MULTILINGUAL ON-LINE BROADCAST MONITORING SYSTEM FOR SLAVIC LANGUAGES

Zdansky Jindrich[1], Blavka Karel[1], Nouza Jan[1], Cerva Petr[1], Silovsky Jan[1], Pazour Petr[2], Svab Jan[2]

[1]SpeechLab, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic
{jindrich.zdansky, karel.blavka, jan.nouza, petr.cerva, jan.silovsky}@tul.cz
[2]Newton Media a.s., Na Pankráci 1683/127, 140 00 Praha 4, Czech Republic
{petr.pazour, jan.svab}@newtonmedia.cz

ABSTRACT

We present a recent version of the complex broadcast monitoring platform that is being developed as a joint research work of the Technical University of Liberec and Newton Media company. Its ultimate goal is to automate on-line recording, transcription, indexation, search and other related services - all being applied simultaneously to tens of TV and radio stations in Central and East Europe.

The collaboration between the two partners started in 2006, when the first prototype developed for Czech broadcast monitoring launched its trial run [1]. Since then, the system has been significantly enhanced, namely with respect to the transcription accuracy and speed, by allowing multiple broadcast channels to be processed simultaneously, and last but not least, by including several other Slavic languages in the system's portfolio.

The main modules that make the platform are shown in Fig. 1. The whole system runs on a computer cluster (recently 25 PCs, each with 4 CPU cores). Its control interface allows a human supervisor to select media sources that are to be continually monitored (usually data coming from DVB-T or internet streams), or those processed on demand (previously recorded multimedia data). The task manager is responsible for allocating available CPUs to each of the input audio signals. These are segmented into 60 s long chunks (with 10 s overlap), which are sent to the decoding modules running in parallel on individual CPU cores. The partial transcripts are assembled together using the technique named SDROLA [2]. Its main advantage is that it allows for processing of very long streams with an acceptable latency and with a decoder whose real-time factor may be greater than 1. The text outputs (and their descriptors, such as time stamps, speaker identity, etc) are stored and regularly indexed in the database module. Stored are also the original multimedia data.

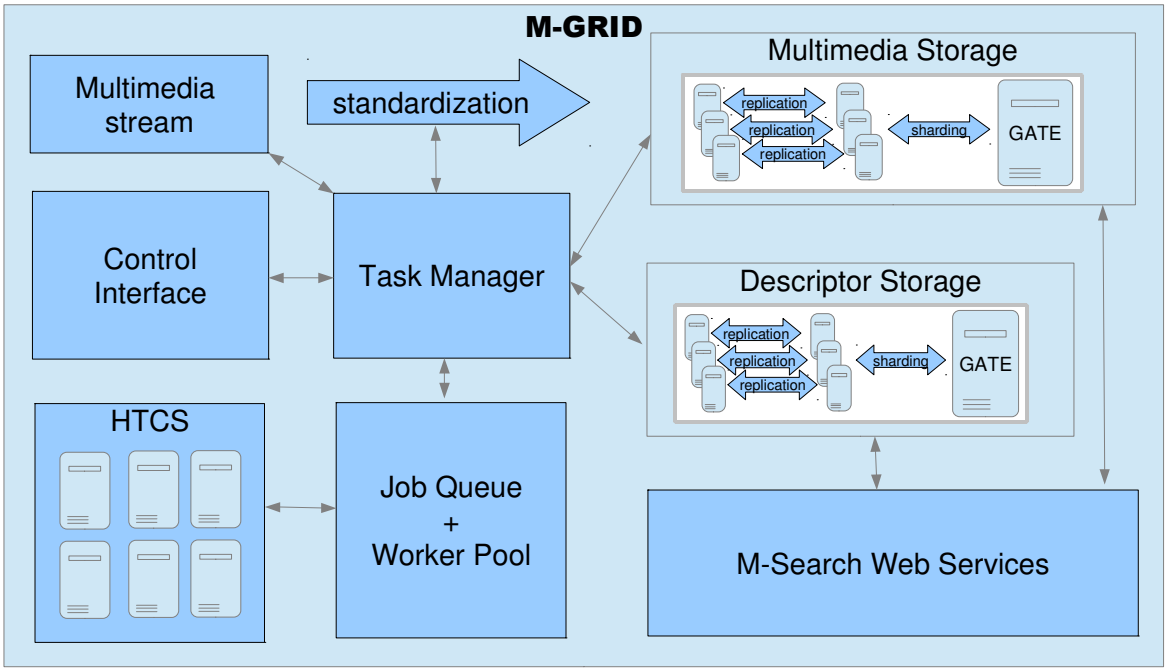

Fig. 1 *Grid (multi-PC) platform for parallel processing of multiple broadcast streams*

All the stored, processed and indexed data can be used for several purposes. The most straightforward one is full-text search. Using a web application like that shown in Fig. 2, one can search for words, word-forms,

phrases and even for speakers. The found records (or their parts) can be played, edited and exported. In this way, the system offers a service analogous to press cutting. Another potential application is a fast alert service operating with a list of client required terms. In the standard setup, broadcast data is available for search (and also for alerts) with a delay that is in range from 3 to 5 minutes. (This latency is a sum of times needed for signal preprocessing including feature adaptation, SDROLA-based transcription and regular re-indexation.) Although, for this specific service, the latency can further drop by shortening the segments used in the SDROLA scheme.

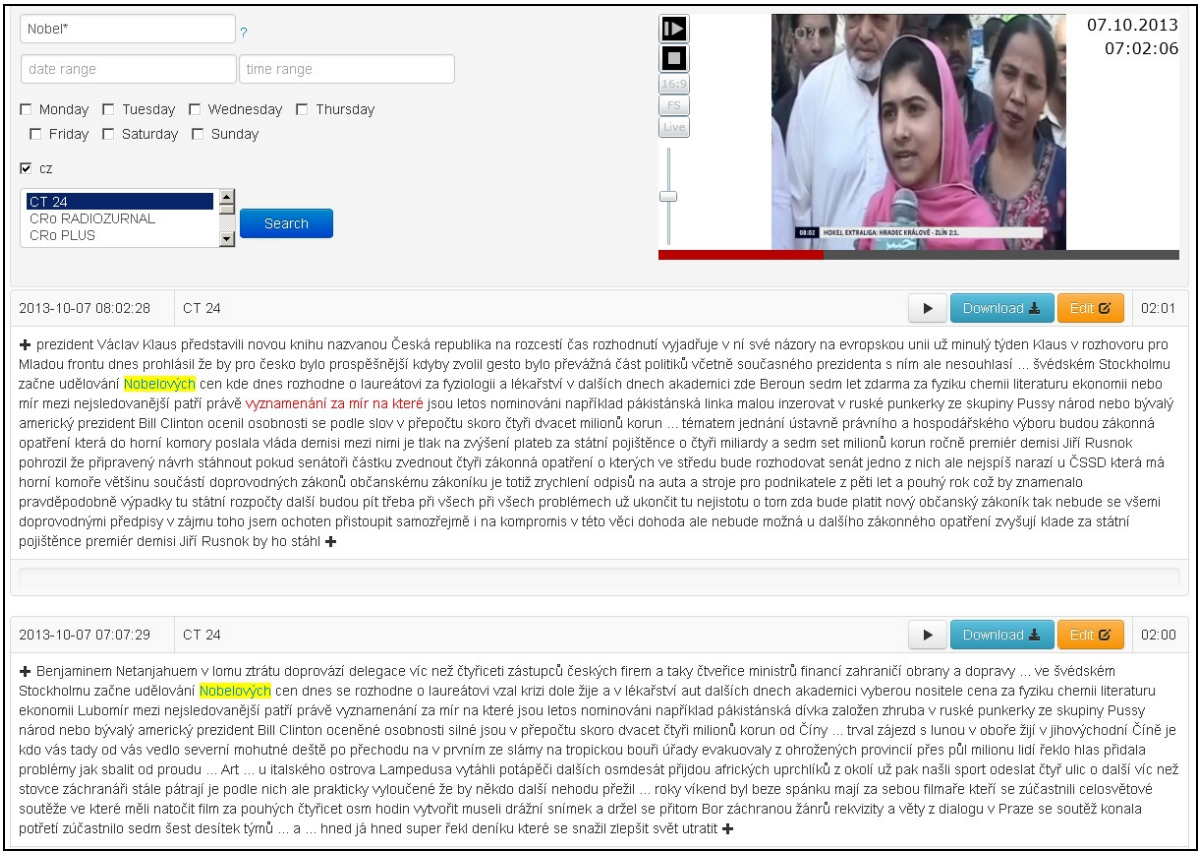

Fig. 2 *Web application for search in the transcribed data. The screenshot shows the most recent occurrences of word "Nobel" (and its inflected word-forms) The first recording was broadcast just 5 minutes before the search.*

Recently, the system is used for monitoring more than 30 broadcast stations in the Czech Republic, Slovakia and Poland. Versions for Russian, Croatian and Serbian languages are currently tested and several other Slavic languages are in a preparation stage. All these languages are challenging as they exhibit a high degree of inflection and require very large vocabularies. The size of used lexicons is in range from 260K (Croatian) to 550K words (Czech). On the other side, our design can benefit from the fact that all the languages have rather regular pronunciation rules, use similar phonetic inventories and employ analogous grammatical structures. The lexicons and language models have been created (and are regularly updated) from publicly available internet resources, mainly electronic newspapers and web pages of major broadcasters. For acoustic model training we use an automatic speech data collection technique and iterative model re-training scheme presented in [3].

ACKNOWLEDGEMENTS

This work has been supported by the Technology Agency of the Czech Republic (project no. TA01011204).

REFERENCES

[1] Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: "Continual On-line Monitoring of Czech Spoken Broadcast Programs". Proc. of Interspeech 2006, Pittsburgh, Sept., 2006, pp. 1650-1653

[2] Zdansky, J.: SDROLA: An Efficient Strategy for Distributed, Accurate Indexing of Spoken Documents, Proc. of 6th Int. Conf. on Informatics and Systems, IEEE, Cairo, 2008 pp. PAR 24-28

[3] Nouza, J., Cerva, P., Kucharova M.. Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages. RADIOENGINEERING, 22(3), pp. 866-873