COMBINING CONFUSION NETWORKS WITH PROBABILISTIC PHONE MATCHING FOR OPEN-VOCABULARY KEYWORD SPOTTING IN SPONTANEOUS SPEECH SIGNAL

Shan Jin, Thomas Sikora

Department of Telecommunication Systems, Technical University of Berlin
Einsteinufer 17, D-10587, Berlin, Germany
phone: + (49) 30 31428505, fax: + (49)30 31422514, email: shan@nue.tu-berlin.de
www.nue.tu-berlin.de

ABSTRACT

In this paper, we study several methods for keyword spotting in spontaneous speech signal. Novel method combining probabilistic phone matching (PSM) approach with word confusion networks (WCN) is proposed for open-vocabulary keyword spotting task. This method runs keyword spotting on multi-level transcriptions (WCN and phone-onebest). We propose to use classical string matching for word spotting on WCN. At the same time probabilistic string matching is used for acoustic word spotting on phone-onebest transcription. It is verified that the novel hybrid method outperforms WCN-based and PSM-based approaches in-vocabulary and out-of-vocabulary (OOV) keywords.

1. INTRODUCTION

Keyword spotting is the task dealing with the identification of keywords in speech data. It is widely used in automatic control, human-computer interface, information retrieval in spoken document etc. With a growing amount of accessible online audio-visual material getting and using the data in an effective and reliable way has become a key issue. Audio streams of multimedia documents often contain spoken parts. We focus on the keyword spotting techniques applied for spoken document retrieval (SDR) in this paper.

Generally SDR systems consist of two parts: indexing and retrieval. Indexing tools transcribe speech data in word or sub-word representations. Based on the representation retrieval tools compute a similarity score for each document according to the query. Query could be seen as a sequence of keywords. Hence SDR could also be treated as a special multi-keyword spotting task.

Recently, SDR approaches could be classified in four main categories according to the type of indexing units: word-based approaches, subword-based approaches, phone/phoneme-based approaches and combined approaches. Word-based approaches [1] rely on large vocabulary automatic speech recognition (ASR) systems that transcribe spoken data in a word sequence. Text string matching algorithms can then be used to find the information. Even though such strategies are able to achieve a reasonable performance, the size of the recognizable vocabulary restricts the number of words in queries. In [5], Logan et. al. reported that about 13% of user queries contain out-of-vocabulary (OOV) words. Moreover, OOV words pose a serious problem in a word-based SDR system, particularly in domains where new words appear frequently over a short period of time. Some experts try to overcome this restriction with sub-word units like VCV [2]. The main drawback of this method is that the sub-word units are extracted directly from text without taking their acoustic properties into consideration. Some other experts try to overcome the OOV-problem with phone/phoneme-based issue [8]. It was verified that a phone/phoneme-based matching algorithm could address the issue of OOV-words encountered in a word-based matching algorithm. However, its performance depends heavily on the accuracy of the phonetic transcription. Typically, a phone recognizer can achieve only an accuracy of 50% as compared to 80% accuracy of a domain dependent word recognizer. Hence approaches that combining different indexing sources are investigated ([3], [4]). James [4] combined word and phone recognition in a complete recognition system in which the phone recognizer is only used to spot the OOV-words. This combination improves the retrieval effectiveness of a SDR system but it needs more training data and more efforts for building two recognizers.

Most approaches announced before deal with one-best ASR output. It is reliable with low error rate. However with increasing error rate, some important information could be lost in one-best transcription. Some techniques have been developed dealing with multiple hypotheses from an ASR system, in which word and/or sub-word lattice are used to index speech data [12]. Techniques are also investigated to reduce the size of multiple hypotheses provided by an ASR system, e.g. confusion network proposed by Mangu [7]. Confusion network has the most compact structure representing multiple hypotheses while keeping the order of symbols along the time axis. It is verified that using confusion network for indexing could achieve more robust keyword matching in erroneous recognition hypotheses.

In this paper, we propose a method that combines WCN-based and phone-onebest based approaches for in-vocabulary (INV) and out-of-vocabulary (OOV) words spotting in spontaneous speech data. The proposed system runs matching on multi-level transcription (WCN and phone-onebest). Only 20k word ASR is built for transcribing the speech data into word-onebest and word lattice transcription. With help of a pronunciation dictionary phone-onebest transcription is extracted from word-onebest ASR output directly. Classical string matching algorithm is used to find INV-keyword in WCN. Acoustic matching for both INV and OOV-keywords is realizd with PSM on phone-onebest transcription extracted directly from word-onebest with help of pronunciation dictionary. The results of WCN-based string matching and phone-onebest based PSM are then be combined together with a match's confidence score. This method outperforms

WCN-only and PSM-only baseline systems.

2. HYBRID KEYWORD SPOTTING METHOD

The proposed hybrid keyword spotting approach is shown in figure 1. It consists of three main modules: multi-level indexing, hybrid Matching and scoring/ranking.

2.1 Multi-level indexing

In this paper, we try to use the multi-level information (phone one-best, and WCN) to improve the performance for both INV and OOV words. It is verified that the phone-based indexing method is effective especially for OOV keywords. However it yields generally a lower precision for INV queries than word-based indexing. The benefits with combined word and phone hypotheses has been shown in recent works [6]. It is reported that the combination of word and phone confusion networks is effective to achieve high retrieval performance for both INV and OOV queries.

We propose a method that combines WCN with phone one-best indexing to improve the detection rate of both INV and OOV keywords. Word-based ASR takes as input the speech signal and generates word one-best transcription and word lattice. Phone one-best transcription is constructed by a word-2-phone module with the help of the pronunciation lexicon. Each word in the one-best transcription is replaced with its corresponding phone sequence.

A word lattice is generated for INV word spotting and is reduced with Mangu's Algorithm [7]. As a more compact representation of lattice, WCN is constructed step-by-step:

- First the posterior probability is computed for all edges in the word lattice.
- Second the edges with posterior probability far below one are removed from the lattice.
- Then edges corresponding to the same word instance with time-overlap are grouped into one cluster. Cluster posteriors are set to the sum of all clustered edges' posteriors.
- In the last inter-word clustering which groups different words that compete around the same time interval with similar phonetic properties.

2.2 Hybrid matching

Different matching strategies are used for word spotting on multi-level transcriptions. The probabilistic phone matching algorithm [8] is applied for phone-based word spotting. A classical string matching algorithm is used to find keyword in WCN.

The probabilistic phone matching algorithm is based on a 1-best phone transcription and consists of search term location and search term weighting. The task of the slot-detection component is to find possible slots in transcription which may contain the keyword sequence. It is assumed that most errors in transcription are substitution errors. The slots which have a sufficient conformity with the keyword phone sequence are estimated. The conformity is measured as the number of common phones, the same phone occurring at the same position within the keyword phone sequence and slots. A slot is verified when its number of common phones is greater than a pre-defined threshold value.

Phone confusion information is used for slot probability estimation. Phone confusion information in form of a matrix contains statistics on the substitution errors, insertion errors and deletion errors of a phonetic recognizer. In our case, there are 39 phone classes. The dimension of our confusion matrix obtained by running the phone recognizer on development data set is 40*40. The upper left sub matrix with dimension 39*39 including all substitution errors is called the substitution matrix, in which a component $C(r,l)$ corresponds to the number of phone r recognized as phone l. The last column contains the number of deletion errors that indicates the probability that a phone is spoken but not recognized. The last row contains the number of insertion errors for each phone, which gives the probability that a phone is not spoken but recognized. Slot probability is then estimated as follows [8]:

$$
\begin{aligned}
sim(s_0,q_v) &:= P_{sub}(q[v],s[0]); && (1)\\
sim(s_u,q_0) &:= P_{sub}(q[0],s[u]); && (2)
\end{aligned}
$$

$$
sim(s_u,q_v) = max\begin{cases} sim(s_{u-1},q_{v-1}) + P_{sub}(q[v],s[u]); \\ sim(s_u,q_{v-1}) + P_{del}(q[v]); \\ sim(s_{u-1},q_v) + P_{ins}(s[u]); \end{cases} \qquad (3)
$$

where S_u is the sub-string of the $u+1$ first phones of s (slot detected); q_v is the sub-string of the $v+1$ first phones of q (keyword phone sequence); $q[v]$ is the $v+1$ phone in q; $s[u]$ indicates the $(u+1)$ phone in detected slot; $P_{sub}(q[v],s[u])$ is the probability that phone $q[v]$ is substituted with phone $s[u]$; $P_{del}(q[v])$ indicates the probability that the phone $q[v]$ is deleted; and $P_{ins}(s[u])$ is the probability that the phone $s[u]$ is inserted. The slot-probability estimation is implemented using dynamic programming.

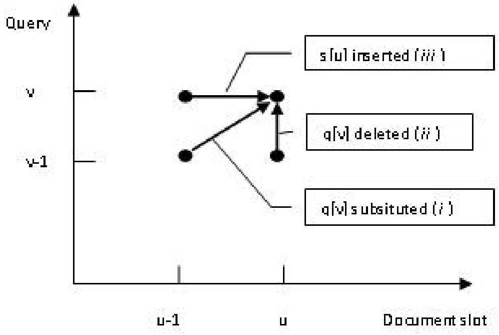

Figure 2: DP transitions used

The three alternatives defined in equation (3) are illustrated as arrows in a two-dimensional grid defined by the slot phone sequence (x-axis) and the keyword phone sequence (y-axis). Figure 2 shows a simpler and more straightforward recursive scheme used in the work described here.

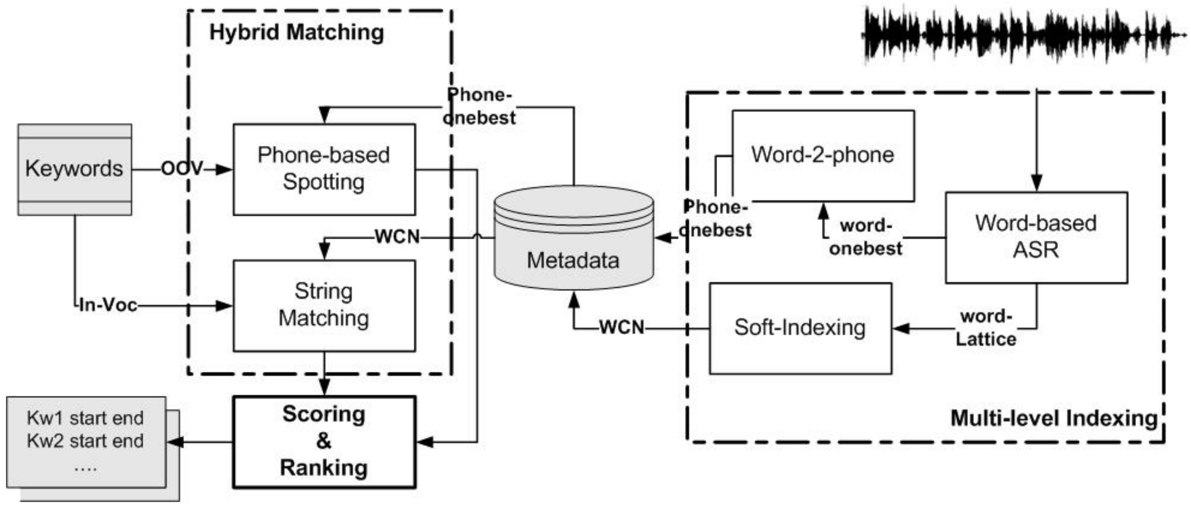

Figure 1: WCN-PSM keyword spotting sytsem structure

2.3 Scoring and ranking

If the given keyword is found in WCN with a string matching algorithm, the score of this candidate will be set to 1, otherwise 0. For the single-keyword spotting task the ranking score of INV word's candidate is estimated as the sum of WCN-based string matching score and the PSM-score. The ranking score of OOV-candidates is set to its PSM-matching score.

For multi-keyword spotting task the ranking score is set to:

$$rank\_score = \frac{1}{rank\_in\_lst} + \lambda * single\_rank\_score. \qquad (4)$$

where *rank_in_lst* is the rank in the result list of single-keyword spotting task; The weight λ is selected experimentally; *single_rank_score* is the ranking score estimated for single-keyword spotting task.

3. EXPERIMENT

3.1 Experiment setup

Wall Street Journal (WSJ) corpora si-dt-s2 data is selected to evaluate the spotting performance of the proposed system. This spontaneous clean speech data consists of a set of single-sentence documents covering ten different domains. There is a total of 207 sentences spoken by 10 persons (5 female and 5 male). 200 INV-words with 304 occurrences and 40 OOV-words with 52 occurrences are selected for the evaluation procedure.

The Word-based ASR (20k) is trained on 16kHz mono audio speech (WSJ). A Gaussian mixture model is used to model 8,000 tied states (16 Gaussians per state). The acoustic model was initialized with the TIMIT corpus [11] and trained on WSJ training data (all). The bigram language model is trained on NoV-92 LM training data (20k vocabulary). This word-based ASR achieves an accuracy rate of 60% on the test data set with an OOV-rate of 7%. This word-based ASR provides the word-onebest and word lattice at the same time for further analysis. Phone onebest transcription is constructed directly from the word-onebest transcription with the help of the CMU pronunciation dictionary [10]. The WCN is constructed using the method described in section 2.1.

Precision, recall and mean average precision (mAP) are selected metrics for system performance evaluation [9]. As defined in formula (5) precision is the number of detected true hits over the total number detected candidates. Recall is defined in formula (6) as the number of detected true hits over total true hits in the collection of documents.

$$Precision = \frac{Detected\_true\_hits}{Nr\_of\_detected\_Candidates} * 100\%. \qquad (5)$$

$$Recall = \frac{Detected\_true\_hits}{Total\_hits\_in\_test\_data} * 100\%. \qquad (6)$$

Sometimes it is troublesome to compare the performance of different retrieval systems using precision-recall curves. Therefore single performance measure, mAP, is commonly used. The mAP value is obtained by averaging the precision values across all recall points. The mAP value can be interpreted as the area under the precision-recall curve.

3.2 Results and discussion

The objective of the experiments was to examine whether the keyword spotting performance can be improved by combining WCN and phone-onebest based PSM techniques. The following keyword spotting approaches were compared: WCN-based string matching, phone-onebest based PSM only and the proposed method that combines both of them. Because that the mean average precision is a single number, and many performances information may be hidden. Two keyword spotting systems with different precision-recall curves can have the same mAP value (the area under the precision-recall curve is the same). Therefore we generally use both of them, i.e. the precision-recall curve and the mAP value, to evaluate the performance of our system.

Table 1: mAP (%) obtained with different keyword spotting approaches

	INV	OOV	Total
WCN	70.0	0	62.2
PSM	46.5	18.8	41.2
WCN+PSM	73.8	18.8	63.7

Table 1 shows the mAP values obtained using 240 selected keywords. It is observed that WCN-based string matching approach reached mAP value of 70% even with high WER (40%) in the word transcription. None of the OOV-words could be detected with this method and consequently mAP of 62.2% for all INV- and OOV-keywords was obtained. The phone-onebest transcription includes a lot of errors even though it was extracted directly from ASR word-onebest output. Hence phone-onebest based PSM algorithm could detect OOV-words but with degrading effectiveness (mAP of 41.2% for all INV- and OOV- words). It can only reach a mAP value of about 46.5% for INV-words and 18.8% for OOV-words. The PSM approach yielded a better performance for INV-words than OOV-words because the phone-onebest transcription was extracted directly from ASR word-onebest output with the help of the pronunciation dictionary. The chance of an exact matching between INV-keyword phone sequence and a slot detected in the phone-onebest transcription is much higher than one with OOV-keywords. It could be observed, too, our proposed method combining both WCN and PSM strategy (as explained in 2), obtained the best performance with a mAP of 63.7% for all INV- and OOV-words. The mAP is improved by about 2.4% relative to the WCN-based approach. The novel WCN/PSM system achieves an improvement in the mAP value of about 54.6% relative to a PSM-only approach.

Figure 3 shows a similar behaviour in form of recall-precision curves. At almost all levels of recall the combination of WCN and PSM could yield better precision values. Especially at a recall level of 60% an improvement of about 40% (in precision value) could be observed.

4. CONCLUSION

In this paper, we investigated different keyword spotting techniques. We have successfully combined WCN-based string matching and phone-onebest based PSM for keyword spotting in spontaneous speech signal. Despite high word error rates in word-level transcription, WCN-based approaches could obtain mAP of about 70%. OOV-keywords could be detected with PSM approaches. Experiments showed that the combination of these two systems could improve the keyword spotting performance.

5. ACKNOWLEDGMENT

This research was supported by the Europenan Commission under the contracts IST-1-038398-VISNET II and FP7-216444 PetaMedia. The views expressed in this paper are those of the authers and do not necessarily represent the views of the commission.

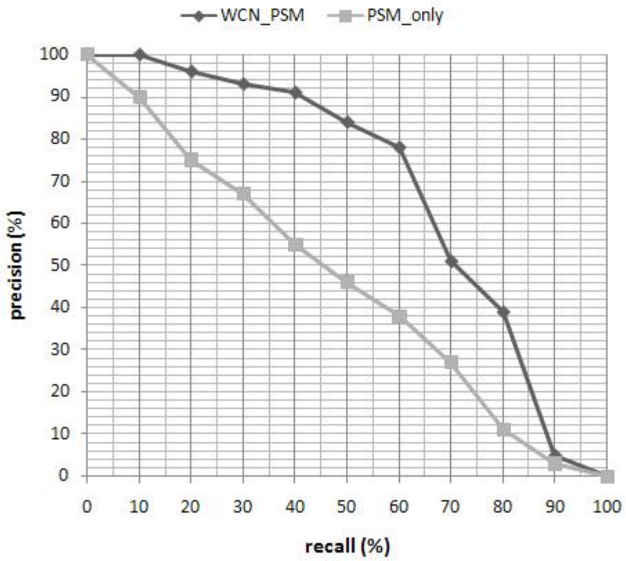

Figure 3: Precision/Recall plot of the phone-based and multi-level keywords spotting sytsem

REFERENCES

[1] B. Zhou, J.H.L. Hansen, “SpeechFind: an Experimental on-line Spoken Document Retrieval System for historical Audio ,” in *Proc. ICSLP-2002*, Denver, Co USA, Sept. 2002, vol 3 pp. 1969-1972

[2] U. Glavitsch,*A First Approach to Speech Retrieval*, technical report TR 238, Information Systems, Swiss Federal institute of technology (ETH), Zurich.

[3] K. Ng, “Information Fusion for Spoken document retrieval, ” in *ICASSP’00*, Istanbul, Turkey, 5-9. June 2000, vol.4 pp.2405-2408

[4] D. James, *The application of classical information retrieval techniques to spoken documents*, PhD Thesis, university of Cambridge, UK.

[5] B.M. Logan, T. Pedro, J.-M. v. Whittaker,“An experimental study of an audio indexing system for the Web. ” in *ICSLP-2000*, Beijing, China. vol.2, pp. 676-679

[6] T. Hori, I. L. Hetherington, T. J. Hazen and J. R. Glass,“Open-vocabulary spoken utterance retrieval using confusion networks,” in *Proc. ICASSP 2007*, Honolulu, Hawai’i, U.S.A, April 15-20, 2007, vol 4, pp. IV-73–IV-76.

[7] L.Mangu, E. Brill and A. Stolcke, “ Finding consensus in speech recognition: word error minimization and other application of confusion networks”, *Computer speech and language*, vol. 14, pp. 373–400, 2000.

[8] M. Wechsler, “Spoken document retrieval based on phoneme Recognition,” Ph. D. thesis, Swiss Federal institute of Technology(ETH), 1998

[9] D.K Harman, *Sixth Text Retrieval conference (TREC-6)*. Gaithersburg MD, USA. National institute for Standards and Technology, NIST-SP 500-240.

[10] B. Weide, *The Carnegie Mellon Pronouncing Dictionary*, Department of Computer Science,

Carnegie Mellon University. Available by Internet as ftp://ftp.cs.cmu.edu/project/fgdata/dict.

[11] W.M. Fisher, G.R. Doddington and K.M. Gaudi-Marshall, “The DARPA Speech Recognition Research Database: Specifications and Status,” in *Proceedings of DARPA Workshop on Speech Recognition*,1986 pp. 93-99

[12] M. Saraclar and R. Sproat, “lattice-based search for spoken utterance retrieval,” in *Proc. HLT-NAASCL*, 2004.