A HYBRID METHOD WITH CONFUSION NETWORK FOR INDEXING SPOKEN DOCUMENTS IN E-LIBRARIES

BENDIB Issam and LAOUAR Mohamed Ridda
LAMIS Laboratory, University of Tebessa
Route de Canstantine, 12002, Tebessa Algeria
bendib2012@gmail.com, ridda_laouar@yahoo.fr

***Abstract*—**The technological development of storage techniques and retrieval methods excite the managers of E-libraries to integrate this resources in their systems. In practice, the creation of these systems imposes the definition of document indexing techniques. In this effect, indexing the spoken content of audio recordings requires the use of automatic speech recognition. Although the research areas of spoken words and audio retrieval has been well addressed, but still significant limitations to achieve, especially in terms of multimedia resource available today. In this paper we propose an indexing procedure for spoken document in field of E-libraries. Our method use the word confusion network as being a representation of alternative recognition candidate by aligning mutually exclusive terms and by giving the posterior probability of each term. The rank of the competing terms and their posterior probability is used to estimate term frequency for indexing. In parallel, we calculate the posterior probability of terms in their specific position in lattice generated with the transcription provided by a large vocabulary continuous speech recognition system. A validation of this approach of indexing and information retrieval is in the course of validation for the field of the E-libraries.

Keywords—Spoken document retrieval; Word confusion networ; LVCSR; PSPL; insert (key words)

1. INTRODUCTION

Actually, the automatic systems for indexing, archiving, searching and browsing of large amounts of spoken documents have become a reality in the last decade. As more and more spoken information is produced and archived, there is an increasing need for indexing and retrieving audio material based of their content. At this effect, the indexing of multimedia documents, in particular, the detection of keywords in audio files currently was attracting great interest both in terms experimentally and theoretically. However, despite the progress made in the field of indexing speech, much remains to be done notably for the key word search in spontaneous speech. Further, this evolution has allowed to pass from indexing structured data (Radio broadcasts the documentary), the indexing of unstructured data (such as spontaneous speech). In literature, among the indexing systems of unstructured data, we find the indexing systems of vocal emails, video conference and voice annotations of images [1], [2]. These systems have introduced new problems such as hesitations, false starts, poor structuring of sentences…etc. This has encouraged researchers to suggest implementing new techniques to deal with these problems.

Several retrieval techniques dealing with multiple hypotheses from an ASR system have been proposed, in which word and/or sub-word lattices are used to index each utterance. In this context, a very successful new approach of indexing speech information with a very compact structure: Position-Specific Posterior Lattices (PSPL) has been recently proposed [3]. This approach efficiently considers all possible paths in the recognized lattice, as well as word proximity information within the lattice. However, the out of vocabulary (OOV) problem is still left unaddressed in PSPL; that is, OOV words generally do not appear in the recognized lattice. However, sub-word based representations such as phone lattices are crucial especially for OOV words. It is also effective to combine word and sub-word lattices to achieve high retrieval performance for both IV and OOV queries since subword-based indices generally yield a lower precision for IV queries compared with word-based ones.

Although, a more compact representation of a lattice called word confusion network (WCN) has been proposed [4]. Each word in a confusion set is associated with its posterior probability i.e. the probability of the word given the signal at the time interval. Sentence hypotheses can be generated by freely combining hypotheses at each alignment position.

In general spoken utterance retrieval (SUR), the process is separated into two parts. The first one is indexing and the second one is search. Indexing is an offline process in which linguistic information is extracted from all speech data in the archive, and an index table is built. A speech recognizer is used to extract such linguistic information. The index table maps each extracted linguistic symbol (word, subword or phrase) to a set of utterances that the symbol matches. Search is an online process in which users' queries are accepted and utterances that each query matches are found. The system finds the target utterances efficiently using the previously built table. The index table helps to search with a desirable speed that is almost independent of the size of the archive.

This paper is organized as follows, in section 2 we give an overview of related work, focusing on methods dealing sub-word and lattice. We present our approach of indexing in spoken document using PSPL and WCN in the field of E-libraries. Experimental process is given in Section 4. Finally, our conclusions are presented in Section 5.

2. RELATED WORK

Jones et al. describe a system that combine a large vocabulary continuous speech recognition (LVCSR) system and a phone-lattice word spotter (WS) for retrieval of voice and video mail message [5]. Srinivasan and petkovic introduce a methode for phonetic retrieval based on the probablistic formulation of term weighting using phone confusion data [6]. Logan et al compare three indexing methods based on words, syllable-like particle, and phonemes to study the problem of OOV queries in audio indexing systems. they have give an alternate approach to the OOV query problem by expanding query words into in-vocabulary phrases while taking acoustic confusability and langage model scores into account [7].

James et al propose a new method for speech file indexing for an unlimited vocabulary. This approach consists in generating in differed time a phoneme lattice for each audio file using a modified version of the Viterbi algorithm [8]. The detection of keywords is performed by dynamic comparison between the search word and phoneme sequences in the lattice.

Saraclar and Sproat proposes an approach that builds an inverted index from ASR lattices-word or phone (sub-word) level-by storing the full connectivity information in the lattice; retrieval is performed by looking up strings of units. This approach allows for exact calculation of n-gram expected counts but more general proximity information (distance-k skip n-gram, $k > 0$) is hard to calculate. No compression of the original [9].

Retrieval performance can be increased by extracting alternative hypotheses from the recognizer in addition to the most probable (1-best) candidate. A lattice is a graph containing a number of most probable hypotheses considered by the recognizer and can be used as a source for extracting additional terms. A more compact representation for the hypotheses is the word confusion network (WCN), which offers a convenient representation of competing terms along with the posterior probability for each term. Mamou et al. have shown improvements of SDR performance in low accuracy conditions by indexing and weighting terms in confusion networks based on their probability and rank among competitors [10].

Chelba et al. present their method of indexing spoken documents, which uses a lattice "simplified" for each segment. The PSPL is in fact a lattice representation of speech recognition will not register the arcs, but the likelihood that an arc that generates a word is placed at a position in the lattice. This enables a rich, compact representation of lattice, which allows approximate accounts for n-grams. The position information will also allow taking into account the proximity of search terms in the measure of relevance [11].

3. METHODS

Digital libraries have now become the most popular area for users of documentary resources. Thus, today, and with the great advances of web technologies, structuring, and access to information and multimedia resources by the speech recognition has become an important necessity for managers and users of these libraries. It is this necessity that encouraged us to propose an approach for a fast and efficient access to multimedia resources [12]. This approach will then be the basis for developing an indexing system for multimedia digital libraries (E-Library).

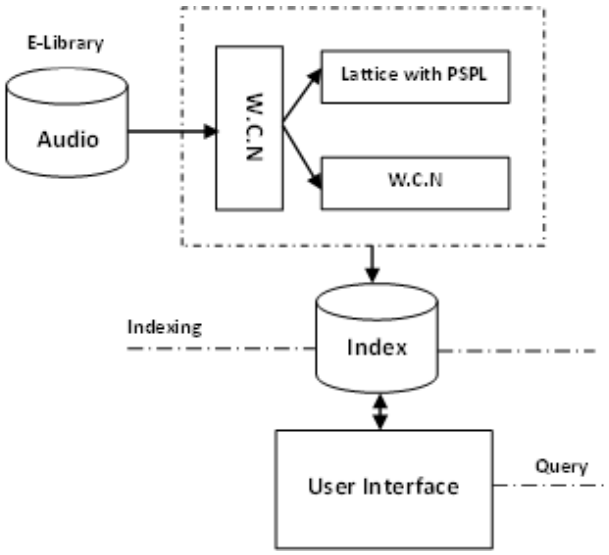

Fig. 1. System Architecture.

In this effect, we present the global structure of our system and explain details of the techniques used in our work as shown in Figure 01. The system contains three main components. First, the ASR component is used to convert speech into a lattice representation, together with timing information. Second, this representation is indexed for efficient retrieval. These two steps are performed off-line. Finally, when the user enters a query the index is searched and matching audio segments are returned.

3.1.Large Vocabulary Continuous Speech Recognition

We use acoustic models with HMM, he consists of decision tree state clustered tri-phones and the output distributions are mixtures of Gaussians. The language models are trigram models. The pronunciation dictionaries contain few alternative pronunciations. The ASR systems used in this study are single pass systems. The recognition networks are represented as weighted finite state machines (FSMs). The output of the ASR system is also represented as an FSM and may be in the form of a best hypothesis string or a lattice of alternate hypotheses. The labels on the arcs of the FSM may be words or phones, and the conversion between the two can easily be done using FSM composition. The costs on the arcs are negative log likelihoods. Additionally, timing information can also be present in the output.

3.2.Indexing With Position-Specific Posterior Lattices (PSPL)

The basic idea of PSPL is to calculate the posterior probability prob of a word W at a specific position *pos* in a lattice for a spoken segment *d* as a tuple *(W, d, pos, prob*). Such information is actually hidden in the lattice *L* of d since in each path of *L* we clearly know each word's position. Since it is very likely that more than one path includes the same word in the same position, we need to aggregate over all possible paths in a lattice that include a given word at a given position.

A variation of the standard forward-backward algorithm can be employed for this computation. The forward probability mass *α(W, t)* accumulated up to a given time *t* at the last word *W* needs to be split according to the length *l* measured in the number of words:

$$\alpha(w,t,l) \doteq \sum_{\substack{\pi:\text{a partial path ends at time } t,\\ \text{has last word } W \text{ ,and inludes}\\ l \text{ ssubword units}}} P(\pi)$$

Where π is a partial path in the lattice. The backward probability *β(W, t)* retains the original definition [2]. The elementary forward step in the forward pass can now be carried out as follows:

$$\alpha(W,t,l) = \sum_{W'} \sum_{\substack{t':\exists \text{ edge } e \\ \text{starting at time } t', \text{ ending at time } t,\\ \text{and with word}(e)=W}} \begin{bmatrix} \alpha(W',t',l') \cdot P_{AM}(w) \\ \cdot P_{LM}(w) \end{bmatrix}$$

Where $P_{AM}(W)$ and $P_{LM}(W)$ denote the acoustic and language model scores of *W* respectively; e is a word arc in the lattice and *word(e)* means the word entity of arc *e*. The position specific posterior probability for the word *W* being the l^{th} word in the lattice is then:

$$P(W,b,b+Sub(W)-1|L) = \sum_{t} \frac{\alpha(W,t,b+Sub(W)-1)\cdot\beta(W,t)}{\beta_{start}} \cdot Adj(W,t)$$

Where βstart is the sum of all path scores in the lattice, and *Adj(W, t)* consists of some necessary terms for probability adjustment, such as the removal of the duplicated acoustic model scores on *W* and the addition of missing language model scores around *W* [11]. In this paper, we regard the tuples *(W, d, pos, prob)* for a specific spoken segment *d* and position *pos* as a cluster, which in turn includes several words along with their posterior probabilities.

3.3.Indexing With Word Confusion Network (WCN)

A Confusion Network (CN) is a weighted directed graph with a start node, and end node, and word labels over its edges. The CN has the peculiarity that each path from the start node to the

end node goes through all the other nodes as shown in Figure 02.

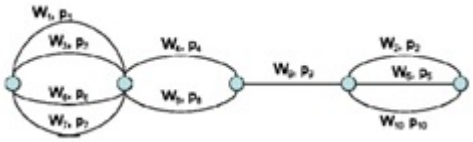

Fig. 2. Structure of Confusion Network

For indexing, confusion networks offer a convenient source for expanding the transcript with alternative recognition candidates. Confusion networks are more compact than lattices and they also provide alignment for all the terms in the lattice. With confusion networks, it is easy to rank locally competing terms by their posterior probability and use the information for indexing [13]. Generally, the algorithm for transforming a lattice to a confusion network composed of the following steps:

- Compute the posterior probability for all edges in the lattice
- Pruning: remove all edges with posterior probability below some threshold
- Intra-word clustering: merge together edges corresponding to the same word instance and sum their posterior probabilities
- Inter-word clustering: group different words which compete around the same time interval and have similar phonetic properties to form confusion sets

In this usage, let *D* be a document modeled by a confusion network. We use two pieces of information in the confusion network for each occurrence of a term *t* at position *o*: its posterior probability *Pr(t|o,D)* and rank among competitors *rank(t|o,D)*. Posterior probability tells how confident the recognizer is that the term occurs in the signal at that position. Rank of the term reflects the importance of the term relative to the other alternatives. In retrieval, document with a higher probability and/or rank of a term should be preferred to one with lower values.

The classical vector space model with *tf-idf* weights and cosine distance relevance measure is used for ranking the search results [13]. Normally, term frequency *tf* is the number of times a term occurs in a document. In our case, we need to estimate a value for term frequency based on the posterior probabilities and ranks of each occurrence of the term in the confusion network of a document.

The term frequency is evaluated by summing the posterior probabilities of all of its occurrences in the confusion network. This means, that if the recognizer is confident that the term at a location is correctly recognized (posterior probability close to one), term frequency is added by (close to) one as in the case of indexing error free text documents. Less weight is given to terms with less confidence. Thus, the term frequency of a term t in a document *D*, *tf(t,D)* is defined:

$$tf(t,D) = \sum_{i=1}^{|occ(t,D)|} \Pr(t|o_i, D)$$

The inverse document frequency *idf* indicates the relative importance of a term in the corpus. Traditionally, *idf* is a function of the number of documents in the collection the term occurs in. In our case, we count the number of confusion networks that the term occurs in at any position. In other words, term occurrence *o(t,D)* was estimated by:

$$o(t,D) = \begin{cases} 1, & if\ tf(t,D) > 0 \\ 0, & otherwise \end{cases}$$

Now, the inverse document frequency for a term *t* is

$$idf(t) = \log(N \Big/ \sum_{D} o(t,D))$$

where *N* is the number of documents in the collection.

In the equation for *o(t,D)*, the value of *tf(t,D)* could also be thresholded by using a value greater than zero to eliminate the effect of terms with low estimated frequency.

4. EXPERIMENTS

4.1.Evaluation Techniques

We use to evaluate the ASR performance the standard word error rate (WER) as our metric. For evaluating indexing system performance we use precision and recall with respect to manual transcriptions. Let *Correct(q)* be the number of times the query *q* is found correctly, *Answer(q)* be the number of answers to the query *q*, and *Reference(q)* be the number of times *q* is found in the reference.

$$Precision(q) = \frac{Correct(q)}{Answer(q)}$$

$$Recall(q) = \frac{Correct(q)}{Reference(q)}$$

In addition to individual precision-recall values we also compute the F-measure defined as

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

And report the maximum F-measure (maxF) to summarize the information in a precision-recall curve.

4.2.Experimental Step

We will perform three steps in our approach. In the first we execute a speech recognition system and we compute the N-best result. There are top N results with maximum a posterior probability. Generally, the top 1 is the result of one best. Here, letting W_1 and W_2 range over the N-best hypothesis output by speech recognizer, then the real result W_c called center hypothesis is found as follow.

$$W_c = arg \min_{i=1,N} \sum_{k=1}^{N} P\left(W_1^{(k)}|A\right) WE\left(W_2^{(i)}, W_1^{(k)}\right)$$

Here, *WE(…)* means the edit distance of two different strings.

During the second phase we generate the word confusion network using the output of previous step then we will prune with different percentiles. And we apply a minimum error training procedure at the goal of weight optimization of the log-linear model. Notice that a separate optimization will perform for each N-best and CN condition.

Finally, we used a standard large vocabulary continuous speech recognition system for generating 3-gram ASR lattices and PSPL lattices. The 3-gram language model used for decoding is trained on a large amount of text data selected based on frequency in the training data. The acoustic model is trained on a variety of wide-band speech and it is a standard clustered tri-phone, 3-states-per-phone model.

5. CONCLUSION

In the field of automatic speech recognition (ASR), existing techniques are mostly based on the recognition of large vocabulary, they offer very good results on structured data, but are still far from being able to handle so successfully for spontaneous speeches and interviews.

In this context, we attempted to propose a hybrid approach combining indexing techniques with lattice and word for spoken document. Our objective is:

- Build an index structure for easier navigation and updated databases
- To adapt the structure to the problem incrementally.
- Reduce the time and complexity of query search of speaker model (document spoken).

Many other objectives of the indexing of multimedia documents are the subject of our research for things like: (video, audio…). As perspectives to our work, we are interested in defining methods to evaluate the performance of indexing approaches and hierarchical classification.

REFERENCES

[1] Park, A., Hazen, T.,Glass, J.: Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling. In: Proc ICASSP, Philadelphia, PA, (2005).

[2] Mills, T., Pye, D., Sinclair, D., Wood, K.: A digital photo management system, In Technical ” AT&T Laboratories Cambridge, Cambridge UK., december (2000).

[3] Lin-shan, L., Yi-cheng, P.: Voice-based information retrieval — how far are we from the text-based information retrieval ?, Automatic Speech Recognition & Understanding, pp. 26-43, November (2009).

[4] Mangu, L., Brill, E.: Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Computer Speech And Language, pp.373–400. (2000).

[5] Jones, G.J.F., Foote, J.T.: Retrieving spoken documents by combining multiple index sources. In: Proc. SIGIR 96, pp. 30–38, Z¨urich, August. (1996).

[6] Srinivasan, S., Petkovic, D.: Phonetic confusion matrix based spoken document retrieval. In: Proc of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 81–87. (2000).

[7] Logan, B., Moreno, P., Deshmukh, O.: Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In: Proc of the HLT, (2002).

[8] James, D.A., Young, S.J.: A fast lattice-based approach to vocabulary independent wordspotting. In: Proc ICASSP. (1994).

[9] Saraclar, M., Sproat, R.: Lattice-based search for spokenutterance retrieval. In: Proc HLT'2004, Boston, (2004).

[10] Mamou, J., D. Carmel, D., Hoory, R.: Spoken document retrieval from call-center conversations. In: Proc of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 51–58, New York, NY, USA, ACM Press. (2006).

[11] Chelba, C., Acero, A.: Position specific posterior lattices for indexing speech. In: Proc of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05). Association for Computational Linguistics, Ann, Arbor, Michigan, pp. 443–450. (2005).

[12] Bendib, I., Laouar, M.R.: Approaches for the Detection of the Keywords in Spoken Documents Application for the Field of E-Libraries. In: Proc of the 19th International Conference on Neural Information Processing (ICONIP2012), pp.196-203. (2012).

[13] Turunen, V.T., Kurimo, M.: Indexing confusion networks for morph-based spoken document retrieval. In: Proc. SIGIR ’07. New York, NY, USA: ACM, pp. 631–638. (2007).