Information Retrieval from Spoken Documents *

Michal Fapšo, Pavel Smrž, Petr Schwarz, Igor Szöke, Milan Schwarz,
Jan Černocký, Martin Karafiát, and Lukáš Burget

Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
speech@fit.vutbr.cz, http://www.fit.vutbr.cz/speech/

Abstract. This paper describes a designed and implemented system for efficient storage, indexing and search in collections of spoken documents that takes advantage of automatic speech recognition. As the quality of current speech recognizers is not sufficient for a great deal of applications, it is necessary to index the ambiguous output of the recognition, i. e. the acyclic graphs of word hypotheses — recognition lattices. Then, it is not possible to directly apply the standard methods known from text-based systems. The paper discusses an optimized indexing system for efficient search in the complex and large data structure that has been developed by our group. The search engine works as a server. The meeting browser JFerret, developed withing the European AMI project, is used as a client to browse search results.

1 Introduction

The most straightforward way to use a large vocabulary continuous speech recognizer (LVCSR) to search in audio data is to use existing search engines on the textual ("1-best") output from the recognizer. For such data, it is possible to use common text indexing techniques. However, these systems have satisfactory results only for high quality speech data with correct pronunciation. In the case of low quality speech data (noisy TV and radio broadcast, meetings, teleconferences) it is highly probable that the recognizer scores a word which is really in the speech worse than another word.

We can however use a richer output of the recognizer – most recognition engines are able to produce an oriented graph of hypotheses (called *lattice*). On contrary to 1-best output, the lattices tend to be complex and large. For efficient searching in such a complex and large data structure, the creation of an optimized indexing system which is the core of each fast search engine is necessary. The proposed system is based on principles used in Google [1]. It consists of **indexer**, **sorter** and **searcher**.

* This work was partly supported by European project AMI (Augmented Multiparty Interaction, FP6-506811) and Grant Agency of Czech Republic under project No. 102/05/0278. Pavel Smrž was supported by MŠMT Research Plan MSM 6383917201. The hardware used in this work was partially provided by CESNET under project No. 119/2004.