# Phonetic spoken term search using topic information

*Shahram Kalantari, David Dean, Sridha Sridharan*

Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia.

## Abstract

The aim of spoken term detection (STD) is to find all occurrences of a specified query term in a large audio database. This process is usually divided into two steps: indexing and search. In a previous study, it was shown that knowing the topic of an audio document would help to improve the accuracy of indexing step which results in a better performance for STD system. In this paper, we propose the use of topic information not only in the indexing step, but also in the search step. Results of our experiments show that topic information could also be used in search step to improve the STD accuracy.

**Index Terms**: spoken term detection, keyword spotting, indexing, search

## 1. Introduction

STD is the process of finding all occurrences of a specified search term in a large volume of speech database. In the first step, all audio documents in the database are indexed into a compact and searchable representation (indexing stage). The second stage is responsible for exploring this intermediate representation to detect the query terms (search stage). Indexing is performed once, as an off-line process, while many searches are later performed within this index. It has been shown that using topic information in the form of topic dependent language models (LM) improves the accuracy of indexing and consequently improves the STD accuracy [1]. In this paper, we propose a novel method to exploit the topic information of the indexed document not only in the indexing stage, but also in the search stage using the $a$-priori probability of the search term with respect to its topic dependent LM (TDLM). This will answer the question if using topic information of the indexed document in the search stage has the same positive effect on STD accuracy as it has in the indexing stage.

Indexing audio documents could be performed using large vocabulary continuous speech recognition (LVCSR) systems [2]. By so doing, a word lattice is produced for each audio segment [3] which could be traced to detect the query term in search stage. LVCSR systems are able to accurately recognize words within their vocabulary. however, STD systems based on LVCSR are not able to detect out-of-vocabulary query terms. In order to solve this problem, sub-word based indexing have been investigated to provide open vocabulary query search. However, recognition accuracy in these systems is not as good as LVCSR systems.

The dynamic match lattice spotting (DMLS) technique [4] has been proposed as a phonetic STD approach to search and detect query terms in recognized lattices of audio segments which are created using a phone recognition engine based on hidden Markov models (HMM). This technique has continued to be used as a state-of-the-art approach for STD up to the present day [5, 6]. DMLS was further improved by Wallace et al. [7] and became faster by putting phone sequences of the recognized lattices into a phonetic sequence database (SDB) as an off-line process and then the sequence of phonemes in the search term (target sequence) can be searched through the phonetic SDB. This technique is explained in Section 2 and will be used as our baseline framework.

Language modelling is one approach that has been shown to help indexing accuracy in phonetic STD systems [5]. Topic dependent LMs are also used to make use of topic information of the decoding documents in indexing stage [1]. A set of topic dependent LMs were trained and used in addition to acoustic models to index the audio documents in the database. The search stage however did not make use of the topic information in this approach.

In order to find a search term in an audio database, a set of putative occurrences as well as their scores are output by the STD system. By thresholding these scores, best ones are selected to output as hits, and the rest of them are discarded. In this paper, we will make use of topic information in the search stage. Suppose that we know the topic of a document is “sports” and we are searching for term “goal”. Then if the score of a putative occurrence is below the threshold value, it could be increased by adding a value which is in direct relationship with the LM probability of that term in “sports” topic. This could also be done to decrease the score of rare terms regarding topic dependent LM probability of the search term in case of searching for a word that is less likely to happen in the indexed document. We believe that this approach will compensate some of the errors that are introduced by the recognition engine during indexing stage.

The remainder of this paper is organized as follows: Section 2 presents the DMLS system and its components. In Section 3, our method of using topic information for STD in search step is proposed. An overview of the datasets and experimental set-up is presented in Section 4 followed by the results and discussions on the experiments in Section 5. Finally a conclusion of our method is proposed in Section 6.

## 2. DMLS system

The phonetic STD system developed by Wallace et al. [7] which is based on the DMLS system [4] is used as our baseline system. In this system, indexing is run once to create a database from recognized lattices of phonemes and in the search phase, this database is explored to find the best match with query term.

### 2.1. Indexing

The purpose of indexing is to construct a database that provides fast search. First, phonetic speech recognition is performed to decode each speech segment in the database which results in producing lattices of multiple phone sequence recognition hypotheses. In the next step, these lattices are traversed by means of Viterbi dynamic search method to extract all phone sequences with a predefined fix length, *N*, that terminate at each

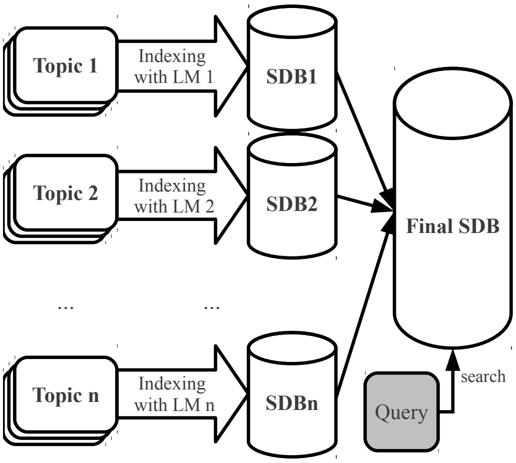

Figure 1: *Topic dependent indexing for STD*

node in the lattice. All these phone sequences are then collected into a SDB. For indexing each audio segment, its topic dependent LM is used. This database can be considered as a look-up table that returns the location of each occurrence of a particular unique *N*-gram phone sequence. In this paper, we used the value of *N* = 11, which provides a reasonable trade-off between index size and simple retrieval of long phone sequences [8].

### 2.2. Search

In search stage, the query term is decomposed into its phoneme constituents using a pronunciation lexicon. Letter to sound rules are applied in case of out-of-vocabulary search terms. The difference between the target phone sequence and each indexed phone sequence is calculated using the minimum edit distance (MED) criteria. If the difference is lower than a pre-specified threshold value, then the putative occurrence is emitted as a detected occurrence.

The MED is defined as the minimum possible sum of the costs of phone substitution, insertion and deletion errors that transform the indexed phone sequence into the target phone sequence, and is calculated using dynamic programming [4].

## 3. Use of topic information in search stage

Topic information of the indexed document could be used to refine the score of putative occurrences and make a better decision for outputing them as hits. Assuming that the topic of each audio document is known before indexing, a set of topic dependent LMs are used to index each document and phone sequences of each indexed audio segment are put in the sequence database (SDB).

In the search stage, the usual MED criteria is used for computing the distance between the target phone sequence (the sequence of the phones in the query term) and each phone sequence in SDB. In this work, a new method is introduced to refine the MED values based on the topic of the phone sequence in the SDB. The log-likelihood of the search term belonging to the topic of the document it was indexed with, is used to produce a new distance value between the search term and the phone sequence in the database.

### 3.1. MED refinement based on topic information

The MED value could be refined based on the topic dependent LM probability of the search term. In order to do that, first we compute the log-likelihood of the search term using its 1-gram topic dependent LM probability. We also compute the log-likelihood of the search term using its 1-gram general LM probability. If we subtract the second probability value from the first one, it results in the log-likelihood ratio ($LLRH$) of the search term regarding topic dependent LM and general LM.

$$LLHR(w) = LogP_{TDLM}(w) - LogP_{GLM}(w) \qquad (1)$$

$LLHR$ can be interpreted as commonness of a word in a particular topic compared with general text. A higher value for $LLHR$ indicates that the word happens more in that particular topic compared with general text. After computing the $LLHR$ of all terms, their normal distribution is computed and In the next step, all these values are brought to [-0.5 0.5] range using the following formula:

$$LLHR_n(w) = \begin{cases} -0.5, & \text{if } (LLHR - \mu) < -3\sigma \\ \frac{LLHR-\mu}{6\sigma^2}, & \text{if } -3\sigma < (LLHR - \mu) < 3\sigma \\ 0.5, & \text{if } (LLHR - \mu) > 3\sigma \end{cases} \qquad (2)$$

where $\mu$ is the mean of $LLHR$ values, $\sigma$ is the standard deviation, and $LLHR_n$ is the normalized $LLHR$ which is a value between $-0.5$ and $0.5$. Now the refined distance value could be defined as:

$$NewDistance = MED * (1 - LLHR_n)^C. \qquad (3)$$

Where $C$ is a constant value indicating the weight of the second term and $1 - LLHR_n$ is a value between $0.5$ and $1.5$. Therefore, if the $LLHR$ value of the search term is high, then the distance between the search term and the phone sequence in the SDB will be decreased. This could potentially increase the number of hits. On the other hand, if the $LLHR$ value of the search term is low, then the distance between the search term and the phone sequence in the SDB will be increased and by doing so, the number of false alarms are likely to be decreased.

## 4. Dataset and experimental set-up

### 4.1. LM training

In order to train an $n$-gram LM for indexing, $n$-gram probabilities need to be estimated from a set of training transcriptions. The English text database of the second phase of topic detection and tracking (TDT2) project is used for training a GLM as well as TDLMs. The TDT2 English audio and text database is a collection of broadcast resources in the form of audio recordings and corresponding transcriptions and also new-wire data, which is generally used for the purpose of topic detection and tracking [9]. Broadcast data are useful for the purpose of topic dependent STD, as they are audio files with corresponding word-level transcription and also the topic information for each individual document. Each document in this corpus is tagged with one of 96 topics.

For the purpose of topic dependent STD task, there were some further annotations done on the TDT2 database. First, documents which belong to the broadcast data were selected for this study. TDT2 defines topics as a specific event or activity, along with all directly related events and activities. This definition makes topics to be quite specific. For example, “the

| Topic Id | Topic name | Hours |
|---|---|---|
| 1- Vio | Ongoing violence | 15.00 |
| 2- Sca | Scandals | 12.71 |
| 3- Fin | Finance | 7.18 |
| 4- Leg | Legal cases | 5.81 |
| 5- Ele | Elections | 5.02 |
| 6- Sci | Science news | 2.74 |
| 7- Spo | Sports | 2.73 |
| 8- Acc | Accidents | 1.45 |
| 9- Dis | Natural disasters | 0.86 |
| 10- Law | New laws | 0.23 |
| 11- Mis | Misc. news | 3.56 |

Table 1: *The modified TDT2 database and the hours of speech contained in each topic*

financial crisis in China and its effects on Asian countries" is considered as an entire topic. However, such events could be generalized into broader topics. In this case, we can categorize this event as a financial topic. Among TDT2 data, the number of documents belonging to each individual topic (based on TDT2 definition) was limited which causes the TDLMs to be under-trained. Therefore, as shown in Table 1, in the second step 96 topics were categorized into 11 broader topics. This procedure was done manually and all of the documents were manually tagged with a cluster id based on TDT2 annotation guidelines. This resulted in a final database with 11 different new topics, organized into 11 clusters and each cluster has a set of audio and transcription files. For the rest of this paper, the word "topic" refers to this set of 11 broad topics.

Each topic was randomly divided into two parts: 70% of the data is used as STD development data to train phone errors and consequently insertion, deletion, and substitution costs. In the next step, a word LM was created from development data based on the transcription files for each topic. The well-known SRILM toolkit [10] with default Good-Turing discounting and Katz back-off for smoothing was used to create 1, 2, 3, and 4-gram word LMs. Evaluation is performed on the remaining 30%. A total of 1200 search terms are chosen randomly from a pool of words that occur at least once in the evaluation data in each topic, with 400 words selected for each of the lengths of 4 phones, 6 phones, or 8 phones.

It is worth mentioning that for the best performance, it is necessary to find the best topics which represent the dataset more accurately and divide documents into more suitable topics.

### 4.2. AM training

For acoustic modelling, a monophone HMM is trained for each phoneme class, which is a 32 mixture mono-phone HMM, with 3 emitting states. These HMMs are trained using TIMIT, WSJ1, and 160 hours of speech from Switchboard-1 Release 2 (SWB).

There are a number of parameters that must first be tuned on tuning data. For each LM, we tune the parameters of the decoder on a small set of held-out training data from the TDT2 corpus, and select the parameters that provide for the best phone recognition accuracy to achieve the best performance of indexing.

For each LM type, token insertion penalty and grammar scale factor are optimized for 1-best phone recognition accuracy, and an $n$-gram order of up to 4-gram is considered. This was done by first decoding initial lattices with up to a 2-gram LM and then applying up to a 4-gram LM during lattice rescoring with the HTK tool, HLRescore. Phonetic indexing is performed by decoding a lattice of words, then expanding these tokens into their corresponding sequences of phones using a pronunciation lexicon, whilst maintaining lattice structure. While higher order $n$-gram LMs are possible, this is not considered here to avoid training data sparsity problems [11]. The results of tuning found that the best phone recognition accuracy was achieved by decoding with a full vocabulary and with 4-gram LMs. Therefore, this configuration is used in all experiments in the following sections.

### 4.3. Evaluation

STD performance, is evaluated in this paper using figure of merit (FOM). FOM is a widely used metric to evaluate the performance of STD systems [7, 12, 13, 14] and is defined as the average detection rate at each integer value between 0 and 10 false alarms per search term per hour.

## 5. Experiments and results

In order to investigate the effect of using topic information in the search stage, we applied MED refinement on the test set for each topic and compared the result of STD accuracy with that of a usual STD system which works based on MED criteria. Figure 2 shows the results of these two systems.

In this figure as it was shown in our previous work [1], we can see that using topic information for indexing a document will improve the STD accuracy compared with using general LM which does not care about the topic of the documents. It is also can be seen that in all topics, using topic information in search stage based on our approach improves the accuracy of STD system. In fact, STD accuracy is improved by relative 7% in average when refining MED between the target phone sequence and the search term comparing with the average accuracy of topic dependent indexing for STD. For some topics, topic information based search provides more improvement than using topic information for search in other topics. For example, the improvement we get by using topic information for search in "Scandals" and "Science" topics is much more than that in "Ongoing violence" and "Elections" topics. This is probably because of better representation of documents with topic dependent LMs in those topics. Also, it could happen due to the poor labelling of some the training data according to their topics which again makes them less representative of their documents. However, results in all cases show that topic information not only helps better indexing which helps the overall STD performance, but also it could be used in the search stage to refine the distance between candidate phone sequences in the database and the search term.

## 6. Conclusions and future work

In this paper, usefulness of topic information for STD in search stage was investigated. Topic dependent LM probability of the search term in the indexed document was used to refine the MED between the candidate phone sequences in the database and the query term. Experiments show that this method improves the accuracy of STD system compared with the case when usual MED criteria were used. This shows that our method increases the number of hits by accepting more term which are assumed to be a good match with the target phone

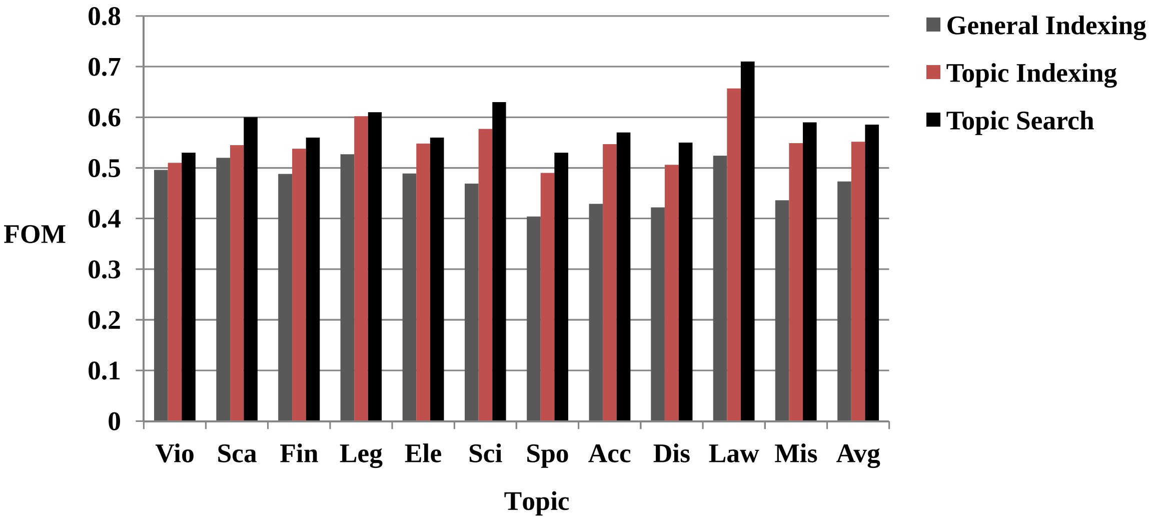

Figure 2: STD accuracy (FOM) using general LM based indexing, TDLM based indexing, and TDLM based indexing-search

sequence. Also, the distance between the search term and the phone sequence in the SDB will be increased in cases where the search term is not likely to happen in the indexed document and by doing so, the number of false alarms are decreased and these two factors improve the STD accuracy.

As a future work, the topic of each document could be inferred from the audio information to determine its front-end effect on the final STD system. In this paper our set of search terms is unbiased (as they are chosen randomly from the transcripts) and is provided online for reproduciblity. The mapping between TDT2 topics and the broader topics used in this paper as well as the development and evaluation data and train/test divisions and also the search term list for each topic will be provided at https://wiki.qut.edu.au/display/saivt/ at the time of publication of this paper. The TDT2 corpus is also available through linguistic data consortium (LDC). An investigation of the effects of the approach on other kinds of search terms, including out-of-vocabulary terms and rare terms, is an important matter for future work.

## 7. Acknowledgements

This work has been supported by the Australian Cooperative Research Centre for Smart Services.

## 8. References

[1] S. Kalantari, D. Dean, and S. Sridharan, “Topic dependent language modelling for spoken term detection,” in *Europian signal processing conference*, 2014.

[2] S. Young, “A review of large-vocabulary continuous-speech recognition,” *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, 1996.

[3] D. Miller, M. Kleber, C. lin Kao, O. Kimball, T. Colthurst, S. Lowe, R. Schwartz, and H. Gish, “Rapid and accurate spoken term detection,” in *Interspeech*, 2007, pp. 314–317.

[4] K. Thambiratnam and S. Sridharan, “Rapid yet accurate speech indexing using Dynamic Match Lattice Spotting,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 346–357, 2007.

[5] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, “The effect of language models on phonetic decoding for spoken term detection,” in *ACM Multimedia Workshop on Searching Spontaneous Conversational Speech*, 2009, pp. 31–36.

[6] M. Rajabzadeh, S. Tabibian, A. Akbari, and B. Nasersharif, “Improved dynamic match phone lattice search using viterbi scores and Jaro Winkler distance for keyword spotting system,” in *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, 2012, pp. 423–427.

[7] R. Wallace, R. Vogt, and S. Sridharan, “Spoken term detection using fast phonetic decoding,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4881–4884.

[8] ——, “A phonetic search approach to the 2006 NIST Spoken Term Detection evaluation,” in *Interspeech*, 2007, pp. 2385–2388.

[9] National Institute of Standards and Technology, “Topic detection and tracking (TDT),” July 2003. [Online]. Available: http://nist.gov/speech/tests/tdt/index.htm

[10] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *The 7th International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[11] R. Rosenfeld, “Two decades of statistical language modeling: where do we go from here?” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[12] R. Wallace, R. Vogt, B. Baker, and S. Sridharan, “Optimising Figure of Merit for phonetic spoken term detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5298–5301.

[13] A. J. K. Thambiratnam and S. Sridharan, “Dynamic match lattice spotting for indexing speech content,” U.S. Patent 11/377 327, August 2, 2007.

[14] D. Zhu, H. Li, B. Ma, and C.-H. Lee, “Discriminative learning for optimizing detection performance in spoken language recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4161–4164.