A Soundex-based Approach for Spoken Document Retrieval

M. Alejandro Reyes-Barragán, Luis Villaseñor-Pineda
and Manuel Montes-y-Gómez

Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica, México.
{alejandroreyes, villasen, mmontesg}@inaoep.mx

Abstract. Current storage and processing facilities have caused the emergence of many multimedia repositories and, consequently, they have also triggered the necessity of new approaches for information retrieval. In particular, spoken document retrieval is a very complex task since existing speech recognition systems tend to generate several transcription errors (such as word substitutions, insertions and deletions). In order to deal with these errors, this paper proposes an enriched document representation based on a phonetic codification of the automatic transcriptions. This representation aims to reduce the impact of the transcription errors by representing words with similar pronunciations through the same phonetic code. Experimental results on the CL-SR corpus from the CLEF 2007 (which includes 33 test topics and 8,104 English interviews) are encouraging; our method achieved a mean average precision of 0.0795, outperforming all except one of the evaluated systems at this forum.

1 Introduction

Nowadays, thanks to the falling price of storage media and the rising capacities for information generation, the amount of available multimedia repositories is increasing. In such a situation, it is clear that new retrieval methods are required to search through all this information.

In particular, this paper focuses on the task of *spoken document retrieval* (SDR), which consists of searching relevant information for general user queries from a collection of automatic transcriptions of speech. In this case, the transcriptions may come from many different kinds of audio sources, such as radio and television news programs, political discourses, scientific conferences, business meetings, and interviews.

The conventional approach for SDR integrates speech recognition and information retrieval technologies. It first applies an automatic speech recognition (ASR) process to generate text transcriptions from speech (i.e., the spoken documents), and then it makes use of traditional –textual– information retrieval (IR) techniques to search for the desired information.

The main drawback of this approach is that it greatly depends on the quality of the generated transcriptions, which are far from being perfect. Current ASR methods tend to generate several transcription errors (such as word substitutions, insertions and deletions), producing word error rates that vary from 20% to 40% in accordance to the kind of discourse. Particularly, most of these errors are related to the treatment of out-of-vocabulary words, which in the majority of cases are substituted for a phonetically similar –known– word.

Surprisingly, in spite of the frequent transcription errors, the conventional approach for SDR has achieved more or less satisfactory results. As Allan described in [1], it seems that the impact of the transcription errors is minimal when the queries and/or spoken documents are sufficiently large. Nevertheless, for other kind of scenarios (for instance, those considering short queries and/or spoken documents), these errors have a strong influence on the retrieval results. In these cases, the substitution or elimination of some words may really alter the retrieval process. For example, if the speech utterance "Unix Sun Workstation" is incorrectly transcribed by the ASR process as "unique set some workstation", then it would be impossible to retrieve this phrase by the query "Unix".

Most recent works on SDR have paid no attention to the transcription errors and have mainly concentrated on evaluating the usefulness of IR techniques such as query expansion, relevance feedback and information fusion [2, 3, 4, 5]. In contrast with these works, in this paper we propose a new document representation based on a *phonetic codification* of the automatic transcriptions. The purpose of this representation is to reduce the impact of the transcription errors by characterizing words with similar pronunciations through the same phonetic code. In particular, we propose using the *Soundex codes* [6] to enrich the representation of transcriptions. This way, the example transcription phrase "unique set some workstation" will be also represented by the codes U52000 S30000 S50000 W62300, allowing its retrieval by phonetically similar words such as Unix (with code U52000) or sun (with code S50000).

It is important to mention that the idea of using a phonetic codification for indexing automatic transcriptions of speech is not completely new. There was some attempt to use the Soundex codes for indexing names with the aim of finding all their pronunciation variants [7]. Continuing this idea, in this paper we extend this early approach by applying the phonetic codification indiscriminately to all transcription words. To our knowledge, this is the first time that the Soundex codification has been used and evaluated in the task of SDR.

The rest of the paper is organized as follows. Section 2 introduces the task of SDR and presents the evaluation results from the Cross-Language Speech Retrieval Track of the 2007 Cross-Language Evaluation Forum. Section 3 describes the proposed enriched representation for transcriptions as well as the Soundex codification algorithm. Section 4 presents the evaluation results. Finally, Section 5 gives our conclusions and describes some ideas for future work.

2 Spoken Document Retrieval

The task of SDR consists of searching relevant information for general user queries from a collection of automatic transcriptions of speech. Research in this task has been mainly fostered by two international evaluation conferences, initially by the TREC[1] [8, 9] and more recently by the CLEF[2] [10, 11]. The first one used news recordings, whereas the second considers a collection of spontaneous conversational speeches.

In this paper, we use the data from the Cross-Language Speech Retrieval Track of CLEF 2007 (hereafter referred as CL-SR 2007). This corpus included 8,104 transcriptions of English interviews as well as 96 query topics, 63 for tuning and validation, and 33 for evaluation.

It is important to mention that for each interview there were provided three different automatic transcriptions, each of them having a different word error rate (WER): ASR03 with a WER of 44%, ASR04 with a WER of 38%, and ASR06 with a WER of 25%. In addition, each interview transcription was annotated using two different sets of automatically extracted keywords (AK1 and AK2).

Table 1 shows the evaluation results from this track (more details can be consulted in [11]). From these results, it is clear that SDR is a very complex task; the MAP (mean average precision calculated at the first 100 documents) scores are very low, much lower than those from traditional textual IR (which are around 0.4).

Table 1. Evaluation results from the English monolingual task of CL-SR 2007

Participating Team	***Used Information***	***MAP***
University of Ottawa [3]	AK1,AK2,ASR04	0.0855
Dublin City University [2]	AK1,AK2,ASR06	0.0787
Brown University [4]	AK1,AK2,ASR06	0.0785
University of Chicago [5]	AK1,AK2,ASR06	0.0571
University of Amsterdam [12]	AK2,ASR06	0.0444

The complexity of SDR is due to several factors. First, the automatic transcriptions are not perfect, that is, they contain many errors such as word substitutions, insertions and deletions. Second, the used vocabulary tends to be smaller than that from written documents. Third, the indexing units are commonly speech segments rather than complete documents. Regardless these additional complexities, the participating teams at the CL-SR 2007 mainly focused on applying traditional IR techniques. They used different weighting schemes [3], several query expansion and relevance feedback techniques [2, 3, 4], as well as some information fusion approaches [3]. In other words, none of them did something special for tackling the transcription errors. Moreover, different to our proposal, none of them took into consideration any kind of phonetic information from transcriptions.

[1] The Text REtrieval Conference, http://trec.nist.gov.

[2] The Cross Language Evaluation Forum, http://www.clef-campaign.org.

3 An Enriched Representation for Spoken Documents

As we previously mentioned, the proposed approach for SDR relies on a *phonetically enriched representation* of the automatic transcriptions (spoken documents). This representation aims to reduce the impact of the transcription errors by characterizing words with similar pronunciations through the same phonetic code.

The construction of the enriched representations considers the following actions:

1. Compute the phonetic codification for each transcription using the Soundex algorithm (refer to Section 3.1).
2. Combine transcriptions and their phonetic codifications in order to form the enriched document representations. This way, each document is represented by a mixed bag of words and phonetic codes.
3. Remove unimportant tokens from the new document representations. In this case, we eliminate a common list of stopwords as well as the most frequent phonetic codes (refer to Section 4).

In order to clarify this procedure, Table 2 illustrates the construction of the enriched 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 belong to the spoken document with id=VHF31914-137755.013 from the CL-SR 2007 corpus.

Table 2. Example of an enriched document representation

Automatic transcription	…just your early discussions was roll wallenberg uh any recollection of of uh where he came from…
Phonetic codification	... J23000 Y60000 E64000 D22520 W20000 R40000 W45162 U00000 A50000 R24235 O10000 O10000 U00000 W60000 H00000 C50000 F65000 ...
Enriched representation	{just, early, discussions, roll, wallenberg, recollection, came, E64000, D22520, R40000, W45162, R24235}

It is important to mention that, in order to take advantage of the proposed representation in the retrieval process, it is also necessary to construct the enriched representation of queries. For that purpose we apply the same procedure than the one used for constructing the enriched representations of transcriptions. Table 3 shows the enriched representation of a query from the CL-SR 2007 corpus.

Table 3. Example of an enriched query representation

Original query	Eyewitness accounts that describe the personalities and actions of Raoul Wallenberg and Adolf Eichmann
Phonetic codification	E35200 A25320 T30000 D26100 T00000 P62543 A53000 A23520 O10000 R40000 W45162 A53000 A34100 E25500
Enriched representation	{eyewitness, accounts, personalities, actions, raoul, wallenberg, adolf, eichmann, E35200, A25320, P62543, A23520, R40000, W45162, A34100, E25500}

From this example, it is interesting to notice that the usage of the phonetic codifications allows improving the matching between transcription and query, since the words “roll” and “Raoul” were both represented by the same phonetic code (R40000). For this particular case, we presume that the word “roll” was obtained from an incorrect transcription of the word “Raoul”.

3.1 The Soundex Codification Algorithm

Phonetic codifications attempt to represent words with similar pronunciations by the same code. Among all existing phonetic codification algorithms, *Soundex* is the most widely known. It was originally proposed for dealing with the problem of having different spelling variations of the same name (e.g., Lewinsky vs. Lewinskey) [6], and since then it has been applied in several database applications for indexing surnames, for instance, it has been used in the U.S. census.

The Soundex algorithm is based on the phonetic classification of human speech sounds (bilabial, labiodental, dental, alveolar, velar, and glottal), which in turn are based on where we put our lips and tongue to make sounds. The algorithm itself is straightforward since it does not require of backtracking or multiple passes over the input word. This algorithm is as follows:

1. Capitalize all letters in the word and drop all punctuation marks.
2. Retain the first letter of the word.
3. Change all occurrence of the following letters to '0' (zero): 'A', E', 'I', 'O', 'U', 'H', 'W', 'Y'.
4. Change letters from the following sets into the given digit:
   - 1 = 'B', 'F', 'P', 'V'
   - 2 = 'C', 'G', 'J', 'K', 'Q', 'S', 'X', 'Z'
   - 3 = 'D','T'
   - 4 = 'L'
   - 5 = 'M','N'
   - 6 = 'R'
5. Remove all pairs of equal digits occurring beside each other from the string resulted after step (4).
6. Remove all zeros from the string that results from step (5)
7. Pad the string resulted from step (6) with trailing zeros and return only the first six positions. The output code will be of the form <uppercase letter> <digit> <digit> <digit> <digit> <digit>.

Using this algorithm, both "Robert" and "Rupert" return the same string "R16300", whereas "Rubin" yields "R15000".

4 Experimental Results

4.1 Experimental Setup

This section presents some experiments that allow evaluating the usefulness of the proposed representation. In all these experiments, we used the data from the CL-SR 2007 task [11]. In particular, we considered the automatic transcription ASR06 (with a word error rate of 25%), and the sets of automatic keywords AK1 and AK2. Besides, we made use of the Indri search engine [13], which provided us with the functionalities for term weighting and query expansion[3].

It is also necessary to mention that, in accordance to the proposed representation, in all experiments we eliminated a set of stopwords as well as a set of high frequency phonetic codes. Mainly, we decided to eliminate the same proportion of both items. In this way, we eliminated 319 stopwords[4], which correspond to 75% of the word items, and 265 phonetic codes, which approximately represent the same percentage from the whole set of used phonetic codes.

On the other hand, the evaluation was carried out using the *MAP* (mean average precision) and *precision@10* measures. These measures are calculated as follows:

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

$$precision@10 = P(r = 10)$$

where Q is the set of test questions, N is the number of retrieved documents, r indicates the rank of a document, $rel()$ is a binary function on the relevance of a given rank, and $P()$ is the precision at a given cut-off rank. This way, the precision@10 indicates the percentage of correct items from the first ten retrieved documents.

4.2 Results

The first experiment aimed to determine the *pertinence of the phonetic codification*. In order to do that, we performed the SDR considering only textual information (the transcriptions along with automatic keywords) as well as using the phonetic codification by itself. Table 4 shows the achieved results.

[3] We applied a blind query expansion approach that incorporated to the original query the ten most frequent terms from the first ten ranked documents.

[4] The list was taken from the IR resources of the Department of Computing Science at the University of Glasgow (www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words)

Table 4. Text-based vs. Phonetic-based retrieval

	SRD using only text (ASR06 + AK1 + AK2)	**SDR using only phonetic codes** (phonetic codification of ASR06)
MAP	0.108	0.081
Precision@10	25.4%	19.4%

This experiment showed that the phonetic-based retrieval could identify an important number of relevant documents, even though its results were inferior to those from the text-based approach. In addition, this experiment allowed us observing the complementarity of both approaches: together they retrieved 2240 relevant documents, nevertheless they only shared 1674, indicating that their results are complementary in 27%.

Based on this last observation, the second experiment attempted to evaluate the *effectiveness of the proposed representation*, which combines textual and phonetic information. For this experiment, the weight of words was defined as the double than the weight of phonetic codes. Table 5 shows the results from this experiment.

Table 5. Combining transcriptions and their phonetic codifications

	Proposed representation	**Improvement over Text-based SDR**
MAP	0.116	+ 6.8%
Precision@10	28.1%	+ 10.6%

From these results, it is possible to conclude that the combination of textual and phonetic information stated by the proposed representation improves the SDR process. Nevertheless, it is necessary to do an extensive analysis in order to determine the kind of transcription errors that were handled by the phonetic codification.

Finally, a third experiment was carried out to compare our proposal against other current state-of-the-art methods evaluated on the CL-SR 2007 test corpus. In this experiment, we use the test corpus consisting of 33 topics. Table 6 shows the results from this experiment. It can be observed that our proposal achieved a MAP of 0.0795, outperforming all except one of the evaluated systems at this forum. However, it is important to comment that our proposal (using a phonetically enriched document representation) could be used in conjunction with any one of these methods, probably leading to a better performance.

Table 6. Comparison of our results against methods from CL-SR 2007

Team	***Used Information***	***MAP***
University of Ottawa	AK1,AK2,ASR04	0.0855
Our proposal	AK1,AK2,ASR06	0.0795
Dublin City University	AK1,AK2,ASR06	0.0787
Brown University	AK1,AK2,ASR06	0.0785
University of Chicago	AK1,AK2,ASR06	0.0571
University of Ámsterdam	AK2,ASR06	0.0444

5 Conclusions

In this paper, we have proposed an *enriched representation for spoken documents* that is specially suited for the IR task. This new representation is based on a phonetic codification of the automatic transcriptions, which aims to reduce the impact of the transcription errors by characterizing words with similar pronunciations through the same phonetic code.

Experimental results on the CL-SR 2007 data set are encouraging; our proposal achieved a MAP of 0.0795, outperforming all except one of the evaluated systems at this forum. In addition, these results demonstrated the usefulness of the phonetic codification as well as its complementarity with the textual information.

It is clear that it is necessary to perform more experiments in order to conclude about the advantages of the proposed representation. In particular, we plan to:

- Consider other phonetic codifications (such as Daitch-Mokotoff, NYSIIS, Phonix, Metaphone, and Double Metaphone) and perform additional experiments for determining the most appropriate one.
- Accomplish a detailed analysis of current results in order to determine the kind of transcription errors that were successfully handled by the phonetic codification. Particularly we are interested in studying the treatment of out-of-vocabulary words.
- Use the proposed representation in conjunction with different SDR methods such as those participating at the CL-SR 2007 task.
- Explore the usage of the proposed representation at character n-gram level. In this way, it is possible to take away the word segmentation imposed by the ASR process, and therefore, it is easier to tackle the problems of word insertions and deletions.

Acknowledgements

This work was done under partial support of CONACYT (project grant 61335 and scholarship 212715). We also like to thank the CLEF organizing committee for the resources provided.

References

1. Allan J. Perspectives on Information Retrieval and Speech. Lecture Notes in Computer Science, Vol. 2273. 2002.
2. Jones G. Zhang K. and Lam-Adesina A. Dublin City University at CLEF 2007: Cross-Language Speech Retrieval (CL-SR) Experiments. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007.
3. Alzghool M. and Inkpek D. Model Fusion for the Cross Language Speech Retrieval Task at CLEF 2007. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007 .

4. Lease M. and Charniak E. Brown at CL-SR'07: Retrieval Conversational Speech in English and Czech. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007.
5. Levow G. University of Chicago at the CLEF 2007 Cross-Language Speech Retrieval Track. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007.
6. Odell, M. K., Russell, R. C. U. S. Patent Numbers 1261167 (1918) and 1435663 (1922). Washington, D.C.: U.S. Patent Office, 1918.
7. Raghavan H. and Allan J. Using Soundex Codes for Indexing Names in ASR documents. In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at Humal Language Technology Conference and North American chapter of Association of Computa Computational Linguistics, pages 22–27, Boston, MA, USA, 2004.
8. Voorhees, E., Garofolo, J., and Jones, K. The TREC-6 Spoken Document Retrieval Track. Proceedings of the Sixth Text Retrieval Conference (TREC-6). Gaithersburg, Maryland. November 19–21, 1997.
9. Garafolo, J. S., Auzanne, C. G. P., and Voorhees, E. The TREC Spoken Document Retrieval Track: A Success Story. In Proceedings of the RIAO 2000 Conference: Content-Based Multimedia Information Access. Paris, France, 2000.
10. White R., Oard D., Jones G., Soergel D. Huang X. Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005). Vienna, Austria, 21-23 September 2005.
11. Pecina P., Hoffmannová P. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007.
12. Huurnink B. The University of Amsterdam at the CLEF Cross Language Speech Retrieval Track 2007. Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007). Budapest, Hungary, 19-21 September 2007.
13. Strohman T. Metzler D. Turtle H. and Croft W.B. Indri: A Language-Model based Search Engine for Complex Queries. Proceedings of the International Conference on Intelligence Analysis, McLean, VA. May 2-6, 2005.