Voice Technology to Enable Sophisticated Access to Historical Audio Archive of the Czech Radio

Jan Nouza, Karel Blavka, Marek Bohac, Petr Cerva, Jindrich Zdansky, Jan Silovsky, and Jan Prazak

Institute of Information Technology and Electronics, Technical Univesity of Liberec
Studentska 2, 461 17 Liberec, Czech Republic
{jan.nouza,karel.blavka,marek.bohac,petr.cerva,jindrich.zdansky, jan.silovsky,jan.prazak}@tul.cz

Abstract. The Czech Radio archive of spoken documents is considered one of the gems of the Czech cultural heritage. It contains the largest collection (more than 100.000 hours) of spoken documents recorded during the last 90 years. We are developing a complex platform that should automatically transcribe a significant portion of the archive, index it and eventually prepare it for full-text search. The four-year project supported by the Czech Ministry of culture is challenging in the way that it copes with huge volumes of data, with historical as well as contemporary language, a rather low signal quality in case of old recordings, and also with documents spoken not only in Czech but also in Slovak. The technology used includes speech, speaker and language recognition modules, speaker and channel adaptation components, tools for data indexation and retrieval, and a web interface that allows for public access to the archive. Recently, a demo version of the platform is available for testing and searching in some 10.000 hours of already processed data.

Keywords: audio archive processing, speech-to-text, audio search.

1 Introduction

One of the prospective application areas for modern voice technology is automatic processing of audio archives containing speech. The ultimate goal is to transcribe them and store the transcriptions in the database, in which one can search and retrieve a piece of information he or she is interested in. The search needs not to be focused only on the content, but also on linguistic issues such as the speaking style, pronunciation, spoken language evolution, etc.

The systems developed for audio archive processing are very complex. They include speech and speaker recognition modules as well as tools for indexation, full-text search and audio play. They have been designed namely for broadcast documents [1], for national archives of spoken language [2,3] or for special-purpose collections such as, e.g. MALACH [4]. We have been working on this topic since 2005 [5] and some of the developed tools have been already deployed in broadcast data mining [6].

C. Grana and R. Cucchiara (Eds.): MM4CH 2011, CCIS 247, pp. 27–38, 2012.

This paper presents a large applied-research project supported by the Czech Ministry of culture. The project aims at processing the archive of historical and contemporary recordings of the Czech Radio and making its content available for public access. The archive covers almost 90 years of broadcasting and contains hundreds of thousands spoken documents, from which a large portion should be transcribed by a speech-to-text system developed specially for this purpose in our lab. The project started in 2011 and will run for 4 years. During that period we are going to build the complete archive processing and accessing platform and employ it for transcribing about 100.000 hours of audio data. The project offers several major challenges: a) working with huge volumes of data, b) managing Czech language with its highly inflective nature and very large lexical inventory with more than one million word-forms, c) identifying and processing documents spoken also in Slovak, which was the second official language in former Czechoslovakia, d) dealing with the language and lexicon evolving during the 90-year period and influenced by different political regimes, and last but not least e) coping with rather poor quality of most historical recordings.

In 2000s, a large part of the Czech Radio data has been digitized but the individual recordings are stored on tapes, on CDs, or on hard disks. If one wants to search for any particular piece of information, he or she must browse the catalogue where each document is described by its name, the date of recording or broadcasting and several tags or key-words. Then, it is necessary to retrieve the media from the archive and listen to them in hope that the required information will be found, eventually.

The primary goal of the project is to make this search automated and more comfortable. The user should be able to get answers not only to simple queries made of words or phrases, but he or she could search also for e.g. utterances spoken by a selected person, for historically first occurrence of a given word, or for a particular pronunciation of a word. Moreover, the queries can combine various search criteria to answer, for example, a question like: What did person A say about person B within time period T? Hence, the potential users of the system will be not only the people from the Czech Radio itself, but anybody who is interested in media data mining as well as historians, linguists, phoneticians, communication specialists, students, etc.

Though the project is running its first year, now, the concept of the system has been already set up and some of the essential modules already exist as functional prototypes. We have used them to build a preliminary version of the system to test different techniques and approaches, and to demonstrate it to prospective users. At the moment, the system allows for access to some 10.000 hours of transcribed recordings.

2 Audio Archive Processing Platform

Let us present the archive processing and accessing platform (APAP) from the user's point view, first. Its general overview is depicted in Fig. 1.

Anybody who wants to search in the archive needs to have an access to internet. On the dedicated web page, he or she enters the word(s) or phrase(s), he

or she is interested in, optionally sets some search constraints (e.g. time period, program name, speaker name, etc) and clicks the Search button. Immediately after that, the documents meeting the given criteria occur on the screen, being ordered according to the chosen relevancy rate. By clicking on the selected document, the part containing the searched term starts to be played. The user can easily navigate within the text, read it and listen to any part of it.

The associated audio data are streamed from the server located in the premises of the Czech Radio. The link between the audio and the text is provided by the database server. It contains rich text transcriptions created during the process of speech-to-text conversion and indexation. This is the core function of the whole platform as it includes complex audio data processing procedures mentioned below. These procedures are time consuming and they are performed off-line by a computer cluster.

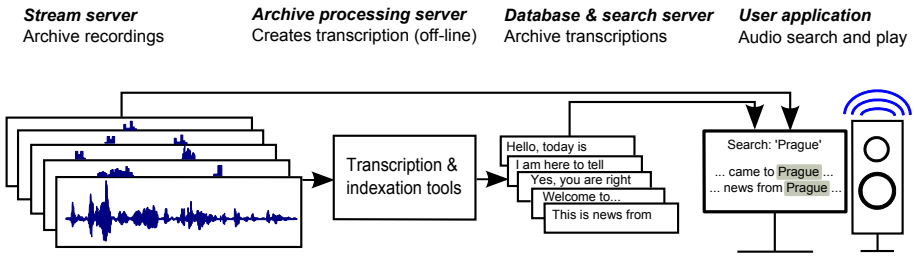

Fig. 1. Data flow in APAP (Archive Processing and Accessing Platform)

The technology behind the platform is much more complex. Its aim is a) to create text documents from the audio ones, b) to establish detailed links between them, c) to store them in the database and d) to allow for their later retrieval. The platform's core is made of several modules as shown in Fig. 2.

The standard procedure for transcribing and indexing an audio document runs as follows: The document passes through an audio processing module where signal samples are converted into feature vectors that are later used in identification, classification and decoding tasks. The next step consists in segmentation of the running signal into speech and non-speech parts (e.g. long silence, noise or music). The speech segments pass into the module that searches for significant changes in signal character, which can be either speaker turns or changes in the signal band-with (usually a part containing telephone talk). These change points are used to split the speech into individual utterances. For each utterance, we try to determine some relevant characteristics such as broad/narrow band signal, clean/noisy speech and male/female speaker. Optionally, we may employ also a speaker identification module operating with a database of a priori known voices. All this kind of information is used to set up the speech recognition module so that it can benefit from employing the proper acoustic model, e.g. the gender or speaker dependent one. The recognition module performs speech decoding using the given (general or topic oriented) lexicon and the corresponding language

model. The output from the decoder is the best sequence of recognized words with their pronunciations and time markers. The latter represent beginning and ending times (measured in milliseconds from the start of the document) for each word and each identified non-speech event, and they serve for aligning the audio signal with its text version. Eventually, the raw output from the recognizer undergoes a post-processing stage where, for example, the sequences of numerals are replaced by digits, capital letters and punctuation are added, etc. The final transcription together with the time markers and complementary information, such as speaker's identity, is indexed and stored in the database. More technical details on the modules and procedures can be found in section 4.

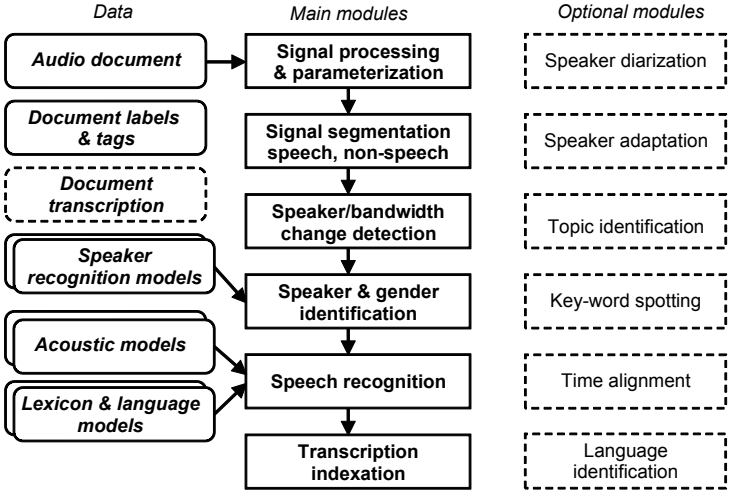

Fig. 2. Data and document processing modules in APAP

Besides this main line of modules, there are also several optional components in the system. The speaker diarization module searches for all segments in the document that belong to the same speakers. This is useful, for example, in the transcription of debate programs where speech segments belonging to speakers unknown to the system can be at least identified as of person A, person B, etc. The speaker adaptation module allows for fitting the available acoustic model to the currently speaking person, which can be employed to improve the accuracy in an optional two-pass decoding procedure. The system performance can be further enhanced if the topic of the document is known and hence a topic dependent lexicon and language model can be employed. The topic can be determined either a priori, from the document tags, or a posteriori during the first-pass speech decoding. For the same purpose we can utilize also a fast performing key-word spotting module. In special situations, where for some documents their transcriptions already exist, the slower speech recognition procedure can be replaced by a

much faster signal-to-text alignment routine whose aim is to generate the missing time markers [7]. This technique is very helpful as many recently broadcast programs already have their text forms. It not only saves computing time, but also allows us to index the already existing human-made transcriptions, which are almost 100 % accurate and which can serve also for feed-back training of the recognition modules. At the end of this brief overview, let us mention also a module that must be developed specifically for this project. Its goal is to identify the language of the currently processed utterance, in our case, Czech or Slovak. The two languages are linguistically similar but with significantly different lexicon and grammar.

3 Archive Data

The data in the archive represent almost 90 years of broadcasting in Czechoslovakia and in the Czech Republic. When founded in 1923, the Czech Radiojournal company was only one year younger than the premier world broadcaster, the BBC. Later, the company was transformed into the Czechoslovak Radio and with the split of Czechoslovakia (in 1993) the Czech Radio took over the service and the historical archive. The Czech Radio is the public broadcaster and recently it runs 8 nation-wide channels and 10 regional programs. Two channels are mainly news oriented, another focuses on science programs, and the others offer a mix of spoken word, music and leisure. The three word-oriented channels produce about 8000 hours of unique documents a year.

The oldest preserved recordings date to late 1920s and since 1945 there is an (almost) un-interrupted series of daily news programs recorded originally on tapes and recently digitized. The archive contains more than 100.000 hours of spoken documents, most of them being news programs, daily commentaries, debates, talk-shows. This is the domain the project focuses on.

The audio data that are to be processed and made available for public search differ in technical as well as social aspects, which makes the project really challenging. The documents (or their parts) may vary with respect to:

- *original storage media:* analog (film, tape) or digital memory,
- *audio band-width:* narrow band (AM or telephone), or wide band signal,
- *digital quality:* CD quality as well as highly compressed loss formats,
- *background noise:* from negligible one in case of studio recordings to large-noise in field records; music or another speech may occur in background,
- *speaking style:* read speech, planned talk, spontaneous conversation,
- *historical and social issues:* contemporary language as well as archaic language from 1930s and 1940s, lexicon of communist era (1945-1989), etc.
- *national language:* mainly Czech, but also Slovak (in particular before 1993).

After a closer survey of the archive data, we have classified them into several historical epochs. Because our strategy is to process the archive from the present time towards the history, we denote the contemporary epoch with index 0 and the previous ones with negative indexes. (The positive indexes will be reserved for future epochs if necessary.)

Epoch E0 (2000 – present). The documents from this epoch are usually in good digital quality and when compressed, the distortion is not severe. The amount of recordings is very high (thousands of hours a year). Moreover, literal transcriptions are available for some of them. This will allow us to perform the automatic alignment procedure to speed up the initial feeding of the database and at the same time to re-train the existing acoustic model on the target data. The official transcriptions will also give us enough information to make a large database of voices for the speaker identification module. The lexicon and language model can be further enhanced by analyzing additional text resources, mainly electronic versions of newspapers. Czech is the major language used in the documents and rarely occurring Slovak can be skipped or neglected – at least in the initial stage of the project.

Epoch E-1 (1990 – 2000). The audio data from this epoch were digitized mostly in times where the available storage capacity was strictly limited and therefore most data is compressed using loss formats (mp3, mp2, etc). There exist no literal transcriptions for the documents, only brief summaries and tags. For lexicon and language model building only a limited amount of newspaper texts is available in electronic form. In early 1990s, Slovak frequently occurs as the second language in the spoken documents.

Epoch E-2 (1968 – 1989). The data from this period were originally stored on analog tapes and later converted into loss audio format. No transcriptions neither electronic texts are available to adjust the lexicon, which was significantly influenced by the ruling communist regime. Here, we hope to get access to at least some amount of scanned and OCRed newspapers from that period. The major type of the processed documents will be the daily recordings of main evening news where Czech and Slovak are mixed regularly.

Epoch E-3 (1945 – 1967). The audio archive contains only one major program per day (evening news). Most data is still stored on analog tapes and it will have to be digitized on demand. The signal has very low quality, partly because of AM broadcasting in those days and partly because of the storage media. For adapting the acoustic model to the given signal quality and for building the proper lexicon and language model, some parts of the archive will have to be manually transcribed. We expect that the language (Czech and Slovak) of this period will be also influenced by Russian, as Czechoslovakia was part of the Soviet political sphere that time.

Epoch E-4 (before 1945). From this epoch, there is only a limited amount of spoken documents. However, they have a great historical value because they include, for example, addresses and talks of the first Czechoslovak presidents, speeches from the parliament, programs broadcasted during WWII and the German occupation, etc. It is almost sure that these documents will require individual care to get them transcribed and indexed in the database.

As mentioned earlier, the archive processing work will go against the flow of time, from the present towards the past. The reason is rational. The transcription system has been developed for contemporary language using a large amount of audio and text data collected during the last decade. When moving backwards we have to adjust the lexicon, the language model as well as the acoustic one towards the character of the data of the previous epochs. In other words, we will have to adapt the speech processing front-end and the acoustic model to the gradually decreasing signal quality. At the same time, it will be necessary to modify the lexicon by adding the words specific to the given period. The language model will be forced to forget continually the contemporary phrases and collocations and learn those typical for the target period. The further to the past we go, the harder this tasks will be for automation, as the amount of data available for training the statistical models will be smaller and smaller. Yet, our preliminary experiments have proved that this way back was feasible.

4 Technical Solutions

All the speech processing modules for the APAP are being developed in our lab. The core components, such as the large-vocabulary continuous-speech recognition (LVCSR) system for Czech, or the speaker identification (SID) tool, already exist and the main task in the project is to adapt them for the target application. In case of the LVCSR, the main focus is put on increasing the size of the operating vocabularies (up to 1 million entries) and on eliminating the impact of lower signal quality. Moreover, the Slovak version of the recognizer must be built. The other components, such as the language identification (LID) module, the speaker diarization module, or the speaker adaptation module are still under development. In the following text, we shall briefly describe the main technical parameters of the platform.

4.1 Audio Processing and Acoustic Modeling Part

Because the data in the audio archive are stored in various formats, it is necessary to convert them into a standard one before they enter the signal processing routines. This standard has been set to 16 kHz, 16 bit, PCM WAV format and we use the popular FFmpeg tool [8] for the conversion. After that, the audio signal is parameterized into a stream of feature vectors. These are 39-dimensional mel-frequency cepstral coefficients (MFCCs) computed every 10 ms. The vector includes 13 static, 13 delta and 13 delta-delta coefficients. Using a 2-second long sliding window, the MFCC features are normalized by the cepstral mean subtraction technique.

The acoustic-phonetic inventory of spoken Czech includes 40 phonemes and 8 noise types. They are represented by continuous-density hidden Markov models (HMM) trained on the database of spoken Czech. Recently, it contains 120 hours of annotated speech from more than 1000 speakers. Approximately one half of that amount is made of read speech recordings containing phonetically balanced

utterances, the other half comes from broadcasting. This second part is continuously growing, being fed by the already processed archive data. Two types of HMMs have been trained on this database: the context-independent units (monophones) and the context-dependent ones (triphones). The latter (currently represented by 75 thousand gaussians) are employed in the main transcription task, while the former (with some 45 thousand gaussians) find use in situations where higher speed and lower memory requirements are important, namely in the word-spotting and text alignment routines. There are separate models for each gender, and for wide-band and telephone signal. The proper type of model is determined in the speaker and channel recognition modules that employ the same feature vectors and gaussian mixture model (GMM) based classifiers.

4.2 Linguistic Part and Speech Decoding

The linguistic part of the LVCSR system is made of a lexicon and a language model (LM). Recently, the universal lexicon used for the transcription of contemporary archive documents contains 340k words and word-forms. These are the most frequent lexical units that occurred in the 40 GB large corpus of texts covering national media since 1990. The number of all distinctive Czech words found in the corpus is higher than 2 millions and we have chosen those that appeared at least 50 times. The lexicon is gradually growing as new words get over the threshold. For most documents, this size of lexicon can assure the out-of-vocabulary (OOV) rate lower than 2 %. If we wished to get below 1 % for the majority of spoken documents, the lexicon should contain at least 800k entries. At the moment, this is not feasible because it would slow down the transcription process significantly. Yet, we plan to reach that size in near future. It should be also noted that more than one tenth of the words in the lexicon have multiple alternative pronunciations. Their total number is 390k, currently. The language model is based on bigrams. In the above mentioned 40 GB corpus of contemporary Czech texts we found 130 million different word-pairs. The unseen ones have been backed-off by the Witten-Bell smoothing technique, which optimally fits to our implementation of the speech decoder.

The decoder has been designed to manage vocabularies up to 1 million distinct words. When it operates with the current set of 390k pronunciation forms, it is able to do it in real time, at least for clean speech. For larger vocabularies, more spontaneous and noisy utterances, the processing time may be doubled, or in an extreme case, even tripled. When required, speech transcription runs in a two-pass mode. In the first one, a smaller vocabulary with approx. 50k words is utilized with the aim to obtain a good estimate of the phonetic transcription of the utterance. Unsupervised adaptation based on the MLLR technique is performed on this data and immediately applied in the second pass, in which the full lexicon is employed. In the two-pass case, the total transcription time is about 1.5 times longer. Some figures illustrating the system performance can be found in section 5.

4.3 Database and Search Part

The modules mentioned above generate results in form similar to that displayed in Table 1. For each document, this data is stored in the database. We decided for the MySQL [9] solution as it optimally fits to the type and size of data. Every word and every speaker occurrence is indexed using the Sphinx [10] platform, which proved to be fast and flexible enough both for the indexation as well as the search tasks.

Table 1. Data generated by transcription system and stored & indexed in archive database. (The presented data are real and they correspond to the document shown in Fig. 3. The start and end times are in milliseconds.)

Document identification

Channel: CRo1 Program: 20 minut Broadcast: 2009-09-23 17:33

Segment/speaker	***Start***	***End***	***Text***	***Phonetic***
Non-speech	000000	004520	-	/jingle/
Helena Šulcová	004520	005140	Lisabonská	lisabonská
	005140	005560	smlouva	smlouva
	...	...	...	...
Václav Klaus	054740	055070	Tak	tak
	055070	055320	jedna	jedna
	055320	055640	věc	vjec
	...	...	...	...
	073230	073480	Bruselu	bruselu

4.4 User Interface Part

The ultimate goal of the project is to allow for public search in the transcribed archive documents. A user just needs to be connected to internet and have a properly set up web browser that supports audio playback. Currently, he or she can use the demo version of the search interface displayed in Fig. 3 and available at [11].

The user has a lot of choices to formulate a query. He or she can search for a word (or its part using the * convention), a phrase or multiple words. Furthermore, the user can specify the speaker, the broadcast channel, the program name or the time period. After the Search button is pressed, the number of found documents is shown and their list is available for reading and playing with the searched terms being highlighted. The user can click on any word in the selected document and the replay starts from that point. During the playback, the words corresponding to the running audio signal are shown in red color. A picture associated with the speaker or the topic of the document can be retrieved from the archive, too.

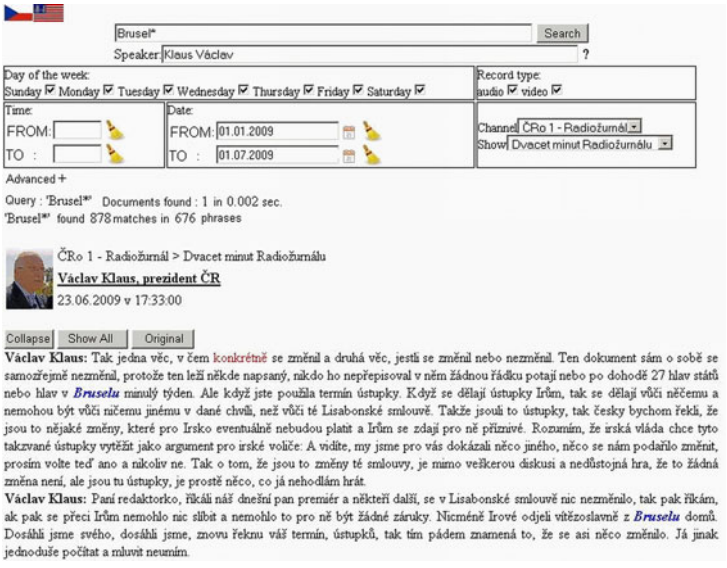

Fig. 3. Web interface for archive search. (The screenshot shows one of the documents containing searched word Brusel* and spoken by Václav Klaus in the given year. The searched word occurs in 73230th millisecond of the talk – compare to Table 1.)

5 Performance Evaluation

In 2011, we have been focusing mainly on setting up the first version of the audio archive platform and on processing the documents from Epoch E0. As explained in section 3, the data from this time period are well covered by the lexicon and language model created previously in our lab [6]. Moreover, for a significant number of documents we have already had their official transcriptions. These can be easily and reliably indexed via the time-alignment procedure mentioned in section 2. An additional benefit is that we can evaluate the performance of the recognition system by comparing its output to the official transcriptions. We do it regularly to investigate the impact of different system settings and the effect of continuously updated acoustic and language model. Some figures illustrating the performance of the current version are summarized in Table 2.

Table 2. Transcription system performance. (The results were obtained with 340k lexicon, bigram LM, 1-pass mode and 10 hours of test recordings from epoch E0 in broadcast quality.)

Document type	Accuracy [%]	OOV [%]	Real-time factor
News from studio	94.67	1.23	1.04
Complete news shows	86.86	1.36	1.26
Talk programs - politics	83.53	1.12	1.43
Talk programs - science	81.61	2.38	1.52

Table 2 shows the basic performance figures for four different types of audio documents. We can see that the transcription accuracy is quite high for news read in studio – 94.67 %. It should be further noted that in this case many recognition errors are deletions of short words (namely one-phoneme prepositions, as 'v', 's', 'z') or minor substitutions in acoustically similar word suffixes due to complex Czech morphology. The results are obviously worse for complete news shows because of their parts recorded out of studio or those with background noise and music. The transcription of discussions and debates suffers mainly due to spontaneous and overlapping speech. We can also see that the debates dealing with more general topics, such as daily politics, are better covered by the lexicon and the language model (which was trained mainly on newspaper texts) than those broadcasted by the science-oriented channel.

Unfortunately, the results get worse if we have to work with loss audio format, such as mp3. This is the real situation because most data in the Czech Radio archive is highly compressed and stored in mp3 files. In that case, the transcription accuracy may decrease by 3 to 5 %. We try to compensate this type of signal degradation in two different ways: The first one utilizes the acoustic models trained on the data that passed through an mp3 decoder and hence better match the compressed audio files. The other approach consists in applying the speaker and channel adaptation technique to each utterance and running the second recognition pass as described in section 4.2.

It should be also noted that the lower recognition accuracy obtained for some parts of the archive does not necessarily mean critical malfunction of the search task. In practice, most queries focus on words or phrases that are more than one syllable long and these have significantly larger chance to be recognized well. So, even though the utterance is transcribed with some errors, most key-words are still searchable and the user usually gets the access to the piece of information he or she is interested in.

6 Conclusions and Future Work

National archives of spoken documents represent a very specific area of cultural heritage. So far they could not be accessed by wide public because their virtual and rather abstract form, together with their specific way of storage, did not allow it. Our project shows that it will be possible soon, thanks to modern information and multimedia technology. At the moment, almost 10.000 documents from the Czech Radio archive have been already transcribed and made accessible. These documents come mainly from the last decade and their processing and publishing was easier compared to what we may expect when moving back towards the historical part of the archive. In order to accomplish the future challenges, we have already started complementary research works, such as scanning historical newspapers and converting them into electronic texts, collecting Czech words and phrases that were typical for specific periods of the Czech and Czechoslovak history and, last but not least, we have begun the development of a module that will be able to process also the Slovak language. The results seem to be promising so far [12].

Although the current project is focused on audio data, the platform is being designed to manage the video broadcast archives as well. This additional feature has been already demonstrated on a small part of the Czech TV archive of news programs [6].

Acknowledgments. This work was supported by project no. DF11P01OVV013 provided by Czech Ministry of culture in research program NAKI.

References

1. Hayashi, Y., et al.: Speech-based and video-supported indexing multimedia broadcast news. In: Proc. ACM SIGIR (2003)
2. Ordelman, R., de Jong, F., Huijbregts, M., van Leeuwen, D.: Robust audio indexing for Dutch spoken word collections. In 16th Int. Conference of the Association for History and Computing, Humanities, Computers and Cultural Heritage, Amsterdam, pp. 215–223 (2005)
3. Hansen, J.H.L., Huang, R., Zhou, B., Seadle, M., Deller, J.R., Gurijala, A.R., Kurimo, M., Angkititrakul, P.: SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. IEEE Trans. on Speech and Audio Processing 13(5), 712–730 (2005)
4. Byrne, W., et al.: Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. Speech Audio Process. 12(4), 420–435 (2004)
5. Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: A System for Information Retrieval from Large Records of Czech Spoken Data. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 485–492. Springer, Heidelberg (2006)
6. Nouza, J., Zdansky, J., Cerva, P.: System for automatic collection, annotation and indexing of Czech broadcast speech with full-text search. In: 15th IEEE Mediterranean Electrotechnical Conference (MELECON 2010), Malta, pp. 202–205 (2010)
7. Nouza, J., Zdansky, J.: Automatic Alignment between Speech Records and Their Text Transcriptions for Audio Archive Indexing and Searching. In: 6th IEEE Conference on Informatics and Systems, pp. MM6–MM12. IEEE, Egypt (2008)
8. FFmpeg converter program, `http://www.ffmpeg.org/`
9. MySQL platform, `http://www.mysql.com/`
10. SPHINX platform, `http://sphinxsearch.com/`
11. Demo of APAP platform, `http://ahmed.ite.tul.cz/demo/`
12. Nouza, J., Silovsky, J., Zdansky, J., Cerva, P., Kroul, M., Chaloupka, J.: Czech-to-Slovak Adapted Broadcast News Transcription System. In: Proc. of Interspeech 2008, Australia, pp. 2683–2686 (2008)