Combining Word and Phonetic-Code Representations for Spoken Document Retrieval

Alejandro Reyes-Barragán[1], Manuel Montes-y-Gómez[1,2] and Luis Villaseñor-Pineda[1]

[1] Laboratory of Language Technologies,
National Institute of Astrophysics, Optics and Electronics (INAOE),
Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico.
{alejandroreyes,mmontesg,villasen}@inaoep.mx

[2] Department of Computer and Information Sciences,
The University of Alabama at Birmingham (UAB),
1300 University Boulevard, Birmingham, Alabama, USA.

Abstract. The traditional approach for spoken document retrieval (SDR) uses an automatic speech recognizer (ASR) in combination with a word-based information retrieval method. This approach has only showed limited accuracy, partially because ASR systems tend to produce transcriptions of spontaneous speech with significant word error rate. In order to overcome such limitation we propose a method which uses word and phonetic-code representations in collaboration. The idea of this combination is to reduce the impact of transcription errors in the processing of some (presumably complex) queries by representing words with similar pronunciations through the same phonetic code. Experimental results on the CLEF-CLSR-2007 corpus are encouraging; the proposed hybrid method improved the mean average precision and the number of retrieved relevant documents from the traditional word-based approach by 3% and 7% respectively.

1 Introduction

The large amount of information existing in spoken form, such as TV and radio broadcasts, recordings of meetings, lectures and telephone conversations, has motivated the development of new technologies for its searching and browsing. Particularly, spoken document retrieval (SDR) refers to the task of finding segments from recorded speech that are relevant to a user's information need [1].

The traditional approach for SDR consists in a simple concatenation of an automatic speech recognition (ASR) system with a standard word-based retrieval method [2]. The main inconvenience of this approach is that it greatly depends on the accuracy of the recognition output. It is well known that recognition errors usually degrade the effectiveness of a SDR system, and that, unfortunately, current ASR methods have word error rates that vary from 20% to 40% in accordance to the kind of discourse.

With the aim of reducing the impact of recognition errors on the retrieval performance, we investigated the helpfulness of using phonetic codifications[1] for representing documents' content. The idea of using phonetic codifications on this task was motivated by two facts. On the one hand, transcriptions errors are not randomly generated; words/phases are commonly substituted by others with similar pronunciation. For instance, the speech utterance "Unix Sun Workstation" would be incorrectly transcribed into "unique set some workstation". On the other hand, phonetic codifications allow characterizing phonetically similar words through the same code. For instance, using Soundex codes, the words "unique" and "Unix" are both represented by the code U52000, whereas the words "some" and "sun" are represented by S50000.

In this paper we propose a retrieval approach that uses word and phonetic-code based representations in cooperation. In particular, we focus on two main concerns. First, we evaluate the usefulness of different phonetic codifications algorithms, namely, Soundex [3], NYSIIS [4], Phonix [5], DMetaphone [6] and DM [7], and second, we analyze the synergy between word and phonetic-code representations. Our results on the CLEF-CLSR-2007 corpus suggest that NYSIIS codes are the more appropriate, and that the combination of word and phonetic-code representations is relevant for SDR and particularly useful for handling short queries.

The rest of the paper is organized as follows. Section 2 introduces some related work on SDR. It particularly presents the major approaches for handling with transcription errors. Section 3 describes our proposed approach for SDR, using word and phonetic code representations in conjunction. Section 4 presents the experimental results on CLEF-CLSR-2007 corpus. Finally, section 5 shows our conclusions.

2 Related work

Due to the limited accuracy of current speech recognizers, several works on SDR have focused on proposing different methods for reducing the impact of transcription errors on the retrieval performance. In general, these methods are of two types: dependent and independent from the ASR system.

From the first kind, we can distinguish two main methods. The first one considers the transcription of speech utterances into phoneme or syllable sequences instead of word sequences by using a phoneme/syllable recognizer [8, 9, 10]. On the other hand, the second method proposes making use of more than the top-1 transcription hypothesis. Particularly, it considers using the n-best hypotheses or the complete word-lattice used internally by the recognizer [11]. As expected, these methods have the disadvantage of requiring access to the inside of the ASR system.

From the second group, we can also differentiate two main methods. One of them proposes using multiple recognizers [12, 13]. It is supported on the idea that

[1] Phonetic codification methods were initially propose for identifying the variants of personal names and for obtaining a canonical or normalized representation of them [20]. Traditionally, these kinds of methods are considered as a kind of approximate string matching technique.

independently developed recognizers tend to make different kinds of errors, and, therefore, that by combining their outputs it might be possible to recover some of them. The second method from this approach proposes reducing the effect of transcription errors by adding some related extra terms to the queries and/or documents [14, 15]. These extra terms can be found by analyzing the transcribed corpus and locating relevant terms based on co-occurrence. However, it has been shown that it is better to use a parallel written corpus, since transcriptions contain recurrent errors and may cause erroneous words to appear as expansion terms.

Similar to the above method, the one proposed in this paper also aims to tackle recognition errors by expanding documents and queries. However, different to this previous approach, it does not achieve this expansion by including some extra words; instead, it proposes to enrich the representation of documents and queries by adding the phonetic codes from the original terms. The purpose of this alternative representation is to reduce the impact of the transcription errors by characterizing words with similar pronunciations through the same phonetic code.

In addition to the previous difference, the proposed method has the advantage of being more portable; it does not require using any external resource (such as a parallel text collection), and, moreover, some phonetic codifications (e.g., Soundex) may be applied with minimal modifications to languages other than English.

Finally, it is important to mention that in a previous work [16] we proposed using Soundex codes to enrich the representation of transcriptions. However, this paper goes several steps forward. First, it presents the evaluation on the use of five different phonetic codifications, namely, Soundex [3], NYSIIS [4], Phonix [5], DMetaphone [6] and DM [7], and second, it explores different ways to combine word and code representations in order to find a reasonable tradeoff between precision and recall.

3 Proposed method

As we previously mentioned, the proposed approach for SDR relies on the use of an expanded representation of automatic transcriptions which combines words and phonetic codes. The following subsections describe in detail two main issues regarding this approach: one the one hand, how to construct the expanded representation, and, on the other hand, how to use this representation through the retrieval process.

3.1 Constructing the combined representation

The construction of the expanded representation considers the following steps:

1. Remove unimportant tokens from transcriptions. Mainly, we consider eliminating a set of common stop words.
2. Compute the phonetic codification for each word from each transcription using a given codification algorithm. A general description and comparison of the codification algorithms used in our experiments can be found in [17], for further details we refer to [3, 4, 5, 6, 7].

3. Combine transcriptions and their phonetic codifications in order to form the expanded document representations. By this combination documents are represented by a mixed bag of words and phonetic codes. Correspondingly, queries need to be represented by their words and phonetic codes.

In order to clarify this procedure, Table 1 illustrates the construction of the expanded representation for the transcription segment "…just your early discussions was roll wallenberg's uh any recollection of of uh where he came from and so…", which belongs to the transcription (spoken document) with id=VHF31914-137755.013 from the CLEF CL-SR 2007 corpus.

Table 1. Example of an expanded document representation using Soundex codes

Automatic transcription	…just your early discussions was roll wallenberg uh any recollection of of uh where he came from…
Preprocessed transcription	… early discussions roll wallenberg recollection came …
Phonetic codification	... E64000 D22520 R40000 W45162 R24235 C50000...
Expanded representation	{early, discussions, roll, wallenberg, recollection, came, E64000, D22520, R40000, W45162, R24235, C50000}

3.2 Using the combined representation

Reports on the TREC's SDR track [14, 1] concluded that traditional word-based representations are good enough for SDR; however, they also indicated that this basic representation has difficulties to effectively handle complex queries, such as small queries or queries containing a large number of out-of-vocabulary (OOV) words.

On the other hand, [16] showed that phonetic codes help to improve retrieval recall but, due to the large number of word coalitions they generate, they tend to decrease precision rates.

Based on this previous evidence, we propose not to use the combined representation in all cases, but only to handle complex queries. We consider the following two criteria for determining –presumably– complex queries.

- *By query length*: a complex query has a length shorter than a given specified threshold.
- *By percentage of OOV words*: a complex query has a percentage of OOV words greater than a given specified threshold.

Figure 1 shows the general architecture of the proposed method. As noticed, it uses two different indexes, one based on words and other on the combination of words and phonetic codes. Both indexes are built offline using the whole document collection. It also includes a module for the online analysis of queries, which allows selecting presumably complex queries that require to be phonetically codified. Finally, it considers a retrieval module which uses the word index or the combined index depending on the form of the given question.

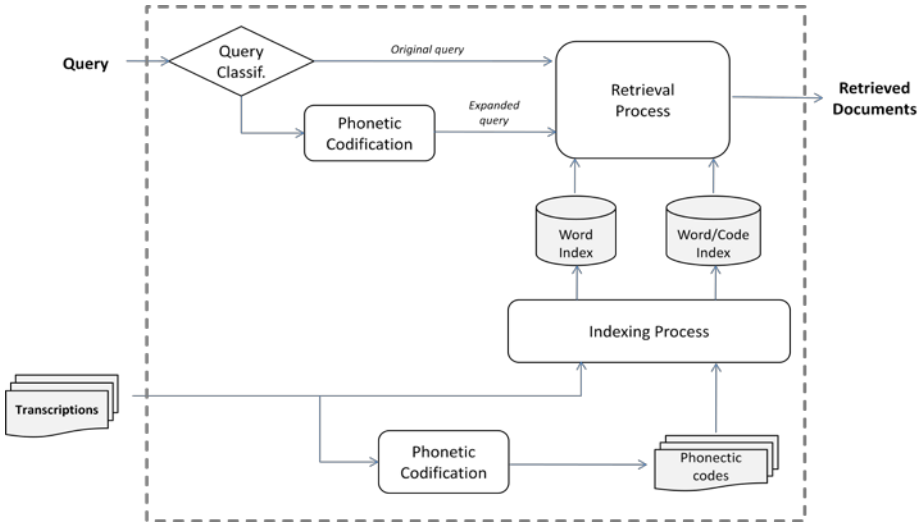

Figure 1. Architecture of the proposed method

4 Evaluation

4.1 Experimental Setup

This section presents some experiments for evaluating the usefulness of the proposed representation. In all experiments, we used the training dataset from the CLEF CL-SR 2007 task [18]. This dataset includes 8,104 transcriptions of English interviews as well as a set of 63 queries.

It is important to mention that for each interview this collection provide three automatic transcriptions (having different word error rates) as well as some sets of automatically and manually extracted keywords. However, in order to get our experiment closer to a real scenario, we decided not to use any set of keywords and to consider only one automatic transcription, namely, the ASR06 with 25% word error rate (WER).

In all the experiments, indexing and retrieval was done by means of the Lemur search engine [19], which was configured to run as a traditional vector space model with tf×idf weights.

On the other hand, the evaluation was carried out using the *MAP* (mean average precision) and the number of relevant retrieved documents (*RelRET*). Both measures were calculated at the first 1000 retrieval results. In particular, the *MAP* is defined as follows:

$$MAP = \frac{1}{|Q|}\sum_{\forall q \in Q} \frac{\sum_{r=1}^{N}\big(P(r) \times rel(r)\big)}{\text{number of relevant document}} \quad (1)$$

where Q is the set of test queries, N is the number of retrieved documents, r indicates the rank of a document, *rel*() is a binary function on the relevance the document at a given rank, and *P*() is the precision at a given cut-off rank.

4.2 Experiments

Experiment 1: assessing different phonetic codifications

The goal of our first experiment was to evaluate the usefulness of several phonetic codifications in the task of SDR. Particularly, we considered the following five codifications: Soundex [3], NYSIIS [4], Phonix [5], DMetaphone [6] and DM [7].

Table 2 shows the retrieval results achieved by these codifications by themselves; that is, these results were obtained by representing documents and queries exclusively by their phonetic codes. This table also shows the results corresponding to the traditional word-based indexing, which is our main baseline.

As expected, due to the generalization caused by the phonetic codifications, their results were worse than those achieved by the word-based indexing. In particular, word-based indexing improved by 5% the *MAP* obtained by the NYSIIS-based representation, which turned out to be the best phonetic-code representation.

An important finding was the number of relevant retrieved documents obtained by the NYSIIS-based representation. It got 1734 relevant documents for the 63 queries, outperforming by almost 10% the result from the word-based indexing. In addition, we noticed that these two representations (using words and NYSIIS-codes) are complementary, since they together may get 1820 relevant documents, and, therefore, they are good candidates for being combined.

Table 2. Results achieved by a phonetic-code-based indexing

		Indexed by				
	Words	Soundex codes	NYSIIS codes	Phonix codes	DMetaphone codes	DM Codes
MAP	0.062	0.051	0.059	0.047	0.037	0.038
RelRET	1578	1529	1734	1539	1567	1483

Experiment 2: using the combined representation for all queries

As suggested by the previous results, we evaluated the effectiveness of applying a combined representation of words and NYSIIS-codes for handling all queries. Table 3 shows the results from this experiment.

Table 3. Results from the combined representation (used for handling all queries)

	MAP	*RelRET*
Words	0.062	1578
Words + NYSIIS codes	0.063	1701
% of improvement over word-indexing	1.6%	7.8%

The obtained results demonstrate the potential of the combined representation, which outperformed the *MAP* and *RelRET* of the traditional approach by 1.6% and 7.8% respectively. However, a deeper analysis showed us that the combined

representation produced worse results than the word-based representation in more than a third part of the queries.

Experiment 3: handling complex queries with the combined representation

The goal of this experiment was to validate the proposed method (refer to Section 3.2), which suggests not to use the combined representation in all cases, but only to handle complex queries. In particular, through this experiment we aimed to evaluate our two different criteria for selecting complex queries.

Table 4 shows the results from this experiment. The first two rows correspond to baseline results: word-based and combined representations. Then, there are the results achieved by applying the proposed combined representation to handle queries of length less than a given threshold. We used three different thresholds, 8, 11 and 14, which correspond to the average minus a standard deviation, the average and the average plus a standard deviation of the lengths from all training queries. Finally, the last three rows show the results obtained by using the proposed combined representation to handle queries having a percentage of OOV words greater than a given specified threshold. We used three different thresholds, 7%, 20% and 33%, which correspond to the average minus a standard deviation, the average and the average plus a standard deviation of the percentage of OOV words from all training queries.

The results from Table 4 once again indicate that using a combined representation is a better alternative than using the traditional word-based indexing. In particular, best results were obtained when the combined representation was used to manage short queries with length lesser than the average length. Using this configuration, the baseline *MAP* and *RelRET* were outperformed by 3.2% and 6.9% respectively.

One important conclusion from this experiment is that the selective use of the combined representation did not show a great advantage over its arbitrary usage, which may point to the necessity of better criteria for evaluating queries complexity.

Table 4. Results from the combined representation (used for handling only complex queries)

	MAP	*RelRET*
Words	0.062	1578
Words + NYSIIS codes (used in all queries)	0.063	1701
Query length ≤ 8	0.062	1670
Query length ≤ 11	**0.064**	**1687**
Query length ≤ 14	0.064	1686
% of OOV words ≤ 7%	0.061	1660
% of OOV words ≤ 20%	0.063	1676
% of OOV words ≤ 33%	0.063	1674

5 Conclusions

In this paper we have proposed a retrieval method specially suited for SDR. This method relies on the idea of using word and phonetic-code based representations in collaboration in order to tackle the effects caused by the transcription errors.

One important contribution of this paper is the evaluation of the usefulness of five different phonetic codifications. Regarding this aspect, our results indicate that NYSIIS is the best phonetic codification for the SDR task. However, they also suggest that phonetic-code-based representations must be used in conjunction with traditional word-based indexing in order to be effective.

The second contribution of this paper is the analysis of the synergy between word and phonetic-code representations. Our results in that direction indicate that the combination of word and phonetic-code representations is relevant for SDR, since using this combination it was possible to outperform the baseline *MAP* and *RelRET* results by 3.2% and 6.9% respectively. Although the combined representation appeared to be more useful for handling short queries, the experimental results suggest that the selective use of the combined representation is not clearly superior to its arbitrary usage.

As future work we plan to explore the usage of the proposed combined representation at character n-gram level. In this way, we think it will be possible to carry away the word segmentation imposed by the ASR process, and, therefore, it will be easier to tackle the problems of word insertions and deletions.

Acknowledgments. This work was done under partial support of CONACYT (project grant CB-2008-106013-Y, and scholarship 204467). We would also like to thank the CLEF organizing committee for the resources provided.

References

1. Garofolo, John S., Auzanne, Cedric G. P. and Voorhees, Ellen M. The TREC Spoken Document Retrieval Track: A Success Story. s.l. : NIST, 1999. pp. 107 - 129. Special publication 500-246.
2. Comas, Pere R. and Turmo, Jordi. Spoken Document Retrieval Based on Approximated Sequence Alignment. Brno, Czech Republic : Springer-Verlag, 2008. Vol. 5246, pp. 285 - 292.
3. Odell, K. M. and Russell, R. C. Soundex phonetic comparison system. [U.S. Patents 1261167 (1918) and 1435663 (1922)].
4. Taft, R. L. (1970), Name Search Techniques, Albany, New York: New York State Identification and Intelligence System. Technical Report, State of New York.
5. Gadd, T. PHONIX: The algorithm. Program: automated library and information systems. 1990, págs. 363-366.
6. Philips, L. The double-metaphone search algorithm. s.l. : C/C++ User's Journal, 2000. Vol. 18, 6.
7. Mokotoff, Gary and Sallyann Amdur Sack, Where once we walked: a guide to the Jewish communities destroyed in the Holocaust, Teaneck, N.J.: Avotaynu. 1991

8. Whittaker, E. W. D., Van Thong, J. M. and Moreno, P. J. Vocabulary Independent Speech Recognition Using Particles. Trento, Italy : s.n., 2001.
9. Siegler, M. Integration of continuous speech recognition and information retrieval for mutually optimal performance. *Ph.D. dissertation.* Carnegie Mellon : Carnegie Mellon University, 1999.
10. Ng, C, Wilkinson, R and Zobel, J. Experiments in spoken document retrieval using phoneme N-grams. Amsterdam, The Netherlands. : Elsevier Science Publishers B. V, September 2000. Vol. 32, 1-2, pp. 61-77.
11. Zhang, Lei, et al. Topic indexing of spoken documents based on optimized N-best approach. Shanghai : s.n., 20-22 Nov. 2009. Vol. 4, pp. 302 - 305.
12. Siegler, M., et al. Experiments in Spoken Document Retrieval at CMU. Gaithersburg, MD, USA : National Institute for Standards and Technology, 1997. NIST-SP 500-240..
13. Nishizaki, Hiromitsu and Nakagawa, Seiichi. Japanese spoken document retrieval considering OOV keywords using LVCSR system with OOV detection processing. San Diego, California : Morgan Kaufmann Publishers Inc, 2002. pp. 157 - 164.
14. Allan, James. Robust techniques for organizing and retrieving spoken documents. New York, NY, United States : Hindawi Publishing Corp, January 2003. Vol. 2003, pp. 103 - 114.
15. Wang, J. and Oard, D. W. CLEF-2005 CL-SR at Maryland: Document and Query Expansion using Side Collections and Thesauri. Vienna, Austria : s.n., September 23, 2005. pp. 744 - 759.
16. Reyes-Barragán, Manuel Alejandro, Villaseñor-Pineda, Luis and Montes-y-Gómez, Manuel. A Soundex-based Approach for Spoken Document Retrieval. [ed.] Springer. México : Springer Berlin, 2008. Vol. 5317, pp. 204-211.
17. Christen, Peter. A Comparison of Personal Name Matching Techniques and Practical Issues. Proceedings of the Sixth IEEE International Conference on Data Mining. September, 2006.
18. Pecina, Pavel, et al. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. [ed.] In Carol Peters et al. Budapest, Hungary : Springer-Verlag, 2008. pp. 674 - 686.
19. Ogilvie, Paul and Callan, Jamie. Experiments Using the Lemur Toolkit. 2002.
20. Gálvez, C. Identificación de Nombres Personales por Medio de Sistemas de Codificación Fonética. Encontros Bibli. Florianópolis, Santa Catarina, Brasil : s.n., 2006. Vol. 11, 22, pp. 105 - 116.