TOPIC DEPENDENT LANGUAGE MODELLING FOR SPOKEN TERM DETECTION

Shahram Kalantari[1], David Dean[1], Sridha Sridharan[1], Roy Wallace[1]

[1]Speech Laboratory, Queensland University of Technology, Brisbane, Queensland, Australia
s1.kalantari@qut.edu.au, ddean@ieee.org, sridharan@qut.edu.au, royw@ieee.org

ABSTRACT

This paper investigates the effect of topic dependent language models (TDLM) on phonetic spoken term detection (STD) using dynamic match lattice spotting (DMLS). Phonetic STD consists of two steps: indexing and search. The accuracy of indexing audio segments into phone sequences using phone recognition methods directly affects the accuracy of the final STD system. If the topic of a document in known, recognizing the spoken words and indexing them to an intermediate representation is an easier task and consequently, detecting a search word in it will be more accurate and robust. In this paper, we propose the use of TDLMs in the indexing stage to improve the accuracy of STD in situations where the topic of the audio document is known in advance. It is shown that using TDLMs instead of the traditional general language model (GLM) improves STD performance according to figure of merit (FOM) criteria.

***Index Terms*—** spoken term detection, language modelling, indexing

1. INTRODUCTION

STD is the process of finding all occurrences of a specified search term in a large volume of speech database. This process usually consists of two steps: indexing and search. In the indexing stage, audio segments are transcribed into an intermediate representation and in the next stage, this representation is searched to detect the query terms. Indexing is performed once, as an off-line process, while many searches are later performed within this index. Better indexing accuracy results in a more accurate STD system. However, the indexing stage is prone to errors due to errors introduced in recognition engine (words, sub-word or phonemes) used in the indexing process.

A common approach to indexing audio segments is to perform word-based transcription by using a large vocabulary continuous speech recognition (LVCSR) system [1] to produce word lattices for each audio segment. This has been shown to be an effective approach for in-vocabulary query terms, but is not applicable for out-of-vocabulary queries, as LVCSR systems are only able to recognize the words within their dictionary. Sub-word based strategies have been investigated to provide open vocabulary query search [2]. The dynamic match lattice spotting (DMLS) technique [3] has been proposed as a phonetic STD approach to search and detect query terms in recognized lattices of audio segments, 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 [4, 5].

DMLS was further improved by Wallace et al. [6] and became faster by putting phone sequences in the 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 system is used as our baseline framework.

One approach that has been successfully applied in phonetic STD, is using language models (LMs) in the indexing stage [4]. While word level information in the form of statistical word based n-gram LMs is helpful, human speech perception system incorporates other information, one of which is the topic of the conversation. When a person knows the topic of the conversation, they may have a better chance to recognize the spoken words. As an example, pronunciation of the words “white house” and “light house” only differs in one phone and they are likely to be confused with each other. However, if the topic of the conversation is politics, then the spoken term is much more likely to be “white house” rather than “light house”. In this paper, we propose topic dependent language modelling to make use of topic information of the speech segments to improve the accuracy of indexing stage and consequently improve the accuracy of phonetic STD.

It has been shown that TDLMs outperform the simple word-based n-gram LMs in terms of perplexity and ASR accuracy [7, 8]. Some of these systems use simple method of building different topic-dependent n-gram LMs and interpolate them to create a single LM [9]. Other systems tend to use more advanced semantic analysis techniques to extract and detect the topic of the document [10]. After that, LMs are created for each topic and are used as prior knowledge to help the acoustic model (AM) recognize the speech.

Presently, there has been minimal investigation in improving STD performance by using topic information. Some examples have been tried using context information using a window of neighbour words for name spotting [11, 12]. These approaches outperform the baseline system in terms of STD accuracy. However, creating context models in the search time is a time-consuming task which is prohibitive for audio docu-

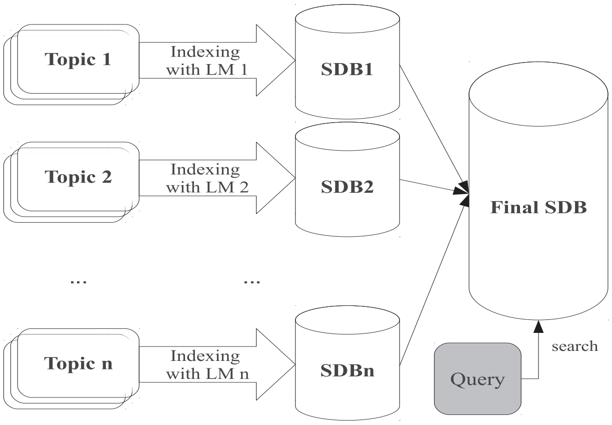

Fig. 1. *Topic dependent indexing for STD*

ment retrieval applications.

2. DMLS SYSTEM

The phonetic STD system developed by Wallace et al. [6] which is based on the DMLS system [3] 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. These lattices are then traversed by Viterbi dynamic search method to extract all phone sequences with a predefined fix length, *N*, that terminate at each node in the lattice. All these phone sequences are then collected into a SDB. 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 [13].

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.

3. TOPIC DEPENDENT SPOKEN TERM DETECTION

Language modelling is commonly used to improve the accuracy of recognizing the sequence of word or sub-word units in the input speech segment [4]. The most popular method

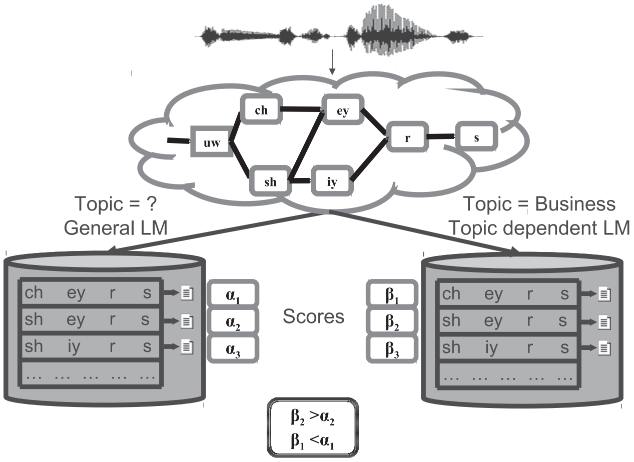

Fig. 2. *An example of TDLM-based indexing for topic dependent STD*

for language modelling is to use n-grams. In this method, it is assumed that $a\ priori$ probability of an observation unit is dependent only on a short history of its $n-1$ proceeding events. This assumption makes the model very simple and it has been shown to be a simple yet effective method for improving speech recognition accuracy [14].

TDLMs have shown to outperform the simple word-based n-gram LMs in terms of perplexity and ASR accuracy [7, 8]. In these systems, each audio segment is considered to belong to a particular topic such as sports, politics, etc. In the simplest approach, a TDLM can be created from documents of each topic. Each speech segment is then decoded using its TDLM.

In this work, assuming that the topic of the speech segments are known in advance, speech segments are decoded using TDLMs and the resulting indexes are put in the final SDB. This approach is shown in Figure 1. We also compare our approach with a system based on a GLM which is trained with all of training data without considering their topic. Our method is useful for situations like political talk shows, sports or finance news, etc., where the topic of the audio document is specified before. It is also applicable if at first a topic detector based on audio information is used to detect the topic of the audio document and then apply topic dependent STD. However, this will impose a front-end effect of topic detection accuracy on the final STD accuracy.

When TDLMs are used for indexing, phone recognition accuracy is increased and the score of each recognized phone sequence will be more accurate. Therefore, a better phone SDB will be provided which consists of recognized phone sequences with more accurate values as their scores which indicate how likely is that the recognition was done correctly. An example of this process is depicted in Figure 2. In this example, the spoken word "share" is passed to the recognition engine. If TDLM is used, it is more likely to be recognized as "shares" rather than "chairs", thus the score of "shares" is higher compared with GLM-based indexing. This will reduce the number of false alarms, while at the same time, increases the number of hits which will provide a more accurate spoken term detection system.

4. DATASET AND EXPERIMENTAL SET-UP

4.1. LM training

As mentioned previously, in order to train an n-gram LM, n-gram probabilities need to be estimated from a set of training transcriptions. For this paper, 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 [15]. 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 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 such as "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 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 in this research we are not seeking to achieve the best STD performance and we just want to investigate if topic information could help to improve STD performance. 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.

Topic Id	Topic name	Hours
1	Ongoing violence	15.00
2	Scandals	12.71
3	Finance	7.18
4	Legal cases	5.81
5	Elections	5.02
6	Science news	2.74
7	Sports	2.73
8	Accidents	1.45
9	Natural disasters	0.86
10	New laws	0.23
11	Misc. news	3.56

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

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 one hour 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. Although these parameters need to be tuned to achieve the best STD performance, in practice, usually indexing stage is tuned to provide best recognition accuracy. Moreover, the aim of this paper is to investigate if indexing improvement using TDLMs provides better STD accuracy.

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 [14]. 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.

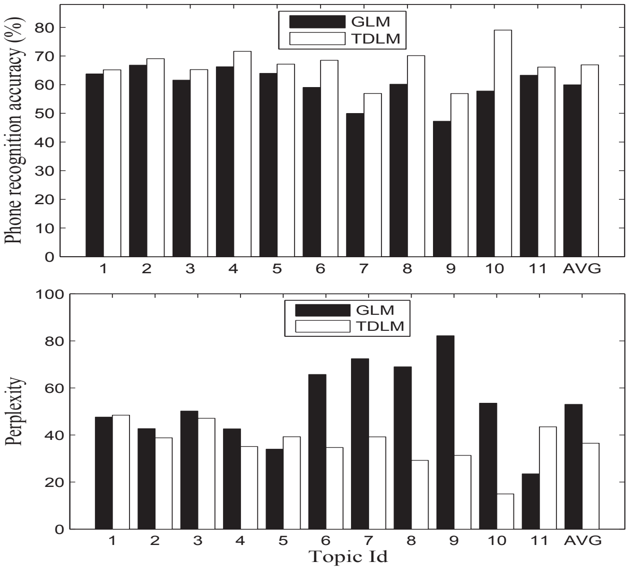

Fig. 3. *Phone recognition accuracy using TDLM versus GLM and the perplexity of TDLMs and the GLM calculated on evaluation set of each topic.*

4.3. Evaluation

STD can be evaluated in two stages. The first stage can test the performance of phonetic lattices through perplexity and phone recognition accuracy. Perplexity is a common way of evaluating LMs and is defined as a probability distribution over entire sentences. Better LMs tend to have lower perplexity which means that they are less surprised by the test samples. Phone recognition accuracy is calculated according to HTK style. The second stage, which actually tests the whole STD performance, is evaluated in this paper using figure of merit (FOM). FOM is used widely to report the performance of STD systems [6,16–18] 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

5.1. Effect of topic dependent LMs on indexing accuracy

In order to investigate the effect of known topics on the STD indexing, the utility of TDLMs and the GLM on phone recognition is investigated in Figure 3. In this paper, we used the SRILM toolkit [19] to calculate the perplexity of TDLMs and the GLM on the evaluation set of topics, which is also presented in Figure 3.

As it can be seen in Figure 3, the phone recognition accuracy of TDLM is increased in all cases. Particularly, if we take average of the accuracy over all of test data shown in the final column, the phone recognition accuracy of the system is increased relatively by more than 10% using TDLMs rather than GLM. This shows, as expected, TDLMs do improve phone recognition. If we have a closer look at this figure, we can see that the perplexity of TDLMs for topics 6 to 10 with less than 3 hours of speech is much lower which means that these LMs are more representative of the data contained in their topic. For example, the perplexity of GLM for the topic 10 with only 0.23 hours of data, is around 54, while this value for TDLM is around 15 which shows about 70% improvement. It is also observable that TDLM perplexity is very close to the GLM for the topics with more hours of data. This suggests that the documents in the topics with more hours of data could be divided into more specific ones to create better TDLMs.

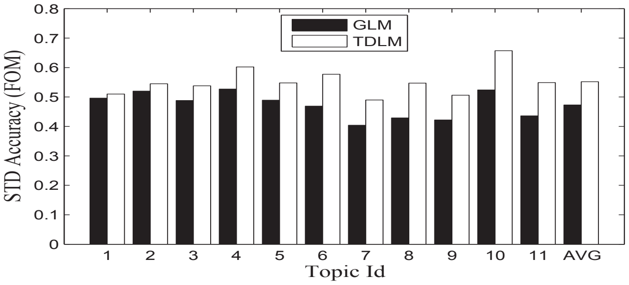

Fig. 4. *STD accuracy (FOM) using TDLM versus GLM*

5.2. Effect of topic dependent LM on STD accuracy

In this section we will investigate if the benefit provided by TDLMs during indexing is also available during the search phase of STD. Figure 4 reports the STD accuracy (FOM) when using topic dependent word LMs and also when using general word LM.

It can be seen that TDLMs have a similar effect on STD accuracy as on the indexing. In average, the FOM is increased approximately from 0.47 to 0.55. However, in general, the improvement for the topics with less hours of data is much more than the ones with more hours of data. For example, topic 10 which according to Table 1 has 0.23 hours of data, in Figure 3 has a high improvement in phone recognition accuracy (around 36%) and in Figure 4 it can be seen that it also has a high improvement in STD accuracy (around 25%). Whereas for topic 1 which has 15 hours of data, the phone recognition accuracy has increased approximately from 63% to 65% and also its STD accuracy is increased by around 3%. By referring to perplexity values displayed in Figure 3, it can be observed that in general, LMs belonging to topics that have less hours of speech, have more improvement in terms of their perplexity compared with GLM and hence represent better models for the documents within their topic. This suggests that better LMs could be created from topics with more hours of data by dividing them into more specific topics. However, despite being low, we still can see FOM improvement for the topics with large amount of data.

6. CONCLUSIONS AND FUTURE WORK

In this paper, topic dependent language modelling was proposed to improve phonetic STD accuracy. It was shown that this approach improves indexing speech segments which results in a SDB with more accurate phone recognition scores for each phone sequence. It was shown that the average FOM was increased. Therefore, topic dependent language modelling improves indexing and this will translate into a more accurate STD system.

In this research, it was assumed that the topic of speech segments in the database was known before indexing. As a future work, this topic could be inferred from the audio information to study its front-end effect on the final STD system. Another area which could be investigated is to find a way to use topic information of the search term in the search phase. TDLM probability of the search term which indicates its membership in each topic could also be used as extra information in STD and perhaps could be used to refine the indexing scores in the search phase for each search term. Furthermore, the results of experiments suggested that topics with equal size data should be investigated for topic dependent STD.

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.

REFERENCES

[1] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, 1996.

[2] Murat Saraclar, Abhinav Sethy, Bhuvana Ramabhadran, Lidia Mangu, Jia Cui, Xiaodong Cui, Brian Kingsbury, and Jonathan Mamou, "An empirical study of confusion modeling in keyword search for low resource languages.," in *ASRU*. 2013, pp. 464–469, IEEE.

[3] K. Thambiratnam and S. Sridharan, "Rapid yet accurate speech indexing using Dynamic Match Lattice Spotting," *IEEE Transactions on Audio, Speech and Language Processing*, 2007.

[4] Roy Wallace, Brendan Baker, Robbie Vogt, and Sridha Sridharan, "The effect of language models on phonetic decoding for spoken term detection," in *ACM Multimedia Workshop on Searching Spontaneous Conversational Speech*, 2009.

[5] 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.

[6] Roy Wallace, Robbie Vogt, and Sridha Sridharan, "Spoken term detection using fast phonetic decoding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[7] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Topic dependent class based language model evaluation on automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2010 IEEE*, Dec., pp. 395–400.

[8] Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa., "Topic dependent class-based n-gram language model," *Audio, Speech, and Language Processing, IEEE Transactions on*, , no. 5, 2012.

[9] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: topic mixtures vs. dynamic cache models," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Oct, vol. 1, pp. 236–239 vol.1.

[10] Yang Liu and Feifan Liu, "Unsupervised language model adaptation via topic modeling based on named entity hypotheses," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4921–4924.

[11] B. Bigot, G. Senay, G. Linars, C. Fredouille, and R. Dufour, "Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech," *Multimedia Tools and Applications*, , no. 2, septembre 2012.

[12] G. Senay, B. Bigot, R. Dufour, G. Linars, and C. Fredouille, "Acoustic person name spotting enriched by lda topic models," in *International conference of the Speech Communication Association ISCA InterSpeech'13*, 2013, p. 1584 1588.

[13] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST Spoken Term Detection evaluation," in *Interspeech*, 2007, pp. 2385–2388.

[14] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," *Proceedings of the IEEE*, , no. 8, 2000.

[15] National Institute of Standards and Technology, "Topic detection and tracking (TDT)," July 2003.

[16] Roy Wallace, Robbie Vogt, Brendan Baker, and Sridha Sridharan, "Optimising Figure of Merit for phonetic spoken term detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5298–5301.

[17] Albert Joseph Kishan Thambiratnam and Subramanian Sridharan, "Dynamic match lattice spotting for indexing speech content," August 2007.

[18] Donglai Zhu, Haizhou Li, Bin Ma, and Chin-Hui 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.

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