for q_k.

We obtain a precision/recall curve by changing the free parameter associated with each thresholding method to simulate different decision cost settings. Right end of these curves fall into the high precision region which is the main concern in our application.

For the case of global thresholding (GT), the same threshold θ is used for all queries. TWV based term specific thresholding (TWV-TST) (Miller et al., 2007) aims to maximize the TWV metric introduced during NIST 2006 STD Evaluations (NIST, 2006).

$$\text{TWV} = 1 - \frac{1}{Q}\sum_{k=1}^{Q}\{P_{miss}(q_k) + \beta.P_{FA}(q_k)\}$$

$$\mathrm{P_{miss}}(q_k) = 1 - \frac{C(q_k)}{R(q_k)}, \mathrm{P_{FA}}(\mathrm{q_k}) = \frac{A(q_k) - C(q_k)}{T - C(q_k)}$$

where T is the total duration of the speech archive and β is a weight assigned to false alarms that is proportional to the prior probability of occurence of a specific term and its cost-value ratio. This method sets individual thresholds for each query term considering per query expected counts and the tuning parameter β. In the proposed method α plays the same role as β and allows us to control the decision threshold for different cost settings.

3.3 Results

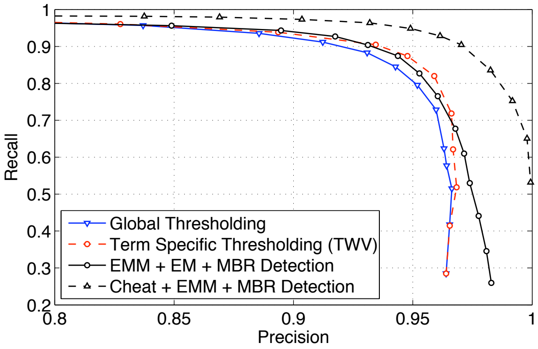

Figure 2: The precision and recall curves for various thresholding techniques.

Figure 2 compares GT, TWV-TST, and the proposed method that utilizes score distributions to derive an optimal decision threshold. For GT and TWT-TST, last precision/recall point in the figure corresponds to the limit threshold value which is 1.0. Both the TWV-TST and the proposed method outperform GT over the entire region of interest. While TWV-TST provides better performance around the knees of the curves, proposed method achieves higher maximum precision values which coincides with the primary objective of our application.

Figure 2 also provides a curve of what happens when the correct class labels are used to estimate the parameters of the exponential mixture model in a supervised manner instead of using EM. This curve provides an upper bound on the performance of the proposed method.

4 Discussion

In this paper, we proposed a TST scheme for STD which works almost as good as TWV-TST. Extrapolating from the cheating experiment, we believe that the proposed method has potential for outperforming the TWV-TST over the entire region of interest given better initial estimates for the correct and incorrect classes.

A special remark goes to the performance in the high precision region where our method clearly outperforms the rest. While GT and TWV-TST methods are bounded around 96.5% precision value, our method reaches at higher precision figures. For GT, this behavior is due to the inability to set different thresholds for different queries. For TWT-TST, in the high precision region where β is large, the threshold is very close to 1.0 value no matter what the expected count of the query term is, thus it essentially acts like a global threshold.

Our current implementation of the proposed method does not make use of training data to estimate the initial parameters for the EM algorithm. Instead, it relies on some loose assumptions about the initial parameters of the likelihood functions and uses uninformative prior distributions. The significant difference between the upper bound and the actual performance of the proposed method indicates that the current implementation can be improved by better initial estimates.

Our assumption about the parametric form of the likelihood function may not be valid at all times. Maximizing the likelihood with mismatched models degrades the performance even when initial parameters are close to the optimal values. In the future, other parametric forms can be utilized to better model the posterior score distributions.

Maximum likelihood estimation with insufficient data is prone to overtraining. This is a common situation with the STD task at hand. With the current data, three or less results are returned for half of the queries. Bayesian methods can be used to introduce priors on the model parameters in order to make the estimation more robust.

Acknowledgments

This study was supported in part by Boğaziçi University Research Fund (BAP) under the project number 05HA202, TÜBİTAK under the project number 105E102 and Turkish State Planning Organization (DPT) under the project number DPT2007K120610.

References

C. Allauzen, M. Mohri, and M. Saraçlar. 2004. General-indexation of weighted automata-application to spoken utterance retrieval. In *Proc. Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL*, pages 33–40, March.

O. Aran, I. Arı, E. Dikici, S. Parlak, P. Campr, M. Hruz, L. Akarun, and M. Saraçlar. 2008. Speech and sliding text aided sign retrieval from hearing impaired sign news videos. *Journal on Multimodal User Interfaces*, 2(1):117–131, November.

D. Can, E. Cooper, A. Sethy, C.M. White, B. Ramabhadran, and M. Saraclar. 2009. Effect of pronunciations on oov queries in spoken term detection. In *ICASSP*, April.

R. Manmatha, T. Rath, and F. Feng. 2001. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01*, pages 267–275, New York, NY, USA. ACM.

D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish. 2007. Rapid and accurate spoken term detection. In *Proc. Interspeech*, pages 314–317, August.

NIST. 2006. The spoken term detection (STD) 2006 evaluation plan http://www.nist.gov/speech/tests/std/.

S. Parlak and M. Saraçlar. 2008. Spoken term detection for Turkish broadcast news. In *Proc. ICASSP*, pages 5244–5247, April.

H. Soltau, G. Saon, D. Povey, L. Mangu, J. Kuo, M. Omar, and G. Zweig. 2007. The IBM 2006 GALE Arabic ASR system. In *Proc. ICASSP 2007*, Honolulu, HI, USA.

D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang. 2007. The SRI/OGI 2006 spoken term detection system. In *Proc. Interspeech*, pages 2393–2396, August.