A SYSTEM FOR SPEECH DRIVEN INFORMATION RETRIEVAL

César González-Ferreras, Valentín Cardeñoso-Payo

Departamento de Informática, Universidad de Valladolid
47011 Valladolid, Spain
{cesargf,valen}@infor.uva.es

ABSTRACT

In this paper we present a system that allows users to search information in a document collection using a spoken query. The system is based on a speech recognizer and on an information retrieval engine. The system works for Spanish language. We evaluated the system using CLEF'01 test set, extended to include spoken queries. We proposed an adaptation of vocabulary and language model, to reduce the out of vocabulary word problem. In order to reduce errors caused by words in a foreign language, we expanded our pronunciation lexicon to include the pronunciation of English words. Experiments showed a relative gain in retrieval precision of 6.34%, a relative reduction in OOV word rate of 24.71% and a relative reduction in WER of 10.87%.

Index Terms— speech recognition, information retrieval, speech driven information retrieval, language model adaptation, foreign words modeling

1. INTRODUCTION

During the last years the use of Internet is increasing. People access the web both at the office and at home, using a personal computer. Recently, there is a growing interest in providing access to the web in mobile environments. There is a proliferation of mobile devices which allow Internet access everywhere and anytime. However, their interface is limited by small screens and input devices (keypad or stylus). The use of speech in that scenario can provide a more usable interaction.

Different approaches have been proposed to allow access to web contents using speech. The main difference with traditional spoken dialogue systems is that textual information lacks the required structure. The limitations of the speech channel are also a problem, because it is not possible to send much information over it. One solution is to extend an existing web browser using speech [1]. Other approaches are based on the automatic generation of dialogue systems for specific web content [2]. Finally, speech driven information retrieval systems use speech as the input to an information retrieval (IR) engine, providing a natural solution to overcome the major limitations of the speech channel.

The objective of speech driven information retrieval is to search information in a textual document collection using a spoken query. A related area of research is spoken document retrieval (SDR), whose aim is inverse: to index and retrieve relevant items from a collection of spoken audio recordings in response to a text query. A lot of effort has been invested in SDR and good results have been obtained [3]. However, speech driven information retrieval is a more difficult task, because spoken queries contain less redundancy to overcome speech recognition errors.

First experiments in speech driven information retrieval were reported in [4], for English language. Results showed that longer queries are more robust to errors than shorter ones. A system designed for mobile devices was presented in [5], for Chinese language. As a relevant conclusion, retrieval precision on mobile devices with high quality microphones was acceptable, although the performance over cellular phones was not satisfactory. Experiments in Japanese were reported in [6]. Better results were obtained using the target document collection to train the LM, and using bigger vocabulary size. Techniques for combining the output of multiple LVCSRs using SVM learning were also experimented with the same test collection [7].

In this paper we describe a speech driven information retrieval system. First, a large vocabulary continuous speech recognizer (LVCSR) transcribes the spoken query. Then, the IR engine retrieves the documents relevant to the given query. The system works for Spanish language. To evaluate the performance of the system we used a standard IR test suite and recorded 10 speakers reading the queries. The experiments showed three types of errors that affect retrieval performance: errors caused by out of vocabulary (OOV) words, errors produced by words in a foreign language and regular speech recognition errors. To reduce the OOV words problem, we proposed a two-pass strategy: in the first pass, documents relevant to the query were retrieved and used to adapt the vocabulary and language model (LM); the adapted models were used in the second pass to obtain the final result. To overcome the foreign words problem, we added the pronunciation

This work has been partially supported by *Consejería de Educación de la Junta de Castilla y León* under project number VA053A05.

of English words to our pronunciation lexicon. The combination of both improvements provided a relative gain of 6.34% in retrieval precision, a relative reduction of 24.71% in OOV word rate and a relative reduction of 10.87% in WER, compared with the baseline system.

The structure of the paper is as follows: section 2 describes the system in detail; in section 3 the results of the experiments are reported; in section 4 we discuss about system performance; section 5 presents conclusions and future work.

2. SYSTEM OVERVIEW

The system is based on a LVCSR and on an IR engine. First, the speech recognizer transcribes the spoken query. Then, the output is used by the IR engine to retrieve the most relevant documents for the given query. In the following sections we describe in detail both speech recognition and information retrieval systems.

2.1. Speech recognition

We used SONIC, the University of Colorado large vocabulary continuous speech recognition system [8]. It implements a two-pass search strategy: the first pass consists of a token-passing Viterbi search and the second pass uses an A* algorithm.

Acoustic models were continuous density hidden Markov models. A standard 39-dimensional feature vector was used for feature representation (12 MFCCs and normalized energy, along with the first and second order derivatives). Gender independent triphone acoustic models were trained using Albayzin corpus [9] (13,600 sentences read by 304 speakers).

We created a word based trigram using SRILM toolkit [10], with Katz backoff for smoothing. The target document collection was used as training data, because this can result in an adaptation of the LVCSR to the given task and provides better system performance [6]. EFE94 document collection is composed of one year of newswire news (511 Mb) and has 406,762 different words. We used a vocabulary of 60,000 words, that was created selecting the most frequent words found in the documents. The pronunciation lexicon was built using a rule based system for Spanish.

2.2. Information retrieval

We used a modified version of an IR engine developed for Spanish [11]. The system is based on the vector space model and on term frequency inverse document frequency (TF-IDF) weighting scheme. A stop word list was used to remove function words and a stemming algorithm [1] to reduce the dimensionality of the space.

[1] Snowball stemmer (http://snowball.tartarus.org).

The similarity between a document d_i and a query q is calculated as follows:

$$sim(d_i, q) = \sum_{t_r \epsilon q} w_{r,i} \times w_{r,q} \quad (1)$$

$$w_{r,i} = (1 + log(tf_{r,i})) \times log\left(\frac{N}{df_r}\right) \quad (2)$$

$$w_{r,q} = (1 + log(tf_{r,q})) \times log\left(\frac{N}{df_r}\right) \quad (3)$$

where $tf_{r,i}$ represents the frequency of the term t_r in the document d_i; df_r denotes the number of documents in the collection that contain the term t_r; $tf_{r,q}$ is the frequency of the term t_r in the query q; N is the total number of documents in the collection.

To improve retrieval performance we applied pseudo relevance feedback. We used Rocchio's method, which is one of the most popular [12]. First, the original query is used to retrieve a preliminary list of documents. Then, the 2 top ranked documents are assumed to be relevant, and the query is expanded by adding 45 top weighted terms from those documents. Finally, the expanded query is used to generate the final result. Rocchio parameters were experimentally set to: $\alpha = 4$, $\beta = 0.4$ and $\gamma = 0.5$.

3. EXPERIMENTS

We made some experiments to evaluate the performance of the system. First, we describe the experimental set-up and the results of our baseline system. Next, we present a technique to reduce OOV words problem, based on vocabulary and LM adaptation. Then, we describe an extension to our Spanish speech recognizer to include the pronunciation of English words. Finally, results of the final system, which incorporates both improvements, are reported.

3.1. Experimental set-up

CLEF'01 Spanish monolingual test suite was used. Cross-Language Evaluation Forum (CLEF) is an evaluation forum similar to TREC, designed to evaluate IR systems operating on European languages, under standard and comparable conditions [13]. The evaluation set includes a document collection, a set of topics and relevance judgments.

The document collection (EFE94) has 215,738 documents of the year 1994 from EFE newswire agency (511 Mb). There are 49 topics and each of them has three parts: a brief title statement, a one-sentence description and a more complex narrative. To construct the queries we used the description field of each topic (mean length of 16 words). We recorded 10 different speakers reading the queries (5 male and 5 female). Headset microphone was used under office conditions, at 16 bit resolution and 16 kHz sampling frequency.

3.2. Baseline system

In the baseline system we used a one-pass strategy: the spoken query was transcribed by the speech recognizer and the best hypothesis was processed by the IR engine to obtain the list of documents relevant to that query. The same methodology of CLEF was used to evaluate the results [13]. For each query, the 1000 most relevant documents (sorted by relevance) were retrieved, and mean average precision (MAP) was calculated using relevance judgments.

Results of the baseline system are presented in table 1. We report the OOV word rate, the word error rate (WER) of the speech recognizer and the mean average precision. We also report the result obtained with text queries for comparison.

We analyzed the results of each individual query, comparing the MAP of spoken queries with the MAP of textual ones. Queries with a relative loss of more than 25% were considered erroneous. We studied the reasons of the degradation and identified three types of errors:

- **Type I**: errors produced by OOV words.
- **Type II**: errors caused by words in a foreign language.
- **Type III**: regular speech recognition errors.

In table 2 we classified each erroneous query according to the type of error (the total number of queries was 490: 49 queries spoken by 10 speakers).

	OOV	**WER**	**MAP**
Text	—	—	0.4851
Baseline	2.17%	18.4%	0.3893
Adaptation	1.62%	16.4%	0.4111
Foreign words	2.17%	18.4%	0.3867
Final System	1.63%	16.4%	0.4140

Table 1. Performance of different experiments (OOV: out of vocabulary word rate; WER: word error rate; MAP: mean average precision).

	Type I errors	**Type II errors**	**Type III errors**	**Total errors**
Baseline	20	27	47	94
Adaptation	14	29	40	83
Foreign words	20	18	54	92
Final System	13	18	44	75

Table 2. Number of erroneous queries for each type of error, from a total of 490 queries (Type I: caused by an OOV word; Type II: produced by a word in a foreign language; Type III: regular speech recognition error).

3.3. Vocabulary and language model adaptation

To reduce type I errors, we proposed a two-pass strategy, as shown in figure 1. In the first pass, we obtained the 1000 most relevant documents to the given query. We used those documents to adapt the vocabulary and the LM:

- **Vocabulary adaptation**: first, we created a list with every word that appeared in the documents retrieved in the first pass. Then, we added the most frequent words from our general vocabulary until we reached a vocabulary of 60,000 words. The number of words from each source depended on the query (the average number of words from the documents was 27,000).
- **Language model adaptation**: first, we trained a new LM with the documents obtained in the first pass. Then, we interpolated this new LM with the general LM, using the adapted vocabulary. We used linear interpolation with an interpolation coefficient of 0.5.

Results using the adapted models are shown in table 1 and in table 2. Better results were obtained because the documents used for the adaptation had a semantic relation with the query. There was a reduction in type I errors, because of a better vocabulary coverage, and in type III errors, because of better language modeling. As a consequence, there was a reduction in OOV word rate, a reduction in WER and an increase in MAP.

3.4. Inclusion of foreign words pronunciation

Type II errors were caused by words in a foreign language. The most frequent foreign words that appear in Spanish are English ones. So, our objective was to enable our LVCSR to understand Spanish speakers saying English words.

There is a serious limitation in training data, because, to our knowledge, a corpus with English words spoken by Spanish speakers does not exist. For this reason, we employed an approach similar to the one Spanish speakers employ: mapping English phonemes to Spanish ones. We used Spanish acoustic models and developed the mapping manually, based on similar sounds of both languages (using International Phonetic Alphabet as a reference). Then, we included the pronunciation of English words in the pronunciation lexicon (it is useful to keep the Spanish pronunciation because some speakers pronounce English words as if they were Spanish ones). We used the CMU pronouncing dictionary[2] to identify English words and to obtain their pronunciation: for each word in our vocabulary that appeared in CMU dictionary, we added the alternate English pronunciation. Our vocabulary had 60,000 words, and we added the pronunciation of 8,891 English words in the pronunciation lexicon.

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

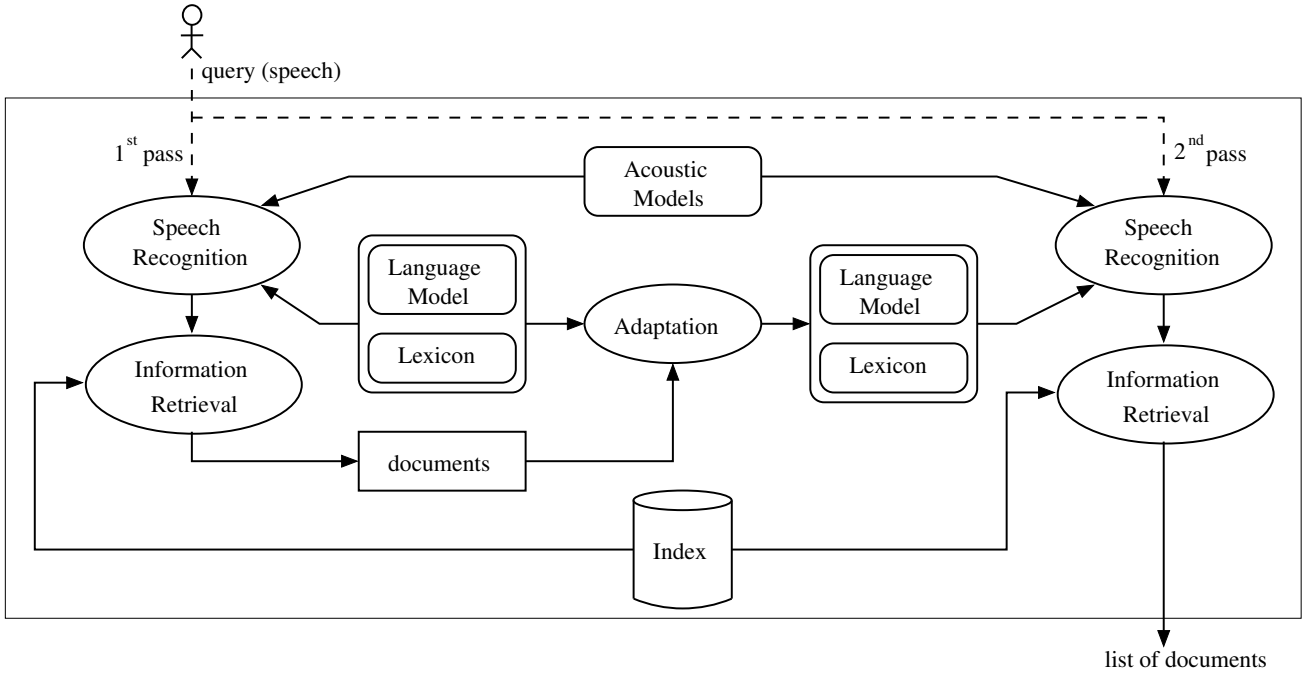

Fig. 1. *Architecture of the system, based on a two-pass strategy.*

One-pass strategy with alternate pronunciations reduced the number of type II errors, however some new type III errors appeared (table 2). As a result, there was no gain in retrieval precision compared with the baseline system (table 1).

3.5. Final system

We combined the two-pass strategy with foreign words modeling. In the first pass, we obtained the 1000 most relevant documents to the query (using the pronunciation lexicon that included English words pronunciation). Then, we adapted the vocabulary and the LM. Next, we expanded the pronunciation lexicon to include English words pronunciation. A different vocabulary was used for each query, so the number of pronunciations of English words added was variable (with an average of 8,640 words).

There was a reduction in the number of erroneous queries, as shown in table 2. There was also an increase in retrieval precision, as shown in table 1 (higher difference in MAP was expected compared with "Adaptation", but variation in average precision across queries is often very high). Finally, comparing the final system with the baseline system, there was a relative gain in MAP of 6.34%, a relative reduction of 24.71% in OOV word rate, and a relative reduction of 10.87% in WER.

4. DISCUSSION

It is not easy to compare our speech driven information retrieval system with other systems in the bibliography. The major difficulty is that each system works in a different language, and thus, it is not possible to use the same test set, which is required for a fair comparison. There is always the possibility to port the system to other language, but this is a costly task, especially for the speech recognizer.

There is a mismatch in the combination of information retrieval and speech recognition technologies. In one hand, information retrieval favors infrequent words (through TF-IDF weighting scheme), because they usually carry more semantic content. On the other hand, speech recognition favors frequent words (higher probability in the LM), because they are more likely to be said. This mismatch is critical with proper nouns, which carry important semantic information, but are a source of errors in speech recognition. We identified two main problems: OOV words and words in a foreign language. The first happens because words with a low frequency are not included in the vocabulary. The second occurs because the speech recognizer is not prepared to recognize words in other languages.

Performance of the final system showed a relative loss in MAP of 14.6%, compared with using textual queries. However, precision loss is not equally distributed among queries: most of the queries did well (small loss of precision) while some queries did badly (high loss of precision). Moreover, each type of error had different impact on precision: type I and type II errors were catastrophic (precision loss of 98.8% on average for type I, and 91.3% for type II), while type III errors produced less degradation in retrieval accuracy (precision loss of 75.0% on average).

5. CONCLUSIONS

In this paper we describe a system for speech driven information retrieval. The motivation of such system is the growing interest in allowing access to web contents in mobile environments. Speech can provide a more user friendly interaction on mobile devices, alone or combined with other input modalities.

System performance was evaluated using a standard IR test suite extended to include spoken queries. Three types of errors were identified: errors produced by OOV words, errors caused by words in a foreign language and regular speech recognition errors. To improve the performance, we proposed a two-pass strategy, which showed effective to reduce OOV word rate and regular speech recognition errors. Moreover, we included pronunciation of English words in our pronunciation lexicon, which reduced errors caused by foreign words. The final system was the combination of both ideas and provided a relative gain in MAP of 6.34%, a relative reduction of 24.71% in OOV word rate, and a relative reduction of 10.87% in WER, compared with the baseline system.

As future work we plan to evaluate the system with real users. We are developing a multimodal dialogue system for a personal digital assistant (PDA) device which allows users to search information using spoken queries. On one hand, the combination of complementary modalities (speech, stylus and small display) can deliver a user friendly interaction, especially in mobile environments. On the other hand, user interaction can provide valuable feedback to improve the retrieval process. Finally, the use of the system in a different environment (PDA) will require an analysis of the impact of acoustic model adaptation on the performance of the speech recognizer.

6. REFERENCES

[1] Bostjan Vesnicer, Janez Zibert, Simon Dobrisek, Nikola Pavesic, and France Mihelic, "A Voice-driven Web Browser for Blind People," in *Eurospeech*, 2003.

[2] César González-Ferreras and Valentín Cadeñoso Payo, "Development and Evaluation of a Spoken Dialog System to Access a Newspaper Web Site," in *Eurospeech*, 2005.

[3] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *TREC-8*, 1999.

[4] James Barnett, Steve Anderson, John Broglio, Mona Singh, R. Hudson, and S. W. Kuo, "Experiments in Spoken Queries for Document Retrieval," in *Eurospeech*, 1997.

[5] Eric Chang, Frank Seide, Helen M. Meng, Zhuoran Chen, Yu Shi, and Yuk-Chi Li, "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 531–541, November 2002.

[6] Atsushi Fujii and Katunobu Itou, "Building a Test Collection for Speech-Driven Web Retrieval," in *Eurospeech*, 2003.

[7] Masahiko Matsushita, Hiromitsu Nishizaki, Seiichi Nakagawa, and Takehito Utsuro, "Keyword Recognition and Extraction by Multiple-LVCSRs with 60,000 Words in Speech-driven WEB Retrieval Task," in *ICSLP*, 2004.

[8] Bryan Pellom and Kadri Hacioglu, "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task," in *ICASSP*, 2003.

[9] Asunción Moreno, Dolors Poch, Antonio Bonafonte, Eduardo Lleida, Joaquim Llisterri, Jose B. Mariño, and Climent Nadeu, "ALBAYZIN Speech Database: Design of the Phonetic Corpus," in *Eurospeech*, 1993.

[10] Andreas Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *ICSLP*, 2002.

[11] Joaquín Adiego, Pablo de la Fuente, Jesús Vegas, and Miguel A. Villarroel, "System for Compressing and Retrieving Structured Documents," *UPGRADE*, vol. 3, no. 3, pp. 62–69, June 2002.

[12] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science*, vol. 41, no. 4, pp. 288–297, 1990.

[13] Martin Braschler and Carol Peters, "CLEF Methodology and Metrics," in *Workshop of the Cross-Language Evaluation Forum (CLEF)*, 2001.