SEARCH IN SPEECH RECORDS

Michal FAPŠO
Master Degree Programme (1), FIT BUT
E-mail: xfapso00@stud.fit.vutbr.cz

Supervised by: Igor Szöke
E-mail: szoke@fit.vutbr.cz

ABSTRACT

This paper describes the designed and implemented system for efficient indexing and search in collections of spoken documents. The system uses an LVCSR (Large Vocabulary Continuous Speech Recognition) system and a phoneme recognizer. The outputs of both recognizers are indexed. Then in the searching phase, both indices are be used. This leads to high retrieval performance for common words and to the ability to search for less common words (which are not in the LVCSR dictionary) as well. Last, but not least, the results from the NIST Spoken Term Detection evaluations are reported.

1 INTRODUCTION

It is very likely that today's success of Google in text search will excite interest in searching also other media. Among these, search in speech is probably the most interesting, as most of human-to-human communication is done by this modality. Although there is a plenty of audio records publicly available on the internet, the only information we can directly get about them is the title or summary at best. For example, in a case we are looking for some specific information discussed in an one hour long meeting, we need to spend a lot of time listening to something that is not interesting until we find what we are really looking for. In this and many other situations, a system capable of search in speech records would be of a great help. In general, search in speech is necessary whenever we need to access information from multimedia records. Thus there is plenty of possible applications in call centres, meeting processing, multimedia data mining, security and defence, etc. Unlike search in text, where the systems deal with text data, search in speech data is a more complex process that needs to address the following points:

- conversion of speech to discrete symbols that can be indexed and searched – LVCSR (Large Vocabulary Continuous Speech Recognition) systems and phoneme recognizers can be used.
- accounting for inherent errors of LVCSR and phoneme recognizer – this is usually solved by indexing and search in recognition lattices [1] instead of 1-best string output.
- determining the confidence of a query – in this thesis done by evaluating the likelihood ratio between the path with searched keyword(s) and the optimal path in the lattice [2].
- capability to search for less common words like proper names, technical terms or mispronounced words [2].

Task	LVCSR	Phoneme	LVCSR + Phoneme
BCN	0.6278	0.3571	0.6541
CTS	0.5186	0.2977	0.5235
MTG	0.0463	0.0078	0.0549

Table 1: The Term Weighted Value on all three source types using LVCSR, Phoneme and the fused LVCSR + phoneme fused system

- processing multi-word queries, both quoted (exact sequences of words) and unquoted.
- providing an efficient and fast mechanism to obtain the search results in reasonable time even for huge amounts of data (Section 3).

2 SYSTEM ANATOMY

The system's architecture is similar to Google and other search systems [3]. It consists of the indexer, searcher and sorter modules (for detailed info see [4]). Indexing and searching is based on combination of LVCSR and phoneme recognizers. The system takes advantage of high accuracy of LVCSR recognizer and the robustness in dealing with less common words of a phoneme recognizer. According to our experiments, the best way to fuse these two systems is to search for common words using the LVCSR system's output, and to use the phoneme recognizer's output for OOV (Out Of Vocabulary - a word that is not present in an LVCSR system) words. This way a fused system will take the best of both subsystems and its retrieval performance will be improved.

3 EXPERIMENTS AND RESULTS

The system was evaluated in The NIST Spoken Term Detection Evaluations (STD) [5]. The task was to find all of the occurrences of a specified "term" in a given corpus of speech data. For the STD task, a term is a sequence of one or more words. The corpus contains three different source types for English language: broadcast news (BCN), telephone conversations (CTS) and round-table meetings (MTG). The overall system detection performance is represented by detection error trade-off (DET) curves (probability of misses as function of probability of false alarms with detection threshold as parameter) and measured by the Term Weighted Value (TWV, the highest possible value is 1) [5].
Now let us have a look on columns "LVCSR" and "Phoneme" in the table 1. Here we can see a higher value for the LVCSR based system, but if we compare it's value with the fused system (the column "LVCSR + Phoneme"), we can see that the fused system is better. It is because of less common words which are not in the LVCSR dictionary (OOV). The LVCSR based system is not able to find any OOV, but the phoneme based system does not make any difference in searching for common words or OOVs.

4 CONCLUSION

A scalable speech search system was developed and evaluated in Spoken Term Detection Evaluations (STD) with very good results. The search system was build in a way similar to Google but it had to be adapted to speech indexing and search. Since the system indexes lattices (graph

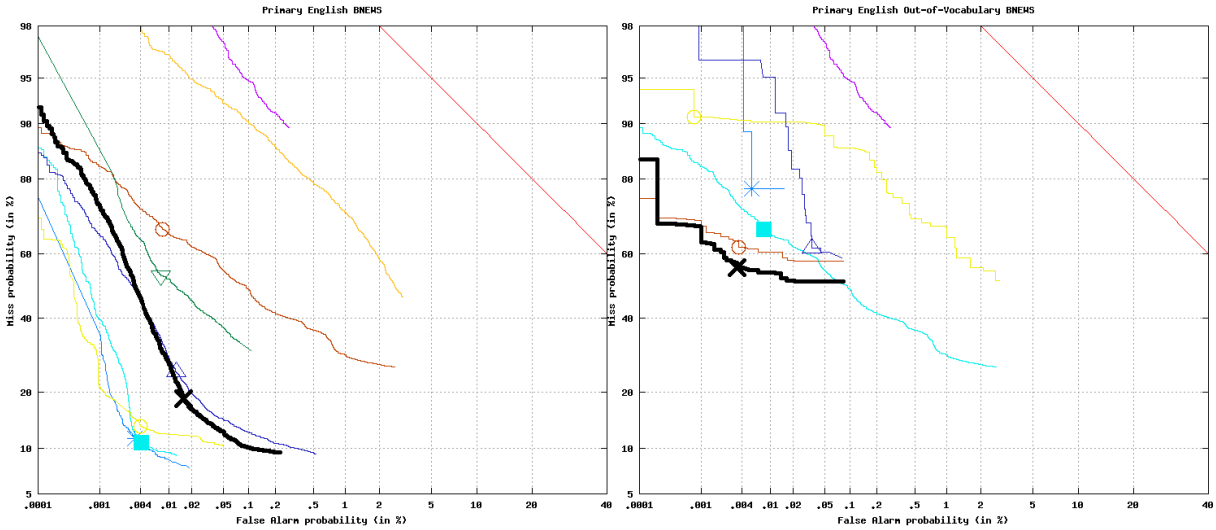

Figure 1: DET curves [5] for English broadcast news systems. Left picture: all terms. Right picture: terms containing OOV. Our system's DET curves are plotted in bold.

of parallel hypotheses), it is able to search even for hypotheses with lower probability. This way a lower miss probability is achieved.

ACKNOWLEDGEMENT

This work was partly supported by European projects AMIDA (IST-033812) and Caretaker (FP6-027231), by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under projects No. 119/2004, No. 162/2005 and No. 201/2006.

REFERENCES

[1] Szöke Igor: Keyword detection in speech data. Concept of Doctoral Thesis, April 2005.

[2] Burget Lukáš, Černocký Jan, Fapšo Michal, Karafiát Martin, Matejka Pavel, Schwarz Petr, Smrž Pavel, and Szöke Igor. Indexing and search methods for spoken documents. In Proceedings of the Ninth International Conference on Text, Speech and Dialogue, TSD 2006, number 4188 in LNCS, pages 351358. Springer Verlag, 2006.

[3] Brin Sergey, Page Larry: The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 30(1–7):107–117, 1998

[4] Fapšo Michal, Smrž Pavel, Schwarz Petr, Szöke Igor, Schwarz Milan, Černocký Jan, Karafiát Martin, Burget Lukáš: Information Retrieval from Spoken Documents, In Proceedings of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2006), Mexico City, MX, Springer, 2006, s. 410-416, ISBN 3-540-32205-1

[5] NIST Spoken Term Detection Evaluations, http://www.nist.gov/speech/tests/std/, 2006