SPOKEN TERM DETECTION FOR TURKISH BROADCAST NEWS

Sıddıka Parlak and Murat Saraçlar

Boğaziçi University
Electrical and Electronics Engineering Department
Bebek 34342, Istanbul, Turkey
e-mail: {siddika.parlak, murat.saraclar}@boun.edu.tr

ABSTRACT

In this paper, we present a baseline spoken term detection (STD) system for Turkish Broadcast News. The agglutinative structure of Turkish causes a high out-of-vocabulary (OOV) rate and increases word error rate (WER) in automatic speech recognition. Several approaches are attempted to reduce this negative effect on the STD system. Sub-word units are used to handle the OOV queries and lattice-based indexing is used to obtain different operating points and handle high WER cases. A recently proposed method for setting term specific thresholds is also evaluated and extended to allow us to choose an operating point suitable for our needs. Best results are obtained by using a cascade of word and sub-word lattice indices with term-thresholding.

Index Terms— information retrieval, speech recognition, spoken term detection, audio indexing

1. INTRODUCTION

In the last decades, speech retrieval has emerged as a new field at the intersection of speech processing and information retrieval (IR). Spoken term detection (STD) is a subfield which deals with locating occurrences of a query in an archive. Classical speech retrieval systems convert speech to text using automatic speech recognition (ASR) and then utilize classical text based IR methods for retrieval. However, such methods are designed for text indexation and not adequate for speech retrieval, especially when the ASR system does not have very high accuracy. This problem is more severe for agglutinative languages such as Turkish and Finnish where the out-of-vocabulary (OOV) rates are considerably high. Not only does the higher OOV rates increase the WER but also more of the queries contain OOV words. Thus, additional methods are required to eliminate the ill effects of OOV queries and inaccurate ASR on STD.

In this work, we develop an STD system for Turkish Broadcast News. Our system is also used in a sign language tutoring application [1] where the task is to display the Turkish Sign Language videos corresponding to each query. This can also be viewed as a sign language dictionary. For this task, we prefer to have high precision, instead of high recall. There might be other applications such as monitoring where having a higher recall is more important. Thus, being able to select the operating point is a great advantage. In addition to improved performance, lattice-based search [2] introduces such an opportunity to the system.

In order to handle the OOV problem, various sub-word units have been proposed for modeling of agglutinative languages [3]. In particular, data driven units called morphs and grammatical decomposition of words into stems and lexical endings have been shown to be effective. In addition to language modeling for ASR, sub-word units can also be used for indexing and retrieval. In [4] and [5], morph-based indexing methods are studied for Finnish. In this paper, we experiment with similar methods for Turkish and demonstrate their effectiveness.

For English, phonetic search has been useful for dealing with OOV queries [6]. Since phone recognition is less accurate than word recognition, it is better to use phone indices obtained by converting words into phone strings. In [2] it is shown that combining word and phone indices by means of a cascade strategy yields the best results. We also investigate this approach for Turkish and show that it does not make a considerable contribution to the system. However, the cascade of word and morph indices improves the performance.

Our approach for lattice-based retrieval is based on the general framework of weighted automata indexation [7]. Alternative ways of indexing lattices are given in [2, 8, 9]. Other representations of the information contained in lattices such as confusion networks [10] and position specific posterior lattices [11] can also be used for building indices. We also investigate the use of confusion networks and show that their performance is similar to lattice based methods.

Detection is based on expected counts computed from the lattices and stored in the indices. These counts are compared with a global variable threshold to obtain various operating points. NIST introduced a novel metric for the 2006 STD Evaluation [12], that allows a closed form computation of optimal term-specific thresholds [9]. We adapted this approach and the relevant metrics in addition to traditional precision-recall measurements. As expected, term-specific thresholds outperform having a global threshold.

This paper is organized as follows. In Section 2, we describe the system and explain the methods in detail for each component of the system. We introduce the setup used for our experiments in Section 3 and present the results in Section 4. Finally, we draw our conclusions in Section 5.

2. SYSTEM DESCRIPTION

The overall system has three main components: ASR, indexation and retrieval. First, the ASR component converts the audio data into a symbolic representation (in terms of weighted automata). Indexation of this representation is done via weighted finite state transducer (WFST) operations. These modules operate off-line and the indices are built before seeing the actual queries. When user enters a query, the retrieval module is activated. The retrieved information consists of program name and date, starting time and duration of the query, as well as a relevance score. The detection is based on comparing the scores to a threshold. We now explain each of these components.

2.1. Automatic Speech Recognition

We use an HMM based large vocabulary continuous speech recognition (LVCSR) system. The acoustic models consist of decision tree state clustered triphones and the output distributions are Gaussian mixture models. The recognition networks and the output hypotheses (one-best or lattice) are represented as weighted automata. The details of the ASR system can be found in [3].

We use word and sub-word based n-gram language models (LMs) for ASR. The sub-word units, called morphs, are extracted using an unsupervised word segmentation algorithm [13]. Morph-based LMs provide approximately 5% improvement in WER [3]. SRILM toolkit is used for building the language models and for confusion network generation [14].

2.2. Indexation

Weighted automata indexation is a general framework for efficient retrieval of uncertain inputs [7]. In our case, alternative ASR hypotheses, together with their probabilities, are represented as weighted automata. These automata are processed to extract all (or possibly a restricted subset) of the possible substrings (called factors) contained in the automata. In this process the automata are turned into transducers where the inputs are the original labels of the automata and the outputs are the index labels. Next, these transducers are combined by taking their union. The final transducer is optimized using weighted transducer determinization, resulting in optimal search complexity — linear in the length of the input string. The weights in the index transducer correspond to expected counts.

We apply this indexing method to different units: words, morphs and phones. Note that, recognition and indexing are independent operations. It is possible to obtain phone indices from word or morph lattices, using a lexicon transducer. This is the way we choose to create the phone indices. It is also possible to convert words into morphs for indexing, but there is not much to gain by doing this.

As an additional method, we use confusion networks for indexing words. In this approach, word lattices are converted into confusion networks (sausages) where the weights on each arc correspond to the posterior probabilities. Although it is possible to build an index directly on this representation we chose to apply weighted automata indexing with confusion networks as inputs.

2.3. Retrieval

The queries presented to the system are also represented as finite state automata, and the search is performed by composing these automata with the index transducer. The output contains the list of all utterance indices where the query appears and the corresponding expected counts. The utterances are ranked using the expected counts, and those exceeding a threshold are selected [7]. After obtaining the utterance indices, we apply forced alignment to identify the starting time and duration of each term.

It is important to note that, by varying the threshold on the expected count, different operating points can be obtained. For our application, it is more convenient to operate at a point where precision is high. A recently proposed approach identifies term-specific thresholds optimizing an evaluation metric [9]. As a novel application of this principle, we varied the operating point by changing a parameter of the evaluation metric to suit our needs.

As mentioned earlier, it is possible to employ both word and sub-word indexes for retrieval. We use two cascading strategies to accomplish this. In the vocabulary-cascade method, OOV queries are composed with the sub-word index and in-vocabulary queries are composed with the word index. Search-cascade method functions as follows: First, the query is composed with the word-index. If no results are retrieved, it is composed with the sub-word index [2].

3. EXPERIMENTAL SETUP

3.1. Evaluation Metrics

The metric for evaluating ASR performance is the standard word error rate (WER) metric. OOV rates are calculated based on both word types and tokens. Retrieval part is evaluated via various metrics. The first two, precision-recall rates and F-measure are relatively familiar metrics and calculated as follows: Given Q queries, let the reference transcriptions include $R(q)$ occurrences of the query q, $A(q)$ be the total number of retrieved documents and $C(q)$ be the number of correctly retrieved documents. Then:

$$\text{Precision} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{C(q)}{A(q)} \qquad \text{Recall} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{C(q)}{R(q)} \tag{1}$$

and

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

Our third metric is the "actual term-weighted value" which is defined in NIST STD 2006 Evaluation Plan [12] as:

$$\text{ATWV} = 1 - \frac{1}{Q}\sum_{q=1}^{Q}\{P_{miss}(q) + \beta.P_{FA}(q)\} \tag{3}$$

where β is a user defined parameter (here taken to be 999.9 unless noted otherwise),

$$\text{P}_{\text{miss}}(q) = 1 - \frac{C(q)}{R(q)} \qquad \text{P}_{\text{FA}}(\text{q}) = \frac{A(q) - C(q)}{T_{speech} - C(q)} \tag{4}$$

and T_{speech} is the total amount of speech. For our experiments the query set consists of all the words seen in the reference transcripts, except foreign words and acronyms.

3.2. Corpora

We use two different types of corpora to investigate the improvement caused by lattice usage for various acoustic conditions. The rest of the experiments are performed on only the second corpus.

The first corpus is the 4 hour test portion of our Turkish Broadcast News corpus (BN), including various acoustic conditions. The second one is the Turkish Broadcast News for the Hearing Impaired corpus (HI), consisting of 10 hours of clean and clearly articulated speech. We present the statistics of the test corpora in Table 1. Note that the OOV rates by type (OOV queries) are over 20%, and even OOV rates by token are quite high.

We use the same ASR system for each corpora. The acoustic model is trained on the BN corpus, which has 100 hours of speech. The language models are trained on 96M words of text.

In Table 2, the word and morph based language models are compared on the HI corpus. A comparison between one-best hypotheses in the lattices and confusion networks is also given. As can be seen, the usage of sub-words and confusion networks reduce the WER significantly.

Table 1. WER and OOV rates for different corpora

		OOV rate		# of words	
corpus	WER	type	token	type	token
BN	40.3%	21.0%	7.9%	8932	35314
HI	23.2%	22.6%	6.3%	12770	68020

Table 2. WER of different methods and LM units on HI

unit	one-best	CN-best
word	23.2%	22.1%
morph	20.4%	20.0 %

4. RESULTS

4.1. Using word lattices

We investigate the effect of lattice based search on two different corpora and make a comparison with using the one-best hypotheses for indexing. We also compare these with indexing the one-best hypothesis obtained by posterior decoding using confusion networks on the HI corpus.

The results in terms of various metrics are given in Table 3. The improvement due to lattice indexing is higher in the case of BN data where the WER is higher. The one-best obtained from confusion networks gives only a very slight improvement over the baseline one-best.

Table 3. Maximum Precision, Recall, F-measure and ATWV values for one-best, CN-best, and lattice on BN and HI corpora

	maxP	maxR	maxF	maxATWV
one-best	82.2	48.0	60.5	46.6
lattice	94.3	63.2	62.9	51.8

(a) BN corpus

	maxP	maxR	maxF	maxATWV
one-best	86.9	62.3	72.4	60.8
CN-best	87.3	62.3	72.7	60.9
lattice	94.6	72.7	73.5	64.5

(b) HI corpus

4.2. Using confusion networks

Despite the slight improvement in WER and retrieval performance of the one-best hypotheses obtained from confusion networks, indexing based on the full confusion networks gave almost identical performance as lattice indexing. However, it is possible to obtain smaller indices by using confusion networks [15].

4.3. Using morph lattices and cascades

Next, we investigate the effects of using morph language models in ASR and morph-based indexing. The precision-recall graphs of both word and morph based approaches are shown in Figure 1. Morph-based recognition and indexing increases the recall significantly, however precision saturates at a lower value than word indexing. Having different characteristics, these two approaches can be combined via strategies mentioned in Section 2.3. The results, presented in the same graph, show that both cascading methods are superior to using each index individually. In the high precision region ($> 90\%$), vocabulary cascade outperforms search cascade, while search cascade is better in the remaining region. In Table 4, we use F-measure and ATWV metrics to compare the cascading strategies. The search-cascade strategy performs slightly better than the vocabulary-cascade in terms of maxF and maxATWV.

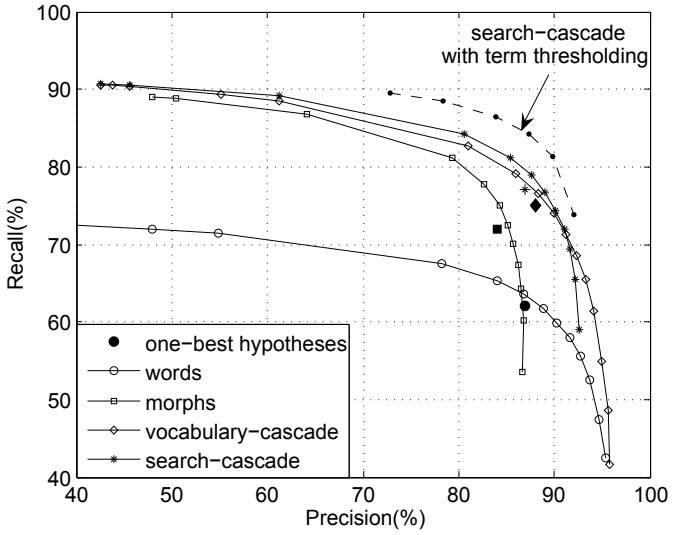

Fig. 1. Comparison of word, morph and hybrid indexing strategies on the HI corpus in terms of precision and recall. Solid single markers indicate the performance of one-best hypotheses.

Table 4. Performance of various methods in maximum F-measure and maximum ATWV (VC:vocabulary cascade, SC:search cascade, TTh: term thresholding)

	word	morph	VC	SC	SC+TTh
maxF	73.5	80.3	82.4	83.3	85.6
maxATWV	64.5	75.8	79.5	81.1	85.7

It is interesting to note that the improvement in retrieval performance is much more impressive than the improvement in WER when morph-based language models and indices are used instead of word-based language models and indices.

4.4. Using phones and cascades

We built two different phone indices: one by converting words to phones and another by converting morphs to phones. Phone indices, individually, contribute slightly to recall at low precision points. However, as precision gets higher, recall degrades dramatically. In other words, although the maximum recall attainable by using phone-based indexing is slightly larger, the maximum precision possible is much lower.

We also experimented with forming cascades of the phone indices with the word index. Cascade usage yields better results than individual phone and word indices, but not better than the word-morph cascade. From these results we conclude that, unlike English phone indexing is not so beneficial for Turkish. This might be due to the fact that Turkish is almost a phonetic language and we base our acoustic models on graphemes instead of phonemes. It could also be argued that the gain from phonetic indexing in the case of English is due to homophones.

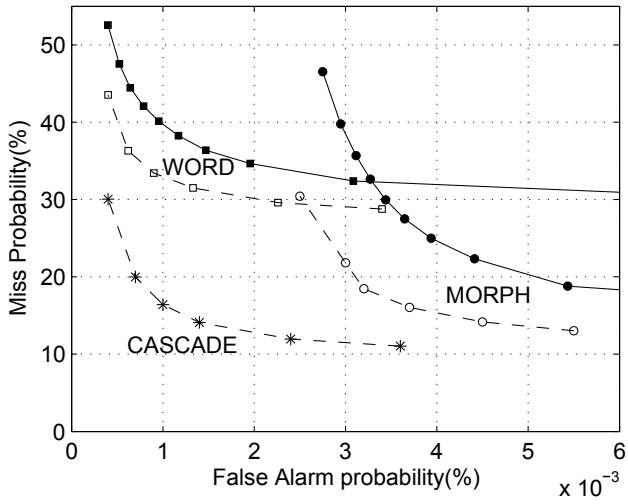

Fig. 2. DET curve for using term-specific thresholds on the HI corpus. Solid lines represent using a global threshold while dashed lines represent using optimal term-specific thresholds.

4.5. Using term-specific thresholds

It was shown in [9, 8] that setting term-specific thresholds outperform using a global threshold. The strategy suggested in [9] is to determine the term-specific thresholds which maximize ATWV for a given β value. The threshold is calculated as:

$$\mathrm{th(q)} = \frac{R(q)}{\frac{T_{speech}}{\beta} + \frac{\beta-1}{\beta}R(q)} \tag{5}$$

The exact value of $R(q)$ is not known but approximating this number by the expected number of occurrences of q in the test corpus is reasonable and works in practice.

For the NIST 2006 STD Evaluation the β value was taken to be 999.9, which yields a high false alarm probability for our task. By increasing the β in Equation 5, we obtain different operating points which can be represented as a curve.

The results of term thresholding approach are presented in Figure 2 for word and morph indices as well as their cascade. Significant gains are obtained with term-specific thresholds in terms of all evaluation metrics, as can be seen in Table 4 and Figures 1 and 2.

5. CONCLUSION

We developed a baseline STD system for Turkish Broadcast News. Various methods were attempted to deal with challenges posed by the agglutinative nature of Turkish. Among these, using a cascade of word lattice based and morph lattice based indices gave the best results. Using confusion networks and phonetic indexing did not yield any improvements. Using term specific thresholds optimizing the ATWV metric resulted in better performance for all methods and evaluation metrics. By varying a parameter of the ATWV metric we were able to obtain operating points which are more suitable for our sign language dictionary tool. Combination of using lattices, morphs, cascading, and term-specific thresholds improved the F-measure from 72.4 to 85.6 and ATWV from 60.8 to 85.7.

6. ACKNOWLEDGEMENTS

The authors would like to thank Ebru Arısoy for the ASR setup and models, Sabanci and ODTU universities for the Turkish text data and AT&T Labs – Research for the software. This research is supported by TUBITAK (Scientific and Technological Research Council of Turkey) (Project codes: 105E102 and 107E021) and Boğaziçi University Research Fund (Project codes: 05HA202 and 07HA201D).

7. REFERENCES

[1] O. Aran, I. Ari, P. Campr, E. Dikici, M. Hruz, D. Kahramaner, S. Parlak, L. Akarun, and M. Saraclar, "Speech and Sliding Text Aided Sign Retrieval from Hearing Impaired Sign News Videos – eNTERFACE-2007 Final Report," `http://www.cmpe.boun.edu.tr/enterface07/outputs/final/p3report2.pdf`, 2007.

[2] M. Saraclar and R. Sproat, "Lattice-Based Search for Spoken Utterance Retrieval," in *Proc. HLT-NAACL*, 2004.

[3] E. Arisoy, H. Sak, and M. Saraclar, "Language Modeling for Automatic Turkish Broadcast News Transcription," in *Proc. Interspeech*, 2007.

[4] M. Kurimo, V. Turunen, and I. Ekman, "An Evaluation of a Spoken Document Retrieval Baseline System in Finnish," in *Proc. Interspeech*, 2004.

[5] M. Kurimo and V. Turunen, "To Recover from Speech Recognition Errors in Spoken Document Retrieval," in *Proc. Interspeech*, 2005.

[6] B. Logan, J. M. V. Thong, and P. J. Moreno, "Approaches to Reduce the Effects of OOV Queries on Indexed Spoken Audio," *IEEE Transactions on Multimedia*, vol. 7, no. 5, October 2005.

[7] C. Allauzen, M. Mohri, and M. Saraclar, "General-Indexation of Weighted Automata - Application to Spoken Utterance Retrieval," in *Proc. HLT-NAACL*, 2004.

[8] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 Spoken Term Detection System," in *Proc. Interspeech*, 2007, pp. 2393–2396.

[9] D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and Accurate Spoken Term Detection," in *Proc. Interspeech*, 2007.

[10] V. T. Turunen and M. Kurimo, "Indexing Confusion Networks for Morph-based Spoken Document Retrieval," in *Proc. SIGIR*, 2007.

[11] C. Chelba, J. Silva, and A. Acero, "Soft Indexing of Speech Content for Search in Spoken Documents," *Computer Speech and Language*, vol. 21, no. 3, pp. 458–478, July 2007.

[12] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan," `http://www.nist.gov/speech/tests/std/`, 2006.

[13] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0," Tech. Rep., Helsinki University of Technology, 2005.

[14] A. Stolcke, "SRILM-An Extensible Language Modeling Toolkit," in *Proc. Interspeech*, 2002, pp. 901–904.

[15] T. Hori, I.L. Hetherington, T.J. Hazen, and J.R. Glass, "Open-vocabulary Spoken Utterance Retrieval Using Confusion Networks," in *Proc. ICASSP*, 2007, pp. 73–76.