Speech-To-Text Technology to Transcribe and Disclose 100,000+ Hours of Bilingual Documents from Historical Czech and Czechoslovak Radio Archive

Jan Nouza, Petr Cerva, Jindrich Zdansky, Karel Blavka, Marek Bohac, Jan Silovsky, Josef Chaloupka, Michaela Kucharova, Ladislav Seps, Jiri Malek, Michal Rott

SpeechLab, Institute of Information Technology and Electronics
Technical University of Liberec, 461 17 Liberec, Czech Republic

{jan.nouza, petr.cerva, jindrich.zdansky, karel.blavka, marek.bohac, jan.silovsky, josef.chaloupka, michaela.kucharova, ladislav.seps, jiri.malek, michal.rott}@tul.cz

Abstract

In this paper, we present the outcome of a 4-year project whose ultimate goal is to develop a complex platform that can transcribe, index and make searchable the historical archive of Czech and Czechoslovak Radio. The archive covers 90 years of public broadcasting and contains hundreds of thousands audio documents. The developed modular platform employs our LVCSR system that has to cope with 2 related languages: Czech and Slovak. Furthermore, it must deal with audio files of varying quality (e.g. recordings originally stored on matrices or tapes, data passed through analog and digital telephone lines, speech recorded during parliament or court sessions, etc.) The system includes speaker and language identification modules, a narrow-band signal detector, a music/song detector, and several other components to enhance transcription accuracy and provide support for multi-optional search. We evaluate the performance on broadcast news test sets grouped according to decades. We show that after acoustic and language model adaptation WER values are in range 8-14% and do not differ much since 1960s to present. We report also results achieved on other types of documents (e.g. talk shows, political debates, public speeches, etc), where the WER is higher but still acceptable for most search tasks.

Index Terms: spoken archive, speech recognition, speaker recognition, language identification, spoken term search

1. Introduction

For four years (2011 to 2014) we have been working on an ambitious project supported by the Czech Ministry of Culture whose aim is to disclose the historical audio archive of Czech Radio (and its predecessor Czechoslovak Radio) to researchers (historians, media experts, linguists, phoneticians) as well as to public [1]. The archive contains several hundreds of thousands of spoken documents (more than 100,000 hours) and covers 90 years of broadcasting in Czechia and Czechoslovakia. During the previous decade, the archive was digitized and now, within the project, it is being transcribed and indexed. By the end of 2014, all the available archive data will be processed and made accessible for search and listening via a dedicated web portal.

Within the project we have designed and implemented a complex platform that includes an LVCSR engine operating in two languages (Czech and Slovak), a language and speaker identification module, a channel type detector and a post-processing unit. This modular system is able to process an arbitrary archive document, find speech, segment it into parts spoken by a single speaker, decide which language and which type of channel should be used by the LVCSR system, and produce transcription whose text can be displayed, read, indexed and stored in a large database. A special web application allows for searching in it. One can search for words or phrases, and queries can be constrained by a speaker name, language, time period, or program name.

In this paper, we introduce the platform and present its modules and functionalities. We focus mainly on those components that are unique to this project and that make our solution different from those designed by other research teams.

2. Related and Previous Work

The first speech transcription systems focused on audio/video archive processing were reported in early 2000s. In the USA, system SpeechFind [2] was developed to enable automatic processing, indexing and browsing of the records in the National Gallery of the Spoken Word [3]. French researchers from LIMSI adapted their broadcast news transcription system to process historic TV and radio documents collected by the Institute National l'Audiovisuel [4]. In the Netherlands, project CHoral was aimed at getting public access to Dutch oral history archive [5]. The MALACH [6] has been a well-known example of a large international initiative whose goal was to collect and transcribe testimonies recorded by survivors and witnesses of the Holocaust. As this collection is multi-lingual, the employed speech processing technology had to be adapted to most of the 32 languages used by the speakers, including, e.g. Slavic languages [7] or Hungarian [8].

The processing of historical spoken archives is challenging because the audio quality of older recordings is often very low compared to recent standards [2]. Moreover, each language evolves in time, and the lexicon, pronunciation and speaking style can differ from one historical epoch to another. In the papers mentioned above, the reported WER values were in range 20-40%, depending on the data and their characteristics. At the time of publishing, these figures were considered acceptable, namely for search tasks. Nowadays, the accuracy could be further improved thanks to recent advances in acoustic and language modeling (e.g. neural networks).

Our project has a similar goal: To develop the technology that can transcribe and index a large historical audio archive. The main difference is that we have to cope with 2 languages, Czech and Slovak, that can be arbitrarily mixed even within a single document. Both were the official languages in former Czechoslovakia and in broadcasting every speaker used his or her native tongue. (This happens even now in recent Czech Radio programs.) Another specific feature of our system is that the used lexicons are really large (559K words for Czech and 308K words for Slovak) due to the inflectional nature of the languages. Moreover, the lexicons and language models (LM) have been built not only for contemporary speech but also for that of previous historical epochs.

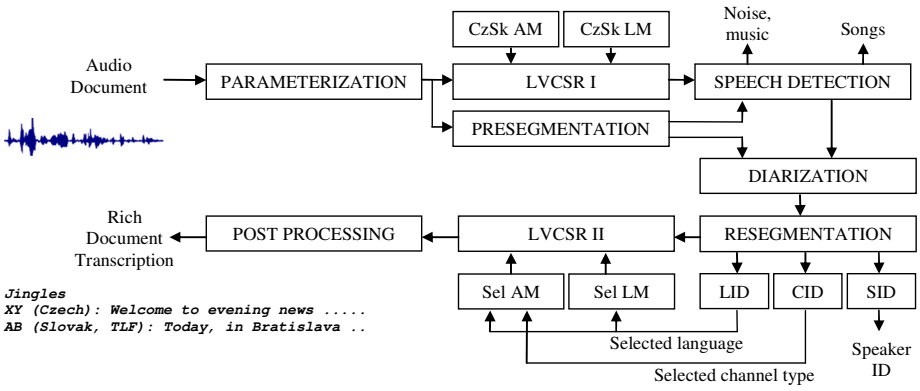

Figure 1: *Diagram of bilingual transcription system with two-pass LVCSR module and modules for language identification (LID), speaker identification (SID) and channel type identification (CID)*

3. Historical Audio Archive

The data in the archive represent 90 years of broadcasting in Czechoslovakia and in the Czech Republic. When founded in 1923, the Czech Radiojournal company was the second oldest broadcaster in Europe (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.

There are hundreds of thousands records in the archive. They represent various types of spoken documents, namely:

- **Broadcast news (BN):** these shows (typically 30 min. long) are available for every day since 1968 till today;
- **Discussions, political debates, talk shows**: contain a lot of multi-speaker spontaneous speech;
- **Professional, scientific and educational documents**: with read, planned and conversational utterances dealing with quite specific rather than general topics;
- **Speeches:** such as addresses read by presidents, political leaders, members of parliaments, judges, etc., often recorded in large meeting rooms;
- **Regional programs:** documents focused on local news and topics, produced by regional stations of CR.

The oldest documents were recorded in 1920s, but the major volume of data comes from the last 5 decades.

In [1] we provide a more detailed description of the archive and split the 90-year period into 5 time epochs, each characterized by its specific historic, social, linguistic and audio recording features. If we look at the archive from the speech technology design point of view, the most essential milestone was year 1990. This was the year, when the communist regime ruling for 45 years in Czechoslovakia was replaced by a democratic system. This had a significant impact on the language of media as it meant the end of propaganda style in reporting. Moreover, in 1990, the first electronic versions of newspapers appeared, followed by internet news that occurred a few years after. Both the sources have made the development of lexicons and LMs for ASR fast and efficient. Unfortunately, for the period prior 1990 no digitized texts had been available and we had to search alternative resources when preparing the lexicons and LMs for the earlier epochs (see section 4.3).

During the 4-year works on the project, a small portion of the data was manually transcribed. We have chosen sample recordings from different periods and from various types of programs. Their total duration was 155 hours. The largest part (100 hours) served for AM and LM adaptation, 30 hours was reserved for development purposes and 25 hours for testing.

4. Transcription System and its Modules

We have been building BN transcription systems applicable to Czech and Slovak languages since 2005 [9]. The significantly enhanced version used in this project is based on the modular platform operating with data flow depicted in Fig.1.

4.1. Two-pass system

Each audio file is processed within two passes. Unlike many other 2-pass systems, where the first pass is used mainly for fast pre-selection of hypotheses and for acoustic adaptation, in our case, the first pass output serves more purposes, namely:

- to support language identification (Czech or Slovak),
- to fine-tune segmentation by synchronizing the output from the change point detector with word boundaries,
- to provide additional features that are used for separating true speech from speech-like audio parts, like songs, and
- to enable speaker/channel adaptation.

4.2. Acoustic processing and modeling

Because the archive data are stored in various formats, they are converted into 16 kHz, 16 bit, PCM WAV, first. Next, the signal is parameterized into a stream of 39 MFCC feature vectors. These are normalized by the cepstral mean subtraction technique applied within a 2-second-long moving window and after that they pass the HLDA transform and enter the decoder. The raw MFCC features are utilized by the modules performing segmentation and speaker/channel identification.

The pronunciation dictionary for Czech ASR uses 41 phonemes. Slovak has 48 phonemes, but we have shown that the specific ones can be mapped on the closest Czech ones with a negligible impact on recognition accuracy [10]. This allows us to use a single phonetic set for both languages.

The baseline AMs are based on GMMs (with 32 mixtures) used for modeling triphone state output pdfs. The Czech AM has been trained on 320 hours of (mainly broadcast) data. For Slovak, 107 hours were available. In the first pass, we utilize a mixed Czecho-Slovak (CzSk) AM trained on 120 hours of Czech and 107 hours of Slovak.

During the project, the initial AMs were used to transcribe and manually correct the training, development and test sets extracted from the archive data (mentioned in section 3). The former set (100 hours) was added to the existing training database and new AMs for Czech and Slovak were trained. The former has 4364 physical states, the latter 3856 ones.

4.3. Lexicons and language models

At the beginning of the project, we worked with the lexicons and LMs created from the texts published after 1990. These performed well for the spoken documents of the same era. However, they were not appropriate for the older archive data because of the reasons mentioned in section 3. Therefore, we have created several sets of lexicons and LMs.

4.3.1. Lexicons and LMs for contemporary language

Both Czech and Slovak are highly inflective languages and large text corpora are needed to built lexicons and LMs with a relevant coverage rate. For Czech we have collected about 13 GB texts published after 1990 and created a lexicon with 551K most frequent items (words, word-forms and multi-words). For Slovak, the lexicon is smaller (due to a smaller, 8 GB corpus), and its size is 303K words. The LMs are based on bigrams. As both the lexicons contain several thousands of multi-word expressions (frequently collocated word strings), a large part of bigrams covers sequences that are three-, four-, five- or even six-word long. The unseen bigrams are backed-off by the Kneser-Ney smoothing technique.

4.3.2. Lexicons and LMs for historic language

Initial ASR experiments done with the contemporary lexicons and historical archive documents showed quite large OOV rates (> 3 %) and worse performance due to inadequate LMs (see section 6). Therefore, we had to search for resources that would allow us to adapt the linguistic part of the system to the to previous historical epochs. As no suitable data had been available in electronic form, we had to find alternative sources and convert them into text, first. It took us almost 2 years and the work is described in [12]. We have utilized namely:

- scanned and OCRed historical newspapers (1945-1989),
- parliament steno-notes (1918-1989),
- subtitles from various TV retro programs, and
- manual transcripts of training archive set (100 hours).

These data (about 1 GB of texts) were used to identify the most frequent OOV words, which were added to the lexicons, and for adapting the LMs. The historic version of the Czech lexicon has 559K and the Slovak one 308K words. These lexicons and corresponding LMs were used iteratively for the transcription of the BN documents from 1968-1989. When the WER measured on the dev_set got below 13%, we added the ASR generated BN transcriptions to the LM training set. The addition of this 182 MB in-domain (though unchecked) text helped to reduce the WER to 11.8% [12].

4.3.3. Lexicon and LM for first pass decoding

Language identification (LID) is done after the first pass. Therefore, at that level, we use a special lexicon made by mixing the *L* most frequent Czech words with the same amount of the Slovak ones [13]. Each word in this bilingual CzSk lexicon has a label that says whether the word is Czech (CZ), Slovak (SK) or common (COM). The corresponding bigram LM is computed on the mix of Czech and Slovak texts and smoothed again by the Kneser-Ney technique.

4.4. First-pass decoding

The goal of the first pass is to get as much information on the processed data as possible. Using the aforementioned Czecho-Slovak lexicon, AM and LM we get a rough transcription of the document. Its accuracy is not high (about 30-40% WER, depending on parameter *L*), but it is sufficient enough for a) locating speech and word boundaries, b) detecting longer pauses and noises of different types, and, in particular, for c) language identification and speaker/channel adaptation. It is the last two mentioned tasks, for which parameter *L* has been optimized. The best ratio between the performance and computation time was achieved for *L*=50K (see [13] and [14]).

4.5. Segmentation and non-speech elimination

The audio segmentation task is one of the most critical ones in the whole process. If it fails, it can deteriorate the performance of the LID, SID (speaker identification) and CID (channel type detection) modules, and consequently, the overall transcription accuracy. Hence, we have put much effort to optimizing its design. It employs a modification of BINSEG algorithm [15], based on the Bayesian Information Criterion (BIC). Unlike its standard usage, in our system, potential speaker/channel change points are searched only at word boundaries found in the first pass. In [14], we proved that this approach is (about 5 times) faster and more robust compared to the classic one (when no prior information about speech content is known).

The output of the first pass includes also information about location of various kinds of noises. (Our system distinguishes between 8 noise types). The duration of them, their type, and the way how they are mixed with adjacent words, all that allows us to detect specific audio events like, jingles, songs, or headlines with background music, and eventually, to remove the unwanted ones (e.g. songs) from further processing.

4.6. Channel type detection

Telephone speech is one of the most difficult types of audio archive data for ASR, especially in situations when the signal passed one or more digital codecs and after that it was stored (once or multiple-times) in a compressed audio format (e.g. MP3). In such a case, the WER achieved with a standard wide-band-signal trained AM can be high (>40 %), even if we apply speaker/channel adaptation. Therefore, we have included a module that detects telephone speech using a pair of GMMs trained on narrow-band and wide-band archive data. When a segment is identified as the former one type, in the second pass it is transcribed with a special AM trained on 60 hours of telephone speech (one third of them being the data from the archive). This helps to reduce the WER by 20-40% relatively.

4.7. Language identification

Czech and Slovak are closely related Slavic languages. They are mutually understandable, even if their lexical inventories share only 23% identical words [13]. In this case, the LID approach based on parallel decoding with a bilingual CzSk lexicon and LM works best. An utterance is claimed Czech if the number of words with CZ label is higher than those with the SK one, and vice versa. In [13] we present the method and show that the LID error rate gets close to 1% when *L*=50K.

4.8. Diarization and speaker identification

The main goal of the diarization component is to cluster the speech segments determined by the segmentation module into

Table 1. *Broadcast news test data grouped to decades with baseline results achieved with GMM-HMM system and contemporary lexicon and LM*

Decade	# Words	Length [min]	WER [%]	OOV [%]
1960s	7239	72	18.0	3.7
1970s	16333	140	18.9	3.5
1980s	27209	225	18.6	2.8
1990s	21325	167	21.5	1.9
2000s	25877	180	11.6	1.1
2010s	18012	123	17.0	0.9

groups belonging to single speakers. We use a method based on i-vectors and cosine distance scoring proposed in [16]. The clustering is beneficial mainly for the speaker identification and adaptation tasks, providing them with more data [14]. The SID module employs again i-vector representation combined with probabilistic linear discriminant analysis (PLDA) [17]. Recently, the list of speakers include mainly Radio employees (news readers, correspondents, reporters, program hosts, etc), whose voices often occurred in broadcasting.

4.9. Second-pass decoding

In the 2nd pass, each segment is decoded as a separate piece of signal. The LVCSR system gets information on the language and channel type, and selects the appropriate AM (Czech, Slovak or the telephone one) and LM. Moreover, the system utilizes the phonetic transcription from the first pass to adapt features using the cluster based global CMLLR approach. This adaptation scheme yields about 20% relative reduction of WER in multi-speaker documents like, broadcast news [14].

5. Information Retrieval System

The transcriptions are stored in a large database that contains information about each document and its segments (language, speaker, channel, signal quality, etc). For each word we store and index its occurrence time, pronunciation and a confidence score). All these data are maintained by two open-source tools, MySQL [18] for the database management, and Sphinx platform [19] for indexation and search.

Search is possible via a web portal currently available at [20]. A user can search for a word, a phrase, a combination of them, or word parts (using the * convention). Furthermore, he or she can restrict the query to a specific speaker, broadcast station, language, program name or time period. The system responds by displaying a list of files ordered according to the criteria that can be individually set and may take into account the number of found terms, their mutual distance, confidence score, etc. By moving a mouse cursor over the displayed time axis with markers, one can get a quick overview of the text fragments containing the searched item. By clicking on the selected marker, the web application switches to the Play mode and the document starts to be played from that position. The user can click on any place on the axis or on any place in the text to listen to that part of document. More details about the search application and its design can be found in [21].

6. System Evaluation and Enhancement

During the works on the project we have created a large set of transcribed documents that have been set aside for testing. Its largest part is made of several tens of complete BN programs representing 6 decades. Their size is shown in Table 1 together with the results from initial tests. In these, we employed the AMs and LMs trained on the spoken and written data from 1990-2013 period. Let us notice that the OOV rates for the 3 oldest decades are quite large and so are the WER values.

Table 2. *The same data used in tests with lexicons and LMs created for 2 epochs (before and after 1990) and 3 approaches to acoustic modeling*

	WER [%]			OOV
Decade	GMM	GMM Adapted	DNN	[%]
1960s	14.0	11.5	10.9	1.4
1970s	10.7	9.6	9.1	1.3
1980s	11.8	9.9	9.6	1.5
1990s	15.7	11.5	10.6	1.2
2000s	10.1	8.7	8.1	1.1
2010s	14.3	13.6	13.2	0.9

Therefore, we put much effort to creating of lexicons and LMs that better fit the older epochs (see section 4.3.2), and also to adapting the AMs to historical audio data (section 4.2). In Table 2 we compare the results achieved for three methods of historical audio acoustic modeling: one based on standard GMMs, the other that employs second pass with cluster-based CMLLR adaptation (section 4.8), and the third based on deep neural network (DNN) concept [11]. At the moment, the latter is available only for Czech. It has 5 hidden layers, each with 1024 neurons, and 4364 output nodes. It is evident that the DNN approach yields the lowest WERs, all in range 8-14%. The larger WER value in the most recent decade is due to the modern style of news presentation with a lot of background music, heavy use of authentic noises, VoIP speech, etc. The same phenomenon was observed also by other authors.

We have evaluated transcription accuracy also on other types of documents. In Table 3, we present results for some of them: a) talks with distinguished scientists), b) political debates with a host and multiple guests, and c) presidential speeches from 1940s (originally recorded on matrices). Due to limited space, we compare only the WERs from the 2-pass GMM-based system with that using the DNN. The latter proved its robustness especially on the oldest audio data, though in this case, the achieved WER is still very high.

Table 3. *Results achieved on other types of documents*

		WER [%]		OOV
Program type	# Words	GMM Adapted	DNN	[%]
Science talk	22104	23.7	23.4	2.7
Political debate	46871	19.4	18.4	1.5
Speeches 1940s	894	68.2	54.6	1.9

7. Conclusions

We present a complex modular platform that employs speech technology to disclose a large historical archive of spoken documents recorded during the last 90 years. So far, the system has been used to transcribe 83,000 hours of data and to index more than 400 million words. The transcriptions have been already utilized in studies focused e.g., on the broadcast language evolution [22], or on pronunciation issues [23].

8. Acknowledgements

This research work was supported by Czech Ministry of Culture (project no. DF11P01OVV013 in program NAKI).

9. References

[1] Nouza, J., Blavka, K., Cerva, P., Zdansky, J., Silovsky, J., Bohac, M., Prazak J.: Making Czech Historical Radio Archive Accessible and Searchable for Wide Public. Journal of Multimedia, 7(2), pp. 159 - 169. 2012

[2] Hansen, J.H.L. et al. SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. IEEE Trans. on Speech and Audio Processing, vol.13, no.5, pp. 712- 730, 2005

[3] Hansen, J.H.L., Deller, J, & Seadle, M. Engineering challenges in the creation of a National Gallery of the Spoken Word: Transcript-free search of audio archives. In Proc. IEEE ACM Joint Conf. Digital Libraries, pp. 235–236, 2001.

[4] Barras, C., Allauzen, A., Lamel, L., & Gauvain, J. L.. Transcribing audio-video archives. In ICASSP 2002, pp. I-13-16, 2002

[5] Ordelman, R. J. F., de Jong, F. M. G., Huijbregts, M. A. H. and van Leeuwen, D. A.: Robust audio indexing for Dutch spoken-word collections. In 16th Int. Conf. of the Association for History and Computing. Humanities, Computers and Cultural Heritage. Amsterdam, pp. 215-223, 2005

[6] Byrne W. et al.: Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Trans. Speech Audio Process., vol. 12, no.4, pp. 420–435, 2004.

[7] Psutka, J. et al.: Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project. In Interspeech, pp. 1349–1352, 2005.

[8] Mihajlik, P., Fegyó, T., Németh, B., Tüske, Z., & Trón, V. Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages–Hungarian ASR for the MALACH Project. In Text, Speech and Dialogue. Springer Berlin Heidelberg, pp. 342-349, 2007

[9] Nouza, J., Zdansky, J., David, P., Cerva, P.,Kolorenc, J., Nejedlova, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In Interspeech, pp. 1681-1684, 2005

[10] Cerva, P., Nouza, J., & Silovsky, J.: Study on Cross-lingual Adaptation of a Czech LVCSR System towards Slovak. In Analysis of Verbal and Nonverbal Communication and Enactment. Springer Berlin Heidelberg, pp. 81-87, 2011.

[11] Dahl, G. E., Yu, D., Deng, L, Acero, A: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 30-42, 2012.

[12] Chaloupka, J., Nouza, J., Kucharova, M.: Using Various Types of Multimedia Resources to Train System for Automatic Transcription of Czech Historical Oral Archives. In ICIAP, Springer Berlin Heidelberg, pp. 228-237, 2013

[13] Nouza, J., Cerva, P., Silovsky, J.: Dealing with Bilingualism in Automatic Transcription of Historical Archive of Czech Radio. In ICIAP, Springer Berlin Heidelberg, pp. 238-246, 2013.

[14] Cerva, Petr, et al.: Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives. Speech Communication 55(10), pp. 1033-1046, 2013

[15] Zdansky, J: BINSEG: An efficient speaker-based segmentation technique. In Interspeech, pp. 2182–2185, 2006

[16] Silovsky, J., Prazak, J.: Speaker Diarization of Broadcast Streams using Two-stage Clustering based on I-vectors and Cosine Distance Scoring. In ICASSP, pp. 4193–4196, 2012

[17] Silovsky, J., Nouza, J., Kucharova M.: Search for Speaker Identity in Historical Oral Archives. Multimedia Tools and Applications. Springer, 2014

[18] MySQL platform available at http://www.mysql.com/

[19] SPHINX platform available at http://sphinxsearch.com/

[20] Search system available at https://demo.ite.tul.cz/cro/

[21] Nouza, J. et al: Large-scale processing, indexing and search system for Czech audio-visual cultural heritage archives. In: 14th International Workshop on Multimedia Signal Processing (MMSP), IEEE , pp. 337-342, 2012

[22] Kucharova, et al. On the Quantitative and Qualitative Speech Changes of the Czech Radio Broadcasts News within Years 1969–2005. In *Text, Speech, and Dialogue* Springer Berlin Heidelberg, pp. 360-368. 2013.

[23] Labus, V. Nisa, or Nysa? (in Czech). *Acta onomastica* (1), pp. 207-218.