Continual On-line Monitoring of Czech Spoken Broadcast Programs

Jan Nouza, Jindrich Zdansky, Petr Cerva, Jan Kolorenc

SpeechLab, Department of Electronics and Signal Processing
Technical University of Liberec, Halkova 6, Liberec, Czech Republic
{jan.nouza, jindrich.zdansky, petr.cerva, jan.kolorenc}@tul.cz

Abstract

In the paper we describe the development of the first practical system that performs automatic on-line monitoring of Czech broadcast stations. It is based on our own speech recognition server that operates with 300K word lexicon and 2.3 RT factor. For true on-line service, several servers are connected to the platform that controls acoustic stream segmentation, distribution of data to the servers, collection of results and production of the final transcription. We show practical results achieved on different types of broadcast programs, such as news (21 % WER), parliament debates (21 % WER) and talk-shows (34 %).

Index terms: information retrieval, broadcast speech, inflected language, distributed speech recognition

1. Introduction

Systems for automatic transcription of broadcast spoken programs (mainly news) have made significant progress during the last 5 years. This was documented in several papers presented at the last Interspeech conference in Lisbon. Word error rates achieved by the most advanced systems developed for languages, like English [1] or French [2], got below 15 % in large evaluation tests. Promising results were reported also for other major languages, like German [3], Mandarin or Arabic.

Our system [4], presented also in 2005, was the first attempt to build an automatic transcription tool for Czech. Its development was a challenging project because Czech belongs to the family of Slavic languages that are known to have very high degree of inflection. In our paper we showed that the typical vocabulary size of 64k words was inadequate for Czech. Only after the lexicon reached over 300k, the OOV rate dropped below 2 %. With our own speech decoder optimized for lexicons of that size we were able to achieve WER about 22 %.

In mid 2005, the system underwent complex trial tests in a company that focuses on information mining in Czech media. Detailed analysis of the results showed that even imperfect output from the automatic transcription of TV and radio news is profitable, because it reduces manual editing work to one half approximately. (That half was needed for correcting recognition errors and for checking some unusual words, most often foreign names.) Moreover, the company showed interest in another application of the system. Besides using it for the previously described semi-automatic transcription of main news programs, they wanted to employ its capabilities for continual monitoring of all broadcast stations. Their number has increased rapidly since 2005 when digital TV was launched in Czechia. Some of these new channels broadcast 24 hours a day and it is not feasible to monitor them in the old (human involved) way.

This set up new challenges for developing a more advanced platform for broadcast monitoring. Its main features can be summarized as follows:

- continual processing of acoustic information from TV and radio stations,
- automatic segmentation and labeling of recorded data into files associated with individual broadcast programs,
- automatic transcription of speech parts within the labeled programs,
- automatic identification of other relevant speech information, e.g. about speakers, channel quality, etc.
- text output that fits both to human-made editing as well as for automatic full-text search.
- transcription should be available with minimum delay so that on-line key-word alarm triggering can be applied.

In this paper we describe a system that fulfills most of the above requirements.

2. System overview

The broadcast monitoring system has form of a multi-level multi-processor platform. It supports parallel processing of speech data that can be distributed to a cluster of computers if these are available. Though, the whole system can run on a single machine as well. Its scheme is depicted in Fig. 1.

2.1. Modular architecture

The system is made of two main parts. The first one cares about acoustic signal acquisition from a TV/radio card. The data is recorded and labeled using the program codes and time stamps provided by broadcasters. On-line signal processing is applied to blocks of signal that are 10 s long. Within these blocks speech is parameterized and a search for speaker/channel change points is performed. When a change point is approved, the segment between this and the previous one is sent to a speech/speaker recognition server. If more servers are available, the load balancer unit optimizes their operation. The server performs speech/non-speech separation followed by speaker identification and verification. For the verified speakers, speech recognition is done with their corresponding acoustic models. For the other, fast on-line adaptation of gender-specific models is applied. The transcription of each segment consists of a sequence of words and their corresponding starting and ending times. This data is sent to the collector unit that performs time sorting, assembles the final transcription and give it form of an editable and searchable XML document.

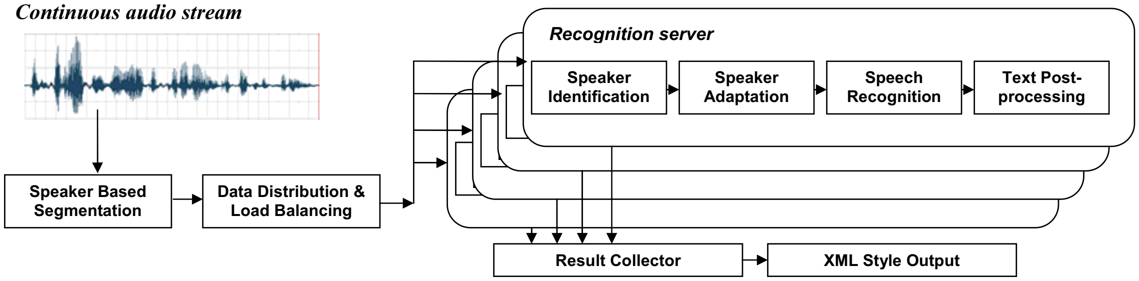


Figure 1 *Schematic diagram of the broadcast monitoring system with distributed speaker and speech recognition*

2.2. Lexicon and Language Model

Our previous research showed that the most significant advances in speech recognition of inflected languages were due to the improvements in the lexicon and language model. Both requested a large and representative text corpus to be collected.

2.2.1. Text corpus

Since 2005 our corpus increased by almost 40 %. Its recent size is 3.6 GB. Most text comes from Czech newspaper articles from the 1996 – 2005 period. Added were also official transcriptions of TV and radio news from the last 4 years and some 80 MB text collected on internet, e.g. e-magazines, novels or parliament archives. In the corpus there are 519M words now, about 2M are distinct. The corpus passed through a large cleaning process consisting in the expansion of abbreviations and numeral expressions, and orthographic standardization, which was applied to the words with variable spelling. (The impact of that was noteworthy, since it allowed us to reduce the vocabulary size by almost 30K words, i.e. by 10 % in the 300K lexicon.)

2.2.2. Lexicon

The baseline lexicon was build from those words that occurred in the corpus at least 10 times. After adding 2100 most frequent multi-word expressions, the total size of the lexicon reached 312K words. These were mapped to some 340K phonetic base-forms. The size of the lexicon was a compromise between acceptable values of OOV and WER, and computation time. These values were measured on a large test set composed of broadcast news (184 min), weather forecast (10 min), broadcast parliament sessions (118 min) and talk-shows (51 min) – altogether 6 hours (51,655 words). We compared the baseline 312K lexicon with its downscaled subsets. To learn whether a larger vocabulary could offer further improvement, we compiled also a 500K lexicon to measure at least the impact on the OOV rate. That was only negligible, as it can be seen in Table 1.

***Table 1**: OOV, WER and speed vs. lexicon size measured on a 6 hour mix of broadcast data (for baseline SI system)*

Lexicon size	OOV [%]	WER [%]	xRT (P3.2 GHz)
64K	5.84	30.6	0.83
102K	3.51	29.4	1.28
195K	1.62	26.3	1.74
312K	**1.09**	**24.9**	**2.42**
500K	0.98	N/A	N/A

2.2.3. Language model

The language model is based on bigrams. From the 519M word corpus we were able to extract 65M different (360M in total) word-pairs belonging to the items in the 312K lexicon. Due to the inclusion of multi-words, 56M bigrams covered sequences of 3 words, 4M bigrams did it for 4 words. In [6] it was documented that these pseudo 3-grams and 4-grams were able to improve WER by some 5-10 % relatively.

2.3. Signal Processing and Acoustic Model

The system accepts 16kHz/16 bit sampled data coming from a common TV/radio card. The signal is framed and coded into classic 39 MFCC feature vectors. Cepstral Mean Subtraction (CMS) is performed after the stream is split into acoustically homogenous segments (section 2.4). The acoustic model consists of 3-state CDHMMs representing 41 Czech phonemes and 7 types of noise. The total number of 13824 gaussians were trained on our 48-hour database of mostly broadcast speech.

2.3.1. Speaker Independent & Gender Dependent AM

The first version of the transcription system employed speaker independent HMMs and produced results that are summarized in Table 1. Later we replaced them by gender dependent (GD) models, which yielded a 2 % WER reduction. In the recent version, even the GD models were put aside and now they serve as initial estimates for speaker adaptation only.

2.3.2. Off-line and On-line Speaker Adaptation

The system utilizes two types of speaker specific models. For the frequently occurring persons (anchors, reporters, top politicians), speaker adapted models are prepared off-line by the traditional MAP approach. For those, who are not in the speaker database or who are rejected in the speaker verification stage, an appropriate model is computed on-line. Here, we use a linear combination of the models belonging to the N-best subjects that were determined in the speaker identification phase. (Details on the combination method can be found in [7]).

The speaker identification and verification modules operate in two steps: In the first one, a fast match using 64-mixture GMMs finds a subset of the most probable candidates. In the second one, GMMs and a UBM (both with 1024 mixture) specify the best speaker and consequently verify his/her identity.

2.4. Stream Segmentation and Parallel Decoding

The stream of the acoustic data is continual and therefore it must be segmented into shorter parts first. These should meet the following criteria: a) they should be acoustically homogenous, namely with the same channel characteristics and possibly with only one speaker, b) they should be long enough in order to capture relevant context in speech, but at the same time c) the segment length should be limited with respect to the optimal operation of the speech decoders working in parallel.

2.4.1. Stream Segmentation

The aim of this module is to detect relevant changes in the audio signal and to use them as break-points for segmentation. For this purpose we developed our own segmentation scheme. It is based on the binary splitting method modified so that it can be applied on-line with minimum delay. In our case, the detector collects acoustic data and waits until a 10-sec long block is available. Then it applies the binary search for change-points. If more than one is detected, the leftmost one is taken. If none is found, the detector waits for the next block of data and then repeats the binary search again. When the block size exceeds a specified threshold (i.e. if there was no change within that block) the detector cuts the signal in the middle of the longest silence. The decision rule for locating and validating a single change point is based on the maximum likelihood approach [8].

2.4.2. Distributed Processing

The unit called load balancer cares about the proper distribution of speech segments to the dedicated recognition servers. Each of them has a complete recognition kit consisting of the speech and speaker recognition modules together with the acoustic, speaker and language models. When assigning the tasks to the servers, the load balancer takes into account their computation power and the parameters of the currently processed jobs. Another unit (usually running on the same machine) collects the results from the servers and puts them into a final document that contains the times, post-processed transcriptions and additional information about speakers, channels, programs, etc.

2.4.3. Decoder

Speech recognition is performed by our own decoder (based on the time-synchronous Viterbi algorithm), which has been optimized for very large vocabularies reaching up to 500K words. The search is done in a single pass. The result has form of 1-best word-sequence with start and end times assigned to each word. Moreover, N-best (usually 10) word-end hypotheses found for each frame can be stored on demand. The standard 1-best transcription as well as the partial hypotheses can be used for a key-word search. In such application the time stamps assigned to the words allow for easy acoustic check by replaying the corresponding signal part.

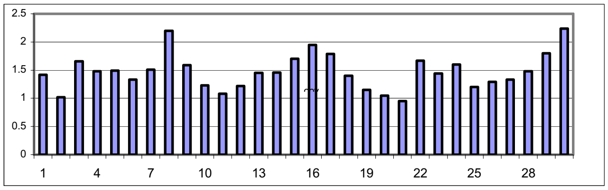
3. Continual Performance Refinement

As the system is intended for continual stream processing of broadcast data, we also had to develop tools that would support regular updating of system resources, namely the acoustic model, lexicon and language model. In the following section we describe these tools and evaluate their effect.

3.1.1. Lexicon and Language Model Updating

It is typical for broadcast news that its vocabulary changes in time more rapidly compared to other spoken programs. New events bring new words, in particular proper names. In inflected languages this phenomenon is even more critical because each new word means that its derivatives will appear sooner or later. In Graph 1, we can see OOV values measured for TV news (3 Czech stations) within one month (November 2005).

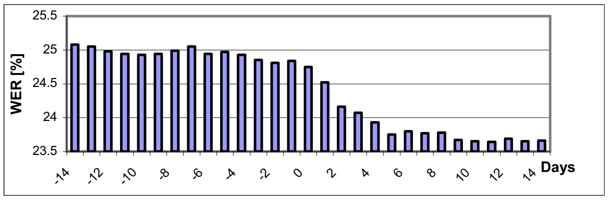
Graph 1*: OOV rates for TV news in Nov. 2005 (Lex312K)*



The tool for regular updating of lexical resources is based on continual collection of texts occurring in newspapers, news releases from press agencies and the latest broadcast transcriptions. Every day at 5 pm (1 hour before the start of the first main news show) the tool makes statistics of the recently collected data and lists the most frequent OOV words. A human operator selects those that are relevant and adds them (and their pronunciations) to the lexicon. The currently acquired text data are added also to the latest version of the text corpus (or better say to its extract in form of listed word-pairs). The complete update of the language model takes less than one hour.

We studied the effect of the regular LM update in experiments performed on TV news recorded in Nov. 2005. Three shows broadcast between Nov. 7 and 12 were transcribed with a series of language models that were updated every day in the period starting 14 days before the first show and ending 14 days after the last show. Before averaging the results, the WER values for each show were shifted in time so that the date of the show had index 0. The results are shown in Graph 2.

Graph 2*: WER values for TV shows achieved with LM updated every day in period -14 to +14 days relative to the date of broadcast (Lex312K, SI models)*



The above diagram shows that the data collected before the date of broadcast had very small positive impact on the performance. On the other side, the impact of adding later data was more evident since the WER value dropped by more 1 %. For on-line processing, the latter fact has no benefit, obviously. However, if we take into account the other application, i.e. the full-text search in automatically transcribed broadcast data, this later re-processing may be worth considering.

3.1.2. Updating of Acoustic Model

If at least some of the broadcast and transcribed programs are checked and corrected by a human operator, these verified data may serve as a source for updating the AM. In our system this is done by applying the procedure of forced alignment that finds the optimal phonetic transcription (including also potential noise events) and prepares the data for retraining the AM.

As we explained in section 2.3.2, the recent system employs speaker specific models only (no SI nor GD ones). Thus the updating of the AM means re-adaptation of the previously made SA models, and also adding new speakers and their SA models to the database. We used the previously mentioned November data in a series of experiments to analyze the effect of increasing the speaker set. The effect is two-fold: If there are more SA models, they can be used both for their 'owners' as well as for those who are not in the database but who rely on the model combination scheme mentioned earlier in section 2.3.2. Table 3 clearly shows that the increasing number of database speakers and their SA models has a considerable effect.

***Table 2**: Effect of acoustic model (Nov. 2005 BN data)*

	Models		SA models - number of speakers			
	SI	GD	13	52	152	307
WER [%]	25.1	23.9	23.6	22.4	21.4	21.3

3.1.3. Overall Performance Evaluation

In Table 3 we give a more detailed view on the results achieved for various types of broadcast programs recorded in 2005. Note that the data called BN-COST2005 are the Czech part of the European Broadcast News Database collected within COST278 project [5]. All the displayed values were obtained with the most recent system, the 312K lexicon, language model updated in November 2005 and SA model of 307 speakers. The figures in the last row can be compared with those in Table 1.

***Table 3**: Results for various types of programs*

Program Type	Time [min]	Words	WER [%]	OOV [%]
BN-COST2005	122	20131	19.6	1.2
BN–Nov.2005	62	9760	21.5	1.1
Weather news	10	1745	9.4	0.4
Talk shows	118	13624	33.81	0.6
Parliament	52	6395	20.8	1.1
Weighted Total	364	51655	21.98	1.07

4. Discussion

When developing the system we had in mind two types of its practical employment. First, we considered its use in a semi-automatic (human-corrected) transcription of selected broadcast programs. The other goal was unsupervised continual screening of one or more broadcast stations aimed at at least approximate monitoring of their spoken content.

Precise transcriptions are usually requested for such broadcast programs, like news, political talks and debates, or parliament speeches. The WER achieved in these jobs varies in larger range as it is shown in Table 3. It is lower for the BN task (at a 20 % WER level) and higher for the other types, like debates, where spontaneous speech dominates.

Here, it should be noted that WER values achieved for inflected languages, like Czech, will always be worse compared to e.g. English. The main problem is the vocabulary, namely its size and confusability. The lexical inventory of the inflected languages contains many word-forms derived from the same lemma that are acousticly very similar or even indistinguishable.

The remaining broadcast programs (those automaticly transcribed but not checked) can serve as source for efficient data mining. When analyzing the most frequent errors in speech recognition of Czech we saw that their majority consisted in a) omitting/inserting short words (1- or 2-letter conjunctions and prepositions) and b) substituting acousticly similar morphologic derivations of the same word [6]. In many cases, these errors are not critical and usually allow a reader to understand the content. Hence, it is possible to use such transcription for full-text search, even if its WER is around 30 %. In a typical case when the searched items are named entities (at least 4-letter long) and when we search a word stem rather then a specific inflected form, the success rate of full-text search in unsupervised transcription gets above 90 % as we show in [8].

Finally, we should mention also the time aspects of the system. If it runs on a single machine, the transcription will take about 2.3 real time on a common 3.4GHz processor. However, if we utilize all the implemented options of the distributed platform, three and more computers will allow for on-line processing of continual data stream. The typical delay between a spoken utterance and its transcription is 20 to 40 seconds. It means that such programs, like broadcast news, are ready for any other processing immediately after their final jingle.

5. Acknowledgements

This research was partly supported by the Grant Agency of the Czech Academy of Sciences (grant no.1QS108040569).

6. References

[1] Nguyen L., Xiang B., Afify M., Abdou S., Matsoukas S., Schwatz R., Makhoul J.: The BBN RT04 English Broadcast News Transcription System. Proc. of Interspeech'05, Lisbon, Sept. 2005.

[2] Gauvain J.L., Adda G., Adda-Decker M., Allauzen A., Gendner V., Lamel L., Schwenk H.: Where Are We in Trancribing French Broadcast News? Proc. of Interspeech'05, Lisbon, Sept. 2005.

[3] McTait K., Adda-Decker M.: The 300K LIMSI German Broadcast News Transcription System. Proc. of Eurospeech'03, Geneva, Sept. 2003.

[4] 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. Proc. of Interspeech 2005, Sept. 2005, Lisbon

[5] Vandecatseye A. et al.: The COST278 pan-European Broadcast News Database. Proc. of LREC 2004, Lisbon, Portugal, May 2004.

[6] Kolorenc J., Nouza J., Cerva P.: Multiwords in the Czech TV/Radio Transcription System. Proc. of Specom2006, St. Peterburg, June 2006

[7] Cerva P., Nouza J., Silovsky J.: Two-Step Unsupervised Speaker Adaptation Based on Speaker Recognition and Model Combination. Proc. of Interspeech2006

[8] Zdansky J.: BINSEG: An Efficient Speaker Based Segmentation Technique. Proc. of Interspeech2006

[9] Nouza J., Zdansky J., Cerva P., Kolorenc, J.: A System for Information Retrieval from Large Records of Czech Spoken Data. Proc. of TSD2006