Making Czech Historical Radio Archive Accessible and Searchable for Wide Public

Jan Nouza, Karel Blavka, Petr Cerva, Jindrich Zdansky, Jan Silovsky, Marek Bohac and Jan Prazak
Institute of Information Technology and Electronics, Technical University of Liberec, Studencka 2, 461 17 Liberec, Czech Republic
{jan.nouza, karel.blavka, petr.cerva, jindrich.zdansky, jan.silovsky, marek.bohac, jan.prazak}@tul.cz

Abstract—In this paper we describe a complex software platform that is being developed for the automatic transcription and indexing of the Czech Radio archive of spoken documents. The archive contains more than 100,000 hours of audio recordings covering almost ninety years of public broadcasting in the Czech Republic and former Czechoslovakia. The platform is based on modern speech processing technology and includes modules for speech, speaker and language recognition, and tools for multimodal information retrieval. The aim of the project supported by the Czech Ministry of Culture is to make the archive accessible and searchable both for researchers as well as for wide public. After the first project’s year, the key modules have been already implemented and tested on a 27,400-hour subset of the archive. A web-based full-text search engine allows for the demonstration of the project’s current state.

Index Terms—audio archive processing, spoken document transcription, speech and speaker recognition, audio search

I. INTRODUCTION

Audio archives collected by national libraries and broadcast institutions have become an important part of cultural heritage. As the first historical recordings date back to the second half of the 19th century and a massive launch of public broadcasting followed about 50 year later, the amount of audio data produced so far and stored on various media has been really huge. The main problem is not only how to preserve them for future generations but also how to access them effectively and how to pick up required pieces of information from them. Thanks to modern technologies for digital audio processing and voice-to-text conversion, efficient solutions already exist.

The first systems focused on spoken archive processing were reported in early 2000s. In the USA, the major event that accelerated the research in this field was the creation of the National Gallery of the Spoken Word [1]. A system called SpeechFind was developed to enable its automatic processing, indexing and browsing [2]. In Europe, a similar initiative named CHoral aims at getting public access to Dutch oral history collections [3]. Another example is international project MALACH that provides access to multi-lingual archive of testimonies recorded by survivors and witnesses of the Holocaust [4].

This paper presents a large applied-research project supported by the Czech Ministry of Culture. It was launched in 2011 and it 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 voice-to-text system developed specially for this purpose in our lab. Within 4 years 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 inflective nature and very large lexical inventory with millions different 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 piece of 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 one could search also for e.g. utterances spoken by a selected person, for the historically first occurrence of a given word, or for a particular pronunciation of a word. Moreover, the queries could combine various search criteria to answer, for example, a question like: What did person A say about person B within time period T? Moreover, the potential users of the system will be not only people from the Czech Radio itself, but also historians, linguists, phoneticians, communication specialists, students and anybody who is interested in media data mining.

II. RELATED WORK

One of the pioneering and most frequently cited papers in this area is that of Hansen et al [1]. It presents the development of system SpeechFind designed for audio
search in the National Gallery of the Spoken Word (NGSW). This internet-hosted gallery aims at providing access to the collection of audio documents important for the history of the USA. It includes a wide range of speech recordings in total length of some 60,000 hours, from Edison's voice to famous Neil Armstrong's words sent back from the Moon, samples of US president speeches, broadcast news recorded for several decades, etc. In the paper, the authors describe the methods that aim at converting these audio data into rich text documents that can be indexed, tagged and searched.

The core technologies of SpeechFind system include an audio stream segmentation module, an automatic speech recognition (ASR) engine, a term indexation component and a web-based information retrieval interface. As the success of the search depends mainly on the accuracy of the voice-to-text transcription, the authors focus particularly on improving the word error rate (WER) of the used ASR system. They employ the CMU Sphinx engine running with 64K-word lexicon supported by trigram language model. The acoustic part is based on hidden Markov models (HMM) trained on 200 hours of (mainly) broadcast data. On the test set made of 3.8 hours of recordings from 1950s to 2000s they report WER values ranging from 20 to 50 %. (Surprisingly, the worst results come from the latest decade.) By applying a speaker/channel adaptation technique, they were able to achieve relative WER reduction in range of 5 - 20 %.

A similar project named CHoral [3] has been running in the Netherlands. Its aim is to disclose Dutch national audio-visual archive both to researchers and wide public. The authors employ technology similar to the SpeechFind platform. More details about its implementation can be found in technical report [24]. ASR is performed by a system developed at the University of Twente. It uses vocabularies up to 64K words that can be modified for different topics and domains. On broadcast news task they report the WER in range 19 to 28 % (depending on the type of broadcast programs.) On historical parts of the audio archive, the WER drops below 50 %.

While the two above mentioned systems are focused on processing historical and contemporary collections of national archives, the MALACH project [4] is specific as it tries to preserve and disclose stories of survivors and witnesses of the Holocaust from many countries. The collection includes almost 52,000 interviews in 32 languages, a total of 116,000 hour of oral documents. The data was collected during the last two decades using a modern audio and video technology. About one tenth of them were transcribed manually and the remaining parts were processed automatically by ASR systems adapted for each language, for specific vocabulary and for individual speakers (using the already existing human-made transcriptions). The WER reported for English was about 40 % (after several adaptations steps). A similar accuracy was achieved also for the Czech part of the collection, which covers 575 hours of audio recordings. The major challenge in this project was processing of speech provided by elderly people and strongly influenced by emotions.

Before we started the current project, we worked on several closely related topics. We have been developing ASR systems for Czech language since late 1990s. In 2006 we presented the first very-large (300K+ word) system for the automatic transcription of broadcast news, which was equipped also by a data mining module [5]. Later, a more advanced system was developed and applied for the full-text search in a collection of Czech TV spoken programs [6]. Due to this experience we were able to design most of the key modules needed for the current project during its first year. Some preliminary results were published in [7]. Recently, we have developed a demo that is used for testing and getting initial feedback from potential users. At the end of 2011, it provided an access to some 70,000 archive recordings.

II. CZECH RADIO ARCHIVE OF SPOKEN DOCUMENTS

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 the second oldest broadcaster on the European continent, after BBC. Later, the company was transformed into the Czechoslovak Radio and with the split of Czechoslovakia (in 1993) the Czech Radio (CR) took over the service and the historical archive. The CR is the public broadcaster and it runs 8 nation-wide channels and 10 regional programs, recently. 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.

A. Data Characteristics

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 radio 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, industrial noise or another speech may occur in background,
- **speaking style:** read speech, planned talk, spontaneous conversation,
- **historical and social background:** contemporary language as well as archaic language from 1930s
and 1940s, politically biased language of the communist era (1945-1989), etc.

- **national language:** mainly Czech, but also Slovak (in particular before 1993).

### B. Historical Epochs

After a closer survey of the archive data, we have classified it into five 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 allows us to perform an 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 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 - 1999).** 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 2nd language in 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 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.

### TABLE I. HISTORICAL EPOCHS IN CZECH RADIO ARCHIVE AND THEIR CHARACTERISTICS

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Period</th>
<th>Amount of data (hours per day)</th>
<th>Language</th>
<th>Audio quality</th>
<th>Document transcriptions</th>
<th>Additional text resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>E0</td>
<td>2000-now</td>
<td>20</td>
<td>Czech</td>
<td>mostly CD quality</td>
<td>for major documents</td>
<td>electronic newspapers</td>
</tr>
<tr>
<td>E-1</td>
<td>1990-1999</td>
<td>5</td>
<td>Czech (Slovak)</td>
<td>compressed (mp3)</td>
<td>not available</td>
<td>electronic newspapers</td>
</tr>
<tr>
<td>E-2</td>
<td>1968-1989</td>
<td>1</td>
<td>Czech+Slovak, regime-biased</td>
<td>AM, compressed</td>
<td>not available</td>
<td>scanned newspapers</td>
</tr>
<tr>
<td>E-3</td>
<td>1945-1967</td>
<td>0.5</td>
<td>Czech+Slovak, regime-biased</td>
<td>AM, noise</td>
<td>not available</td>
<td>scanned newspapers</td>
</tr>
<tr>
<td>E-4</td>
<td>1923-1945</td>
<td>singular documents</td>
<td>archaic Czech+Slovak</td>
<td>very low</td>
<td>not available</td>
<td>not available</td>
</tr>
</tbody>
</table>
C. Epoch Processing

As mentioned earlier, the archive processing work will go against the time flow, from the present towards the past. The reason is rational. The existing transcription system has been developed for the 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 for 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 (mentioned in section VI) have proved that this way back was feasible.

IV. AUDIO ARCHIVE PROCESSING PLATFORM

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

Anybody who wants to search in the archive needs to be connected to internet. On the dedicated web page, one enters the word(s) or phrase(s), he or she is interested in, optionally sets some search constraints (e.g. time period, broadcast channel, program name, speaker name, etc) and presses the Search button. Almost 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 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 running at our institute.

The technology behind the platform is 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 (e.g. a part containing telephone talk). These change points are used to split the speech into individual utterances. For each speaker's 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, speaker or channel 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, the same happens also to some frequently used abbreviations, capital letters and punctuation are added, proper formatting is applied, etc. The final transcription together with the time markers and complementary information, such as speaker's identity, is indexed and stored in the database.

V. METHODS AND 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 on increasing the size of the operating vocabularies, creating appropriate language models 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 describe the key methods and their technical parameters. As we have completed two versions of the system during the project's first year - the initial one in March 2011 (that reported in [7]), and the most recent one from December 2011, we also mention and evaluate the main modifications and enhancements implemented so far. The impact of these already realized improvements is documented in a large-scale series of experiments described in section VI.
A. Audio Signal Processing and Parameterization

The signal of audio archive recordings is determined by several major factors. As we are interested mainly in digital spoken documents, their samples $s_i$ are result of a) speech made of word sequence $W = (w_1, w_2, \ldots, w_N)$ in language $L$, b) voice characteristics of speaker $S$, and c) parameters of recording and processing channel $C$.

$$ (s_1, s_2, \ldots, s_T) = f(W, L, S, C). $$

Given the signal, our goal is to get the most probable estimate of language $\hat{L}$, word sequence $\hat{W}$, speaker $\hat{S}$, utilizing also the information about the parameters of identified channel $\hat{C}$.

Because the data in the audio archive are stored in various formats, it is necessary to convert them into a standard format before they enter the signal processing routines. This standard has been set to 16 kHz, 16 bit, PCM WAV format. We use the popular FFmpeg tool [8] for the conversion.

In the next step, the acoustic signal is parameterized into a stream of feature vectors. These are 39-dimensional mel-frequency cepstral coefficients (MFCCs) computed every 10 ms in 25-ms long frames. Each vector is composed from 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 (CMS) technique. The final step in the parameterization process is HLDA (Heteroscedastic Linear Discriminant Analysis) transform performed by multiplying each feature vector by a $39 \times 39$ HLDA matrix determined during the acoustic model training procedure. After that, the original speech signal is represented by a sequence of vectors $X = (x_1, x_2, \ldots, x_T)$, which are employed in all the following segmentation, recognition and identification tasks.

B. Phonetic Inventory and Acoustic Model

The acoustic-phonetic inventory of spoken Czech used in our voice-to-text systems includes 40 phonemes and 8 types of noise [9]. In [10] we demonstrated that the same 48-unit set could be utilized also for Slovak language - after an appropriate phonetic mapping procedure. All the acoustic-phonetic units are represented by continuous-density hidden Markov models (HMM) trained on the database of spoken Czech.

C. Acoustic Model Training and Refinement

When the projects started, our training database contained about 120 hours of annotated speech provided by some 1000 speakers. Approximately one half of that amount was made of read speech recordings containing phonetically balanced utterances, the other half came from broadcasting. On this data we trained the March 2011 acoustic model (referred in [7]). It consisted of triphones with 2360 physical states and total number of 75520 gaussians (32 per state).

During the work on the project we have automatically processed and manually checked 150 hours of recordings coming from different radio channels and various programs. Moreover, we got access to some 600 hours of archive documents transcribed by an external media monitoring company. While the former data could be used immediately for retraining the acoustic model, the latter source required a procedure that was to check the agreement between audio and text. (The transcriptions provided by the monitoring company are not always verbatim as they usually omit repeated and unintelligible words, hesitations and cross-talks, skip over headlines, etc.) To eliminate these disagreeing parts, we performed ASR on this data and compared its results with the provided transcriptions. The segments in which no difference was found, can serve as another source of annotated training data. Those with minor differences were checked by a human supervisor who either accepted the ASR produced transcription or made a correction in it. In this cost-efficient way, almost one half of the provided transcribed audio data was ready for further use.

Before applying these data in the retraining procedure we had to ensure that speakers in the training set were well balanced. In other words, we had to set up an upper limit for the frequently occurring speakers such as news readers, editors, interviewers, etc. The limit was set up in the way that no speaker was allowed to have more than 20 minutes in the training set. As result, during the first project’s year, the size of the training database almost tripled, with 320 hours of speech from 5200 different speakers. On this data, we trained the Dec2011 acoustic model. It consists if 5562 physical states and 180864 gaussians in total. It is available in gender-independent as well as in gender-dependent (male and female) version.

Besides the triphones we have trained also a context-independent units (monophones). These find use in situations where higher speed and lower memory requirements are important, namely in the word-spotting and text alignment routines.

D. Lexicon and Language Model

The linguistic part of the LVCSR system is made of a lexicon and a language model (LM). When the project was started, the baseline lexicon contained 340K words. During the project’s first year we got access to more text corpora and we could utilize also the already transcribed archive documents. Hence, the Dec2011 lexicon could be
enlarged to reach 483K-word size. These are the most frequent lexical units (words, word-forms and multi-word collocations) 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 10 times. The lexicon is continually growing as new words get over the threshold. For most documents, this size of lexicon assures an 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 [9]. At the moment, this is not feasible because it would slow down the transcription process significantly. Yet, we plan to reach that size in 2012. It should be also noted that almost one tenth of the lexicon entries have multiple pronunciation variants with some of the alternations being specific for Czech phonology [13]. The total number of phonetic forms is 527K, currently.

The language model is probabilistic, based on N-grams. From practical reasons (mainly with respect to the very large vocabulary size), the recent system version uses bigrams. In the above mentioned 45 GB corpus of contemporary Czech texts we found 143 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.

E. Speech Decoding

Decoding is the crucial part in speech recognition. Its aim is to estimate the most likely sequence of words \( \tilde{W} \) from given lexicon and language \( L \). The decoder jointly optimizes the acoustic model score (the measure that the word string fits feature vector sequence \( X \)) and language model score (the word sequence probability):

\[
\tilde{W} = \arg\max_{W \in L} \left[ P(X | W) \cdot P(W) \right]. \tag{2}
\]

The solution to the above equation can be effectively searched by Viterbi algorithm, which is a standard in modern speech recognition systems. As a byproduct we can obtain also vector \( T = (t_1 \ldots t_n) \), where \( t_n \) is the starting time of word \( w_n \) (measured in multiples of frame periods). It should be mentioned that non-speech events, such as silence and noises, are modeled in the same way as words (except for LM values) and they are part of \( \tilde{W} \). Therefore, the end time of \( w_n \) is simply \( t_{n+1} \). Moreover, for each word, the decoder also returns its pronunciation that fits best the given utterance. In this way we obtain not only the text transcription of the spoken document, but also the phonetic one, as well as the precise time position of each word and noise. This is the essential feature necessary for later indexation and search.

The decoder has been designed to manage vocabularies up to 1 million distinct words. When it operates with the current 483k-word lexicon, it is able to do it in almost real time on recent high-end PCs, at least for clean speech. For larger vocabularies, more spontaneous and noisy utterances, the processing time gets longer.

When required, the audio transcription process can run 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 Constrained Maximum Likelihood Linear Regression (CMLLR) technique is performed on this data and immediately applied in the second pass, in which the full lexicon is employed [11]. This allows for 10 to 15 % relative improvement of transcription accuracy, although the total processing time is about 1.7 times longer.

F. Audio and Text Alignment

The decoder can be employed also in the situation when there exist audio recordings that have been already converted into texts, e.g. by professional transcribers. In this case, we do not need to run the complete recognition process. Instead, the decoder solves a much simpler task, known as forced alignment, where eq. (2) transforms into

\[
T = \arg\max_i \left[ P(X | W(i)) \right]. \tag{3}
\]

In practice, however, the task is often complicated by the fact that human-made transcriptions are not always verbatim, especially for spontaneous speech. Also it may happen that some parts of a broadcast program, such as headlines, advertisement, or some words (repetitions or hesitations) are omitted in text. For this purpose, a more flexible approach has been proposed in [12].

G. Speaker Segmentation and Recognition

Audio signal segmentation is another crucial task in a spoken archive transcription system. Its aim is to split the acoustic signal into speech and non-speech sections, and after that, to detect instants where speakers change.

The former sub-task is solved by a speech activity detector (SAD), which is based on signal energy measurements (to identify silence and low-noise parts) and a gaussian mixture model (GMM) classifier trained to distinguish between speech and various types of noise (e.g. music, industrial noise, etc). The latter sub-task is more complex. We have to determine how many speakers occur in the document and where are the change-points between their utterances. The most frequently used method is based on Bayesian Information Criterion (BIC, [14]), which processes feature vector sequence \( X \) and searches for the most prominent changes between neighboring chunks. When the change-points are found, another method performs speaker clustering, whose aim it to group segments for the same speaker together. The whole process is called speaker diarization and its implementation in the APAP system is described in [15]).

In the following step, speaker identification and verification is performed. It is based on discriminatively trained GMMs of all speakers in our database and operates in two stages denoted as identification and verification. In the first one, an ordered list of most likely speakers is created. Then, the score for the best speaker is compared with that achieved with a Universal Background Model (UBM) [16] to avoid false acceptance cases.
The information about the speaker identity (or at least about his or her gender) has twofold usage. First, it is stored together with text as part of the rich transcription format, and provides another searchable feature. Second, it may be employed for speaker-dependent speech recognition, which usually yields another 10 - 15% relative improvement in word accuracy.

**H. Language Identification**

The task for the LID module in the APAP is rather restricted: For each speech segment, it should decide whether it is spoken in Czech or in Slovak. Both the languages come from the same West Slavic tongue branch, they are acoustically similar and share many common words. Yet, there are differences in orthography, morphology and grammar [10].

As mentioned in section V.B, we use the same phonetic inventory for speech recognition in Czech as well in Slovak. This implies that the LID technique based on phonotactics (i.e. frequency analysis of phoneme sequences, [17]) could be adopted easily. We tried it with fair results. Though, in our case, even more precise and more natural implementation of the LID task can be built upon the existing speech recognizer. We simply let a spoken segment recognize with the Czech lexical module (the lexicon and language model) and with its Slovak counterpart, and compare which of the two likelihood scores is higher.

In this approach, we may utilize only a small sub-lexicon with 20k – 50k words to reduce the computation cost and time, while keeping the LID accuracy above 95%. Moreover, this fast pre-recognition step can be immediately used as the first step in the two-pass decoding scheme mentioned in section V.E. This approach is intensively tested, recently. The LID module itself is still under development.

**I. Database and Search**

The modules mentioned previously generate results in form similar to that displayed in Table II. For each document, this data is stored in the database. In the recently build demo version, we decided for the MySQL [18] solution as it optimally fits to the type and size of data. Every word and every speaker occurrence is indexed using the Sphinx platform [19], which proved to be fast and flexible enough both for the indexation as well as for the search tasks.

**J. User Interface**

The ultimate goal of the project is to allow for public search in the transcribed archive documents. A user just

---

The screenshot shows one of the documents containing searched word “Brusel*”. In this case it was spoken by Václav Klaus in talk show ‘Dvacet minut’ on 2009-06-23. The found word occurs in 73230th millisecond of the show – see Table II.

---

**Figure 3. Web interface for archive search.**
needs to be connected to internet and have a properly set up web browser that supports audio playback. Currently, one can try the demo version of the search interface displayed in Fig. 3 and available at [20].

The user has several choices to formulate a query. One can search for a word (or its part using the * convention), a phrase or multiple words. Furthermore, one 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 document topic can be retrieved from the archive, too.

VI. PERFORMANCE EVALUATION

In the first project year, we have been focusing mainly on setting up the initial version of the audio archive platform and on processing the documents from Epoch E0. As explained in section III, the data from this period are well covered by the lexicon and language model created during our previous research [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 V.F. An additional benefit is that we can evaluate the performance of the recognition system by comparing its output to the official transcriptions and obtain an objective measure known as a Word Error Rate. To get it, we need to align the reference sequence $W$ (consisting of $N$ words) with recognizer output $\tilde{W}$ via the Minimum Editing Distance (MED) algorithm, and counting three types of errors: substitutions, insertions and deletions. The WER is calculated as:

$$WER = \frac{N_{sub} + N_{ins} + N_{del}}{N} \times 100\% .$$ (4)

For experimental evaluation, we set aside 580 minutes (almost 10 hours) of radio programs from 2010, which were excluded from acoustic and language model training. The data represent 3 types of spoken documents: a) speech of professional news speakers recorded in studio (140 minutes), b) parts of news recorded out-of-studio (210 minutes containing speech in more or less noisy conditions, including also field interviews), and c) talk shows focused mainly on politics, daily life and science (230 minutes of mainly two-part conversations). To measure the real accuracy of the speech recognition module, these test documents were transcribed very precisely (including repeated words and phrases), and after that manually cut into segments spoken by individual speakers.

Table III summarizes the main figures obtained on the test data when the initial baseline system was set up in March 2011. The system worked with 339,820-word lexicon, an acoustic model trained on 120 hours of previously annotated speech and a bigram language model learned mainly from newspaper texts in the 1990 to 2010 period. We can observe that the WER values are very good for professional studio speech. However, they become significantly worse for out-of-studio recordings, and, in particular, for conversational and spontaneous speech. In the latter cases, not only the WER increases (partly due to a higher OOV rate), but the same is true also for the processing time measured as the real-time factor on a PC with Intel I7 3.40 GHz processor.

During 2011, the transcription module has undergone major improvements. First, the lexicon was enlarged by adding the most frequent OOVs occurring in the already transcribed archive documents. Its recent size is 483,529 entries with the total number of 527,323 phonetic variants. The newly available transcriptions helped to train a more fitting language model, too. As mentioned in section V.C, the latest acoustic model (December 2011) was trained on 320 hours of speech, from which 200 was of the already annotated archive recordings. The impact of the new lexicon and new acoustic and language models is documented in Table IV.

We tested also the two-pass decoding scheme proposed in [11] and described in section V.E. The optimal results were achieved when a scaled-down 50k lexicon was used in the first pass, whose phonetic output enabled for unsupervised CMLLR adaptation of the acoustic model employed in the second pass. As shown in Table V, this 2-pass adaptation scheme yields about 5 - 10 % relative WER reduction. Evidently, it is paid by significantly longer computation time. However, the implementation of this two-pass scheme has not been optimized for speed, yet.

Using the test set we have also evaluated, which types of transcription errors occur most frequently. In many cases, it is mainly very short words, like conjunctions and

<table>
<thead>
<tr>
<th>TABLE III. TRANSCRIPTION PERFORMANCE OF BASELINE SYSTEM (LEXICON 340K, INITIAL AM AND LM; TEST DATA: 10 HOURS OF E0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document type</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Read news from studio</td>
</tr>
<tr>
<td>News recorded out of studio</td>
</tr>
<tr>
<td>Talk shows</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE IV. PERFORMANCE OF SYSTEM Refined on DOMAI N DATA (LEXICON 483K, ADAPTED AM, ADAPTED LM, 1-PASS MODE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document type</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Read news from studio</td>
</tr>
<tr>
<td>News recorded out of studio</td>
</tr>
<tr>
<td>Talk shows</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE V. PERFORMANCE OF 2-PASS SYSTEM WITH CMLLR ADAPTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document type</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Read news from studio</td>
</tr>
<tr>
<td>News recorded out of studio</td>
</tr>
<tr>
<td>Talk shows</td>
</tr>
</tbody>
</table>
prepositions that are either omitted or inserted. (Let us mention that Czech has 8 single-phoneme words, all occurring among the 40 top frequent lexemes.) There is also a large number of homophones in Czech, many of them making pairs which not only have the same pronunciation but also the same meaning. A subtle difference in orthography (e.g. ‘bili’ vs. ‘bily’ or ‘dali’ vs. ‘daly’) has only a grammatical function. Both these two types of errors are rather difficult to be fixed. However, they have almost no impact on the search task, i.e. on the primary function of the archive processing and accessing platform.

Obviously, more transcription errors and more serious word confusions appear in documents containing conversational speech. Here, the vast majority of errors are caused by such phenomena like hesitation, repeated words or their parts, colloquial expressions, sloppy pronunciation and out-of-vocabulary words. To cope with these challenges, the transcription system will have to learn repetitively from the already processed data, which should result in a continuously growing lexicion (with more words and richer pronunciation alternatives) and a more appropriate spoken language statistic model.

When starting the work on epoch E0, we believed that there would be no serious problems with the audio signal quality. We knew that for the last decade the MP3 format was the standard storage format used in the Czech Radio archive (particularly, in its publicly accessible part). This is true but we found that some stations, some studios (namely the regional ones), or external contributors used different compression rates. There was no exception to hear spoken documents sampled at 44 kHz, 16 bit stereo compressed in 1:20 rate or even worse. When these documents were down-sampled to 16 kHz (the default frequency for the ASR module) their audible quality was similar to that of 16 kHz 20 kbit/s (or lower) coded signal. Obviously, such severe compression had a considerable impact on speech recognition. To quantify it more exactly, we have conducted an experiment. A test set containing 30 minutes of clean speech from various radio programs was recorded directly from a wide-band broadcast channel and stored in un-compressed 16 kHz WAV format. Then we applied a professional MP3 converter (built in Audition software) with varying bit rate, starting from 64 kbit/s and going as far as 16 kbit/s. These data were processed by the standard ASR procedures. The results represented by WER values are shown in Fig. 4. We can see that the compression with bit rates lower than 32 kbit/s has a significant impact on the transcription accuracy. Similar results were reported also in [25].

We plan to try to compensate this type of signal degradation in two different ways: The first one will utilize an acoustic model trained on the data that passed through an MP3 decoder and hence it will better match the compressed audio files. The other approach will consist in applying a speaker and channel adaptation technique for each utterance and running the second recognition pass as described in section V.E. This has been already applied for compressed telephone speech.

Table VII shows, political debates, programs focused on science, culture, sport or music, read feuilletons, etc. Some relevant statistics on the processed data are shown in Table VI. In order to manage that large volume of audio documents, we have developed a distributed platform that allows us to utilize several computer classrooms at the university where the computers work on the transcription and indexation tasks during off-class hours.

In 2011 the main goal of the project was to build the first version of the system and to employ it for automatic processing and indexing of epoch E0 data. We have been also given a small amount of sample recordings from the 1960-2000 periods (one broadcast news program per year). We have done very preliminary experiments with these data just to observe where we should expect the main challenges. Surprisingly, in spite of the lower signal quality (mainly a narrower frequency band and noise), the transcription accuracy did not drop too much. One reason

\begin{table}[h]
\centering
\caption{Project Output After First Year}
\begin{tabular}{|l|l|}
\hline
\textbf{Transcribed epoch} & \textbf{E0} \\
\hline
\textbf{Number of transcribed program types} & 117 (from 18 stations) \\
\hline
\textbf{Number of transcribed documents} & 71,585 \\
\hline
\textbf{Number of indexed speech segments} & 3,033,817 \\
\hline
\textbf{Number of indexed words} & 166,890,055 \\
\hline
\textbf{Number of indexed unique speakers} & 5,101 \\
\hline
\textbf{Size of transcribed audio} & 27.401 hours \\
\hline
\end{tabular}
\end{table}
is that speech broadcasted in previous decades was slower, uttered with careful diction and pronunciation. Hence, the main source of transcription errors seems to be a larger number of out-of-vocabulary words.

We have already started complementary research works, which should help us in processing data from the earlier epochs. These activities include, among other, also 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. Our recently achieved results seem to be promising [22].

Even though most of the routine work is done by computers, at least several hundreds of audio documents must be checked and edited manually. This is necessary especially when we start to process new types and new domains of archive programs or when we move into older time epochs. We collaborate with the Department of Czech Language at our university and employ their students. To make the editing work efficient, we have developed a special software tool similar to Transcriber [23]. Our version has some advanced features, such as a special track for phonetic transcription, easier and faster navigation in audio, text and phonetic tracks, a link to a special track for phonetic transcription, easier and faster background and immediately re-align edited texts, etc.

Though the current project is focused primarily 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]. When an archive document includes a video, the visual track is played synchronously with the spoken content. The same feature is available also in the above mentioned transcription editor.

ACKNOWLEDGMENT

This research work was supported by project no. DF11P01OVV013 provided by Czech Ministry of Culture within program NAKI.

REFERENCES


Jan Nouza received his M.Sc. and Ph.D. degrees at the Czech Technical University (Faculty of Electrical Engineering) in Prague in 1981 and 1986, respectively. Since 1987 he has been teaching and conducting research at the Technical University of Liberec. In 1999 he became full professor. His research focuses mainly on speech recognition and voice technology applications (voice-to-text conversion, dictation, broadcast data processing and design of voice-operated tools for handicapped persons). He is the head of SpeechLab group at the Institute of Information Technology and Electronics. Prof. Nouza is member of IEEE and ISCA.

Karel Blavka received his M.Sc. degree at the Technical University of Liberec in 2011. After graduation he joined the SpeechLab team, where he focuses mainly on multimedia data processing, indexation and search.

Petr Cerva received his M.Sc. and Ph.D. degrees at the Technical University of Liberec in 2004 and 2007, respectively. Since 2007 he has been an assistant professor at the Technical University of Liberec. His research work is focused mainly on speaker adaptation and speech recognition.

Jindrich Zdansky received his M.Sc. at the Czech Technical University in Prague in 2002, and Ph.D. degree at the Technical University of Liberec in 2005. He has been an assistant professor at the Technical University of Liberec since 2005. His research work in speech recognition is focused both on algorithm design as well as on implementation issues.

Jan Silovsky received his M.Sc. at the Technical University of Liberec in 2006. In 2011, at the same university, he completed his Ph.D. thesis on speaker recognition and verification.

Jan Prazak received his M.Sc. at the Technical University of Liberec in 2010. He is a Ph.D. student at the same institution, working mainly on speech segmentation and speaker clustering tasks.

Marek Bohac received his M.Sc. at the Technical University of Liberec in 2010. He is a Ph.D. student at the same institution, working mainly on language and lexicon issues associated with speech recognition tasks.