A System for Information Retrieval from Large Records of Czech Spoken Data*

Jan Nouza, Jindřich Žďánský, Petr Červa, and Jan Kolorenč

SpeechLab, Technical University of Liberec, Hálkova 6 461 17 Liberec 1, Czech Republic
{jan.nouza, jindrich.zdansky, petr.cerva, jan.kolorenc}@tul.cz

Abstract. In the paper we describe a complex multi-level system that serves for automatic search in large records of Czech spoken data. It includes modules for audio signal segmentation, speaker identification and adaptation, speech recognition and full-text search. The search can focus both on key-words and key-speakers. The transcription accuracy is about 79 % (for broadcast programs), search accuracy about 90 %. Due to its distributed platform, the system can operate in almost real-time.

1 Introduction

After great success of applying full-text search technology to extract information from electronically available documents, the focus of IT research is moving towards mining in acoustic and video (multimedia) data [1]. Many important facts can be found in broadcast spoken programs (news, commentaries or debates), as well as in movies, sport and cultural documents. However, these multimedia sources must be transcribed into text form before any search for key-words or key-speakers or topics can be started. Recent advances in automatic speech recognition (ASR) made this audio-to-text transcription possible, though its robustness is still a relevant research issue.

Much attention was devoted namely to the search in broadcast news. The earliest systems developed for information retrieval (IR) in this domain occurred around year 2000, see e.g. [2]. Later, ASR techniques were applied also to other types of voice data, such as historical archives of spoken documents [3] or testimonies processed within the MALACH project [4]. Large progress in mining speech data was reported namely for English [1,2,3] though systems for other languages, like French, are also available [5]. Great majority of the IR systems employ the classic approach where speech is transcribed into sequences of words that are directly used as searched items. An alternative method, presented e.g. in [6], consists in translation of speech signal into phonetic lattices that serve as basis for a later search within arbitrary vocabulary.

In this paper we present a system for word-oriented search in automatically transcribed records of Czech spoken documents. It is based on our ATT (Audio Transcription Toolkit) platform developed during the last 2 years. Since it uses a general, very large lexicon covering about 99 % (frequency ranked) Czech lexical inventory, it is suited for a broader range of tasks. In the following, we mention at least those classified by our collaborating partners as highly important:

* This work was supported by the Czech Grant Agency (GACR grant no. 102/05/0278).

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 485–492, 2006.

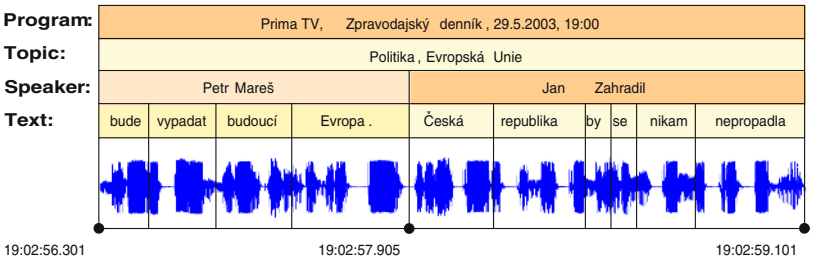

Fig. 1. Hierarchical segmentation of spoken documents

Continual monitoring of TV and radio stations. Here, the interested parties do not insist on perfect transcription of broadcast stream because they know it would need additional human work. Rather, they prefer a formatted text output that is suitable for fast search and allows immediate audio check of found items. An additional demand is to identify speech segments spoken by a particular person.

Searching for specific key-words in regularly broadcast programs. Some clients want to monitor specific TV and radio programs and to search for only a small number of key-words that are a-priori known.

Aligning human-made transcription to audio records of speech. This is the case namely in situations were a formal text record may slightly differ from what was actually spoken (due to artifacts of spontaneous speech) and where this difference may be a source of disputation (e.g. records of parliament talks). Here, the task consists in getting direct access to the audio signal associated with any word or any sentence in the transcription and allowing its easy acoustic check.

2 Task Overview

In general, this project is aimed at automatic search in large spoken data sets (processed off-line) or long acoustic streams (accessed on-line).

In contrast to full-text search in documents, which are already organized into articles, paragraphs, sentences and words, acoustic data is much more complex in its nature. Therefore, one of the key problems in speech data indexing is proper partitioning of the signal. In case of broadcast stream, the most natural segmentation follows the scheme depicted in Fig. 1.

At the highest level, the record is partitioned into individual programs (e.g. news, movies, advertisement, etc). Their identification usually makes no problem because we can use additional information provided by broadcasters in form of program codes and time stamps. At the next level, spoken data is grouped into parts that deal with the same topic. The identification of the topic is a complex task, especially in case of broadcast news where one topic can be covered be several contributions presented by various speakers. Since it utilizes linguistic information, it can be performed as one of the final steps in the data transcription process. In this paper this level is not solved. On the signal level, the first segmentation

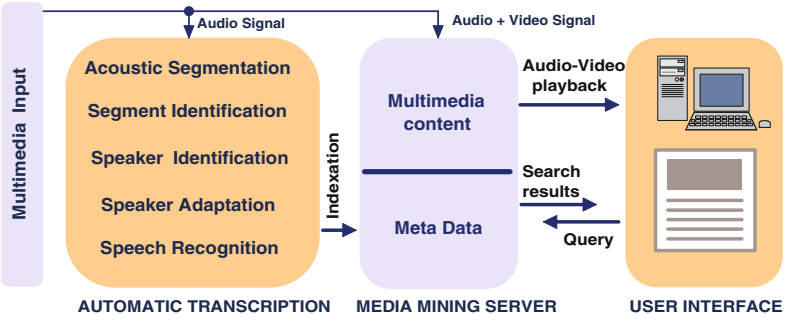

Fig. 2. Modules and functions of spoken data mining system SPOKEMIN

step consists in identifying those time instants where signal changes significantly its long-span character. Usually, these moments correspond to speaker turns, channel switches or speech/non-speech borders. If the segment between two adjacent change points is speech, its data is processed as one signal block (utterance) belonging to a single speaker. It is sent to speaker and speech recognition modules that return back information about speaker identity and speech content. The latter has form of a sequence of words, which are the lowest-level elements of transcription. Program names, topics, speaker info and text, together with their time stamps, make the search space for the information retrieval system.

3 System Architecture

In this section we describe real system called SPOKEMIN (SPOKEn Data MINing) developed in our lab in 2005–2006. A diagram showing its modules and functions is depicted in Fig. 2.

3.1 Signal Processing

The system can process acoustic data that are either stored in previously recorded files or that come directly from on-line sources, such as a microphone, TV/radio card or internet stream broadcasting. The acoustic models in the speech/speaker recognition modules request that the processed data are sampled at 16 kHz rate into 16 bits.

Speech parameterization is done as the first operation applied to the signal. To save time and memory, the signal is processed only once and the resulting features (or their subsets) are used in all modules, i.e. for signal segmentation as well as for speaker and speech recognition. Feature vectors are composed of 40 numbers, classic 39 MFCC parameters and 1 frame energy. Cepstral Mean Subtraction is applied to static MFCCs only after segmentation is done, i.e. for each segment separately.

3.2 Signal Segmentation

The aim of this module is to detect relevant changes in the audio signal and to use them as break-points for segmentation. For this purpose we developed an own segmentation scheme. It is based on the binary splitting method modified so that it can be applied on-line with minimum delay. In our case, the detector collects acoustic data and waits until a 10-second long block is available. Then it applies the binary search for change-points. If more than one is detected, the leftmost one is taken. If no change is found, the detector waits for the next block of data and then repeats the binary search again. When the size of the block exceeds 1 minute (i.e. if there was no change within that block) the detector cuts signal in the middle of the longest silence.

The decision rule for locating and validating a single change point is based on the maximum likelihood approach which was found more accurate then the commonly used techniques based on the Bayesian Information Criterion, as it was shown in [7].

3.3 Speech/Non-speech Detection, Speaker and Gender Identification

This module makes classification of segments into several broader classes. These are: silence (low background noise), noise (loud noise, jingles, music, etc), male and female voice. All these categories are represented by GMMs (Gaussian Mixture Models). The same representation is used also for speakers included in the speaker database. Currently, it includes models of some 300 subjects, mainly TV and radio speakers and top politicians. For each of them we collected at least 75 seconds of speech for training the GMMs and for speaker adaptation purposes. This training data was collected during a longer period (2003–2006) in order to get robust models.

A speech segment is matched to the GMMs using a 2-level classification scheme. On the first level, speech/non-speech separation and speaker ordering is performed by employing 256-mixture GMMs. On the second level, which is applied for speech segments only, speaker verification is done by comparing the best speaker model with a male or female Universal Background Model (UBM). As result, each segment gets one the following labels: non-speech, male, female or a person's name from the speaker list. Moreover, N-best list from the first level is also stored for later use.

3.4 Speaker Adaptation

It is well-known that adaptation of acoustic models to the voice characteristics of individual speakers can improve speech recognition in significant way. Our system utilizes two types of adaptation techniques. For the persons in the speaker database, speaker adapted (SA) models were prepared off-line by a special module that is based on a combination of MAP and MLLR techniques. These SA models are used for those speakers who passed successfully through the verification process. For the rejected ones, a special SA model is computed on the fly by properly mixing the model parameters of the N-best speaker list. For details, see [8].

3.5 Speech Decoding and Transcription

The transcription of the utterance is performed by a speech recognition module. It employs our own LVCSR decoder optimized for large (100K+) vocabularies [9]. The recent version

```
BLOCK=4 BEGIN=36970 LENGTH=2970 IDENTITY=<Daniel_Takáč> <Male>
RAWTEXT=<36970|37100|> <37100|37460|vládní> <37460|37900|představy>
<37900|37960|o> <37960|38430|šetření> <38430|38940|kritizují>...
```

Fig. 3. Example of the SPOKEMIN's transcription format

of the lexicon contains about 312K entries mapped onto some 340K pronunciation baseforms. The lexicon is made of the most frequent Czech words, word-forms and multi-word expressions. It was shown previously that the OOV (out-of-vocabulary) rate for the lexicon of this size was about 1 % for most spoken data [9]. The language model (LM) is based on bigrams estimated from a corpus of some 3.5 GB of Czech text. Acoustic models, in our case 41 phoneme and 7 noise models, are 3-state, 100-mixture HMMs, whose parameters were estimated on a 50-hour speech training database made of broadcast, microphone and telephone records.

Speech transcription is performed by a Viterbi decoder whose goal is to determine the most probable sequence (with unknown length N) of words W from lexicon V together with their time positions (represented by ending times t_n) according to eq. (1):

$$W = \underset{w_n \in V, N, 0<t_n<T}{\arg\max} \left(\sum_{n=1}^{N} \ln(L(w_n|\mathbf{x}(t_{n-1}) \dots \mathbf{x}(t_n))) + \beta \ln(P(w_n|w_n - 1)) \right) \quad (1)$$

where $L()$ denotes likelihood of word w_n for acoustic observation vectors $\mathbf{x}$ in time span t_{n-1} to t_n, $P()$ is the bigram value for word w_n following w_{n-1} and β is LM factor.

The output of the speech recognition module has form depicted in Fig. 3. For each segment (block) it gives its start time and length (measured in milliseconds from the recording start), segment label identification (usually the speaker's name and gender) and a sequence of recognized tokens. (This raw text is further post-processed in a module that cares about capitalization and punctuation.)

3.6 Search in Transcriptions

The above text transcriptions serve for search. In the recent system's version, a query is defined by typing the searched word or its sub-string into the form of the IR interface that is depicted in Fig. 4. The search engine finds all occurrences of the items and makes their time-ordered list. Any of the listed items can be checked visually by reading the transcription of the given segment and acoustically by listening to the whole segment or to the found word only. (Later, also video replay will be available.)

3.7 Implementation

One of the most importing features of an IR system is its operating speed and response time. Much care was devoted to speeding up the operation of all the modules, namely the speech decoder. After many optimizations of the code, now it is able to process an utterance in time which is usually shorter than 2xRT (two times duration of the speech) on a PC with a Pentium 3 GHz HT processor. Moreover, the ATT platform was designed to run in distributed

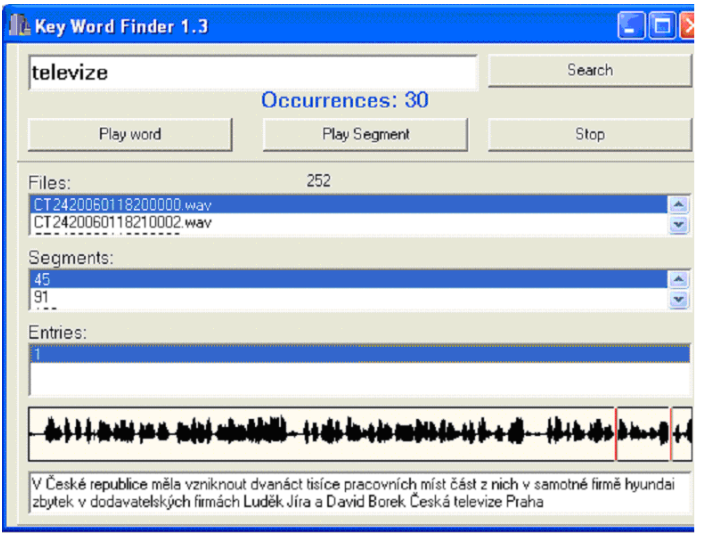

Fig. 4. Interface to the search engine operating over the transcriptions

environment. It means that after the signal is segmented each segment can be processed on a separate machine. Our experience show that if 3 and more client PCs are connected within the ATT network, the transcription time gets shorter than the duration of the processed records.

4 Performance Evaluation

The system and its modules have been exposed to many evaluation tests. In this chapter we mention at least the most relevant ones. The test data used in the experiments were 3 complete programs recorded from CTV 1 station in December 2005. Two of them were main evening news called “Události” (Dec. 9 and 12), the other was “Večerník” of Dec. 7. They were recorded as whole, i.e. with initial and final jingles as well as with opening and intermediate headlines spoken on music background. Three shows with total length of 62 minutes were split into 430 segments (from these 16 contained no speech). Their reference transcriptions included 9760 words. From them, 109 (1.12 %) were OOV words.

4.1 Speech Recognition Results

In the first set of experiments we compared different types of acoustic models employed in speech recognition. Speaker-independent (SI) models served for establishing baseline results (see Table 1.). In the next test we employed gender-dependent (GD) models according to the identified speaker’s gender. The other 2 experiments measured the impact of the speaker adaptation (SA) techniques. In the first test, off-line prepared SA models were used for the verified speakers, the unknown and rejected ones were backed up by the GD models. In the second test, SA models were computed even for the group of the rejected speakers after

Table 1. Word error rate and real time factor for different acoustic models

Acoustic model	SI	GD	GD + SA	SA
WER [%]	24.55	22.80	21.93	21.43
xRT	2.45	2.35	2.23	2.15

employing the unsupervised SA technique mentioned in paragraph 3.4. From Table 1 we can observe that the application of the best scheme yielded 3.43 % absolute reduction of WER when compared to the baseline. Moreover, using the speaker-fitted models had positive effect also on faster recognition as shown in the lower row of Table 1.

4.2 Segment Classification Results

In the second set of experiments, the performance of the segment classification was evaluated. Table 2 presents the results in very detailed way, when explicitly showing all situations of correct and wrong decisions. We can see that most wrong decisions happened when a known speaker (one of the 300 people in the speaker database) did not pass the verification stage but he/she was assigned to the proper gender group. (Detailed analysis showed that the correct speaker was usually the first one on the ordered list, which is a fact that still can be utilized in practical search tasks.) Another interesting observation is that 5 of 16 non-speech segments were labeled as speech. Often this was caused in situations where laugh or cough occurred in these segments.

4.3 Search Results

In Table 3 we present some results from search experiments. We chose a small set of words that were often used in media searching, such as VIP subjects and institutions, company and country names, etc. Our set of 10 entities was composed to cover both short (monosyllabic) key-words as well as longer items. In the analyzed TV programs they occurred 124 times in total. Statistics saying how many times they were found at correct places and how often they were omitted or wrongly identified is shown in Table 3. It should be noted that both the types of incorrect search were often caused by confusions between acoustically very close morphological derivations of the same word (e.g. “policie” and “policii” are two forms equivalent two one English word “police”). Anyway, we can see that about 90 % of all searched words were found correctly in the broadcast programs. Similar performance was observed also for other types of spoken data, e.g. parliament debates, which is shown also in Table 3.

Table 2. Automatic classification of segments into speaker, gender and non-speech groups

Audio segment with	Correctly recognized	Wrongly recognized as				
		KM	KF	UM	UF	NS
Known male - KM	104	1	0	68	0	0
Known female - KF	57	0	2	0	26	0
Unknown male - UM	115	4	0	x	1	0
Unknown female - UF	31	0	2	3	x	0
Non-speech - NS	11	2	1	1	1	x

Table 3. Search results for 10 key-words retrieved from different spoken data streams

key-word	62-minute record of TV broadcast				120-minute file of parliament speech			
	Total occur.	Correctly found	not found	Wrongly found	Total occur.	Correctly found	Not found	Wrongly found
televize	30	29	0	1	2	0	0	2
policie	17	15	2	0	0	0	0	0
soud	15	13	1	1	2	2	0	0
Hyundai	12	8	4	0	0	0	0	0
zákon	11	11	0	0	52	48	2	2
Polsko	10	9	1	0	0	0	0	0
premiér	9	9	0	0	1	0	0	1
president	9	8	1	0	0	0	0	0
ministr	8	7	1	0	13	12	0	1
Klaus	3	3	0	0	1	0	0	1

5 Conclusions and Further Work

System SPOKEMIN is capable of automatic monitoring of Czech spoken data, transcribing its content and allowing a full-text search in the transcriptions. It can operate on-line with a delay shorter than several tens of seconds. It supports parallel processing distributed to a cluster of PCs, which allows for off-line processing of speech records in times that are shorter than their duration. Its recent version has been in trial use in a media mining company since March 2006. First experience shows that the system saves a lot of human work. In near future, its capabilities will be extended towards more advanced search options and a link to video files (e.g. for TV news).

References

1. Semantic Retrieval of Multimedia. IEEE Signal Processing Magazine. March 2006.
2. Makhoul J. et al: Speech and Language Technologies for Audio Indexing and Retrieval. Proc. IEEE, vol. 88, no. 8, pp. 1338–1353, 2000.
3. Zhou B., Hansen J.H. L.: Speechfind: An Experimental On-line Spoken Document Retrieval System for Historical Audio Archives. Proc. of ICSLP 2002, Denver.
4. Byrne W. et al.: Aut. recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. on SAP vol. 12, no. 4: pp. 420–435.
5. Favre B., Bechet F., Nocera P.: Mining Broadcast News Data: Robust Information Extraction from Words Lattices. Proc. of EuroSpeech 2005. Lisbon, Sept. 2005.
6. Kurimo M., Turunen V., Ekman I: An Evaluation of a Spoken Document Retrieval Baseline System in Finnish. Proc. of ICSLSP 2004. Jeju, October 2004.
7. Žďánský J.: Methods for Speaker Detection Change Identification in Audio Signal. PhD thesis (in Czech). Technical University in Liberec. October 2005.
8. Červa, P., David, P., Nouza, J.: Acoustic Modeling Based on Speaker Recognition and Adaptation for Improved Transcription of Broadcast Programs. Proc. of Specom 2005, October, 2005, Patras, Greece, pp. 183–186.
9. Nouza, J., Žďánský J., David, P., Červa, P., Kolorenc, J., Nejedlova, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. Proc. of Interspeech 2005, Lisboa, Portugal, pp. 1681–1684.