

## Automatic collection, annotation and indexing of Czech broadcast speech

In the paper we describe a complex system we developed for automatic acquisition of a large corpus of spoken Czech. The system is capable of continuous monitoring of a selected Czech TV station and providing automatic transcription of its audio track. The transcription is performed by our own speech recognition engine that employs a vocabulary with 320 thousand most frequent Czech words (and word-forms) together with a corresponding language model. Transcription accuracy is fairly good for studio speech (about 90 %), but may drop significantly for noisy recordings and spontaneous speech. Anyway, the system operates without any human supervision and during its operation in 2007 it collected, transcribed, stored and indexed more than 1800 hours of Czech spoken documents. Any word or any combination of words in this corpus can be easily searched by a full-text search engine. The engine finds all occurrences of the queried word(s), ranks them according to several relevancy criteria and prepares them for audio-visual replay (e.g. by the Windows Media Player). The response of the search & play engine is quite fast even though the found programs must be transferred (streamed) via internet from their original data store on the Web pages of the corresponding TV station. (This is fully in accordance with the copyright law.)

During 2007, the system was employed to monitor CT24 channel, which is primarily a news and document broadcasting channel belonging to the state Czech TV. In this way we processed, i.e. transcribed and indexed, more than 50 different types of programs, namely daily news shows, regional news, weather reports, economic and cultural magazines, talk shows, press conferences, broadcast parliament sessions or even some historical castback programs, like "Czech TV 25 years ago". The collection is quite unique because it allows to search in and watch to hundreds of spoken programs broadcast within one year. It allows not only for searching particular facts but also for analyzing speaking styles or pronunciation issues and to evaluate the performance of the recent ASR technology developed for Czech language.

The search engine has an internet interface so that anybody can try it [1]. Its simple layout is displayed in Fig. 1. To make a search, one should type one or more words into the interface's text box. The query can be typed with diacritics or in plain ASCII, wild card symbols (\* or ?) are also allowed. The search can be further narrowed by specifying the program name (using preposition IN), program channel (preposition ON), time period (SINCE and TILL) and even a speaker name (BY). The latter option is enabled by a speaker-recognition module which is part of the complex transcription system. So, the following example query: "*vaclav\* havel OR havl\* SINCE 01.01.2007 TILL 31.03.2007 BY klaus*" can find what current president Vaclav Klaus said about ex-president Vaclav Havel in the specified time period.

The system has been used mainly for demonstration purposes to show current possibilities and limitations of our speech recognition system. However, we believe that it may be useful also for people who are interested in linguistics, phonetics and other aspects of the Czech language.

### References:

- [1] <http://ahmed.ite.tul.cz/>



Fig. 1 – Search engine and its interface

(The query is displayed in the upper part. On the left side, there is a list of found programs in which the searched words occurred. On the right side, the Media Player is playing the selected program together with automatically generated Czech subtitles.)