Czech-to-Slovak Adapted Broadcast News Transcription System

Jan Nouza, Jan Silovsky, Jindrich Zdansky, Petr Cerva, Martin Kroul, Josef Chaloupka

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

{jan.nouza, jan.silovsky, jindrich.zdansky, petr.cerva, martin.kroul}@tul.cz

Abstract

The first broadcast news (BN) transcription system for Slovak is introduced. It employs the same modules as the system we developed earlier for Czech. We utilize similarity between the two languages in efficient lexicon building, in mapping Slovak specific (rarely occurring) phonemes onto Czech ones and in low-resource cross-lingual adaptation of acoustic model. The system uses 166K-word lexicon and on the Slovak part of European COST278 BN database achieves 23.6% WER (which is only 5% less than the original, long-term optimized Czech system). Similar results were achieved also on recently recorded data from four Slovak stations.

Index Terms: speech recognition, cross-lingual adaptation, broadcast news transcription, Slavic languages

1. Introduction

Recently, automatic speech recognition (ASR) technology has become a well established platform for efficient processing of spoken data. One of the major application fields is broadcasting where ASR serves for off-line as well as on-line transcription of broadcast news (BN), parliament sessions or various types of talk shows. Such transcription systems have been developed and employed for major languages like English, German, French, Japanese, Chinese or Spanish. Many of these systems were built specifically for the given target language but there are also examples of platforms that have been designed as multi-lingual, which enables developers to share modules, resources and experience [1].

Since 2003 we have been working on the development of the first BN transcription system applicable for Czech language. The task has been challenging because Czech belongs to the family of Slavic languages, which exhibit a large degree of inflection. In 2005 we completed and presented [2] an off-line version of the Czech BN transcription system operating with 300K-word vocabulary. Since 2006 the system has been used for automatic 24-hour monitoring of major Czech TV stations and indexing the content of selected spoken programs [3]. In 2008 we optimized the decoder and increased its speed so that now it is capable of real-time transcription (subtitling) of BN programs (with a 315K-word lexicon).

It was a natural idea to try to employ the same (or similar) approach and the already developed modules for another Slavic language. To prove the concept, we decided to run a pilot project focused on Slovak. There were several reasons for that decision:

First, Slovak is quite similar to Czech and most Czech people can understand it because both the languages were officially used in former Czechoslovakia. The ability to understand Slovak (though with some limitations) was a great benefit for us because we could do most linguistic and phonetic work by ourselves without hiring external experts.

The second reason was pragmatic. Because of intensive economical and cultural contacts between both parts of former Czechoslovakia, there are many Slovak people working and living in Czechia. Most of them still speak their native language and that is why Slovak quite often occurs in Czech broadcast programs. If this happens, our transcription system is confused and makes additional errors. If we are able to detect a Slovak speaker and have a Slovak ASR module, we can switch to it.

The third reason is that the media monitoring company that uses our software plans to extend its activities also to Slovakia. They asked us for a similar transcription and indexing software that could be applied in the same way for Slovak language.

2. How similar are Czech and Slovak?

Czech and Slovak belong to the West-Slavic branch of European languages. The former is used by some 10 million people living mainly in Czechia and the latter by 5 million Slovak speakers. The 2 languages are considered very similar and closely related because in the past both of them were official languages used within one state. Anyway, since 1993, when Czechoslovakia split, in the succession states a new generation of young people has grown who have problems to understand the language of the other nation. This makes evident that the difference between the two languages is larger than it was commonly thought. It may be of similar scale as in case of West-Iberian languages Spanish and Portuguese.

2.1. Difference and similarity in lexicons

In order to quantify the degree of dissimilarity we performed two types of linguistic measurements. Since both the Czech and Slovak Republics are members of the European Union we could find a lot of EU documents published in both national languages and employ them as parallel corpora. From EU Web pages [4] we took several relevant texts (such as the Treaty of EU, Lisbon Treaty and Amsterdam Treaty) and compiled lists of words used in their Czech and Slovak versions. The Czech texts contained 215K words, from which 12,434 were distinct, while the Slovak ones contained 218K words (12,217 distinct). The two lists of distinct words included 2,925 common items (i.e. 23 %). Later, in another test we compared the Czech and Slovak vocabularies compiled for our ASR systems. Both had (approximately) the same size of 170K words selected via the word frequency criterion. After removing the proper names, the size of the two lexicons decreased to 145K items, from which 29K (20 per cent) were common for both the languages.

Though the tests showed that about 80 % of the lexical inventories were different, the actual measure of dissimilarity is not so high if we perform a more detailed orthographic comparison of corresponding word pairs. Many differ only in
one or two characters, or in suffixes due to slightly different morphological patterns. To demonstrate this, let us compare Czech and Slovak versions of a sentence saying ‘Children have done exactly what teachers told them’.

CZ: Děti udělali práci, co jim řekli učitelé.
SK: Dieťa udělalo prácu, čo im řekli učiteľia.

2.2. Difference and similarity in phonetics

In spoken Czech, usually 41 individual phonemes are distinguished [5]. Some phoneticians recognize also three additional diphthongs. (Yet, in ASR system these can be modeled by concatenated corresponding phone models without loss of recognition accuracy.) The list of Czech phonetic inventory with SAMPA symbols is in Table 1.

<table>
<thead>
<tr>
<th>Groups (11)</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>a, c, e, i, o, u, î, ě, î, o, u, @ (schwa)</td>
</tr>
<tr>
<td>Consonants</td>
<td>p, b, t, d, c, J, k, g, ts, dz, tS,dZ, r, l, f, v, s, z, S, Z, X, j, h, Q, Q̄, m, n, N, J, F</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>o, u, a, u, e, u</td>
</tr>
</tbody>
</table>

For Slovak, literature offers several lists of phonemes. The authors of [6] distinguish between 57 different phonetic units. In [7], 48 individual phones and 4 diphthongs are recognized. We preferred the latter approach and used it to compile Table 2. It is organized in the way similar to Table 1, so that the distinctive items (printed in bold) can be seen.

<table>
<thead>
<tr>
<th>Groups (11)</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>a, c, e, i, o, u, î, ě, î, o, u, @ (schwa)</td>
</tr>
<tr>
<td>Consonants</td>
<td>p, b, t, d, c, J, k, g, ts, dz, tS,dZ, r, l, f, v, s, z, S, Z, X, j, h, Q, Q̄, m, n, N, J, F</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>î, o, î, e, î, U, u, o</td>
</tr>
</tbody>
</table>

A more detailed analysis of [7] shows that symbols ‘r=’ and ‘l=’ stand for syllabic versions of ‘l’ and ‘r’. These two syllable making phonemes occur also in Czech where they are considered as allophones of ‘l’ and ‘r’. Symbols ‘r=’ and ‘l=’ are just longer forms of them. Similarly, Slovak phonemes denoted as ‘G’ and ‘I^’ are considered as allophones of ‘h’ and ‘j’ in Czech. The remaining Slovak-specific phones and diphthongs can be approximated (at least for ASR purposes) by their acoustically closest Czech counterparts as we show in section 4.

3. Text and speech resources for Slovak

3.1. Text corpus

For creating a representative lexicon of contemporary Slovak and for computing a corresponding language model we were provided by a corpus of newspaper articles and broadcast news transcriptions from the 2005–2007 period. Its size was 1.9 GB, from which BN texts occupied about 10%.

The first analysis of the corpus discovered a serious complication when we found that not a negligible portion of the corpus had been Czech texts. This is a side effect of very close cultural contacts between the two formerly joined countries. Many articles from Czech newspapers are reprinted (without translation) in the Slovak ones. Also, when an interview with a Czech person is made, his or her answers often occur in the original language.

To cope with this phenomenon we decided to compile the lexicon by utilizing only the BN part of the corpus. It is formatted in the way that allows for identification of speakers occurring in BN shows and for at least rough estimation of their nationality. The first draft of the lexicon contained 180K items (words and their frequently occurring inflected forms). It was the words that were seen at least 3 times in the BN sub-corpus. In the next step, all of them passed through a Slovak spell-checker (aspell [10]) that helped us to discover about 14K non-Slovak (mostly Czech) words. After removing them, we got the final version of the lexicon with 166K words.

This lexicon was used for computing the corresponding bigram language model. In this case we utilized the whole available 1.9 GB text corpus.

3.2. Audio data

During the research and development process we have employed three sets of Slovak speech data.

In order to perform acoustic model adaptation (see section 4.2) we recorded 8 hours of TV and radio news shows. Recording was done either via satellite broadcasting (STV and TA3 stations) or via internet transmission (JOJ TV station and Slovak Radio). From this raw audio data, we extracted 6 hours of mostly clean speech to be included in the acoustic model retraining procedure.

The development works were done by exploiting the European database of broadcast news, which had been collected within the COST278 action in 2003 [8]. This database contains, among others, 3 hours of Slovak and the same amount of Czech BN recordings. The former was used for development and the latter for comparison tests.

The final evaluation was performed on a set of recent Slovak BN shows recorded in March and April 2008. The set covers broadcasting of three major Slovak TV stations (TA3, STV1, JOJ) and 1 nation-wide radio station (Slovensko). The evaluation set contains 8 complete BN shows (two from each station). This set (audio and reference annotations) has been made publicly available on our Web.

4. Adaptation of BN system to Slovak

When developing a BN transcription system for Slovak we wanted to utilize the existing platform that proved well for Czech language. Its block scheme is depicted in Fig. 1.
The system processes input data at several levels. At the first one, it segments audio stream into speech and non-speech parts. The former are split according to speaker turns into utterances. These pass through a speaker identification and verification stage that is used to select the appropriate gender-specific or speaker-specific acoustic model (AM). The AM is based on 39-MFCC representation and 3-state HMMs of 41 phonemes and 7 types of noises. The decoder is capable of operating with lexicons whose size goes up to 500K words. Since 2006, when the system was introduced [3], it has been further optimized so that it allows for real-time operation. Having this functional platform, our strategy was to utilize it as it was and modify only the language specific data, i.e. the lexicon, the AM and the LM.

4.1. Lexicon, pronunciation and phonetic map

The lexicon extracted from the text corpus contains 166,535 most frequent words and word-forms. To get pronunciations for them, we modified our grapheme-to-phoneme converter by including Slovak specific phonetic rules described in [7]. Since we knew that the available amount of Slovak acoustic data had not been large enough to train a separate Slovak AM, we had to utilize (and later adapt) the existing Czech AM. Therefore, it was necessary to make another conversion, from Slovak phonetic inventory to the Czech one. This was done by mapping the Slovak-specific phones and diphthongs into their Czech counterparts, either single phonemes or phoneme strings. The mapping rules are summarized in Table 3. It should be noted that for only 3 Slovak phones (‘L’, ‘l=’ and ‘r=’) there is a subtle audible difference between the original ones and their Czech replacements.

Table 3: Slovak-specific phonemes (with corresponding orthography) mapped onto acoustically closest Czech ones

<table>
<thead>
<tr>
<th>SK letter(s)</th>
<th>SK phoneme</th>
<th>CZ phoneme(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ľ</td>
<td>ľ</td>
<td>l</td>
</tr>
<tr>
<td>l</td>
<td>l=</td>
<td>1</td>
</tr>
<tr>
<td>ř</td>
<td>ř=</td>
<td>r</td>
</tr>
<tr>
<td>v</td>
<td>U ^</td>
<td>u</td>
</tr>
<tr>
<td>v</td>
<td>w</td>
<td>v</td>
</tr>
<tr>
<td>h</td>
<td>G</td>
<td>h</td>
</tr>
<tr>
<td>j</td>
<td>J</td>
<td>j</td>
</tr>
<tr>
<td>ô</td>
<td>ò, o</td>
<td>uo</td>
</tr>
<tr>
<td>ia</td>
<td>ï, ã</td>
<td>ja</td>
</tr>
<tr>
<td>ie</td>
<td>ï, e</td>
<td>je</td>
</tr>
<tr>
<td>ia</td>
<td>ï, ï\U Sl</td>
<td>ja</td>
</tr>
</tbody>
</table>

Thanks to this conversion and language similarity, the usual problem with phonetic transcription of foreign proper names could be solved simply by copying the pronunciations from the existing Czech lexicon.

Initial experiments showed that the above described mapping worked quite well. However, we have noticed that a significant number of word substitutions happened because of phonetic assimilation, namely at word boundaries in fluent speech. For Slovak, this phenomenon (change of voiced sound into unvoiced and vice versa due to phonemes that follow) is very typical and frequent. We solved it by adding alternative pronunciation to all the words that end with pair consonants. Alternative phonetic forms had to be added also to homonyms and to those words whose pronunciation depends on ambiguous morphological classification (e.g. words like “krásne” or „citová”, where letter ‘n’ can be converted into phoneme ‘n’ or ‘J’). After all these refinements the lexicon included 196,994 phonetic forms (1.18 per word).

4.2. Acoustic model

When all phonetic transcriptions have been mapped onto the Czech phoneme inventory, we could try to utilize the existing Czech AM in initial ASR experiments. The results were surprisingly good (see Table 4). Anyway, the next natural step was AM adaptation. The training part of COST278 data included 2 hours of annotated Slovak BN speech, which was not so much, and therefore we recorded another 4 hours of training data (BN shows from 3 Slovak TV stations). In this way we got 3 hours of male and 3 hours of female speech that could be used for re-training of Czech gender-specific AMs. This was done by adding the Slovak training data to the Czech one (34 hours of male and 27 hours of female speech) and repeatedly running the Baum-Welch reestimation procedure. Hence, the resulting gender-specific models were made of approx. 90 % Czech and 10 % Slovak data.

4.3. Language model

The LM was trained on the test data described in section 3.1. In the 1.9 GB corpus we found 234M occurrences of the 166K-word lexicon items. A bigram LM was computed from 32M seen word-pairs by applying Witten-Bell smoothing.

5. Experimental evaluation

Two types of experiments were conducted. In the first one we focused on searching for optimal configuration with regard to lexicon and acoustic modeling. We also tried to compare the results achieved for Slovak with those received previously for Czech. In the second series of experiments we evaluated the developed system on a larger set of Slovak data.

5.1. Development experiments

In these experiments we used the test sub-set of the COST278 database. Its Slovak part is about 50 minutes long. After checking all recordings, correcting mistakes in reference annotations and removing several utterances spoken in Czech, the test set consisted of 154 utterances with 3921 words.

In the first series of experiments we measured the impact of lexicon size and pronunciation modeling. These initial tests were done with the original Czech AM. Table 4 clearly shows that a lexicon with less than 100K words is inappropriate for a language like Slovak. Even increasing its size to 166K words cannot assure an OOV rate below 2 %. Anyway, it was possible to reach almost 25 % WER if multiple phonetic forms were included to those words mentioned in section 4.1.

Table 4: Impact of lexicon size and pronunciation

<table>
<thead>
<tr>
<th>Lexicon type</th>
<th>OOV [%]</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>64K words, single pronuc.</td>
<td>5.63</td>
<td>32.61</td>
</tr>
<tr>
<td>166K words, single pronuc.</td>
<td>2.12</td>
<td>27.31</td>
</tr>
<tr>
<td>166K words, multiple pronuc.</td>
<td>2.12</td>
<td>25.08</td>
</tr>
</tbody>
</table>

In the next series of experiments we evaluated the effect of acoustic model adaptation. Table 5 compares WER values obtained for the original Czech (gender-dependent) models with those we got after having re-trained the Czech models using 2 hours of COST278 training data and 4 hours of recently acquired BN data. We can see that the first addition contributed to 1 % improved WER, while the next (double) amount of Slovak data helped to reduce WER by only 0.5 %.
Table 5: WER values achieved with original Czech acoustic model and models adapted to Slovak language

<table>
<thead>
<tr>
<th>Acoustic models (all gender dependent)</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech AM</td>
<td>25.08</td>
</tr>
<tr>
<td>Adapted Czech AM (2 hours of Slovak)</td>
<td>24.12</td>
</tr>
<tr>
<td>Adapted Czech AM (6 hours of Slovak)</td>
<td>23.64</td>
</tr>
</tbody>
</table>

Since the system for Slovak BN transcription has been developed by a straightforward modification of the existing Czech system, we wanted to compare the performance of both. This time we used the Czech part of COST278 database, which contains 70 minutes of BN speech (7451 words in 309 utterances). The test data was submitted to two configurations of the Czech system. One was its latest version that operates with a 315K-word lexicon, performs speaker identification, employs GD as well as SA acoustic models and its language model was trained on 4 GB of texts. The second configuration was a down-scaled version of the first one, with parameters similar to those of the Slovak system, i.e. a lexicon limited to 170K words and GD models only. Test results are shown in Table 5. If we compare them with the performance of the Slovak system (the last row in Table 5), we can see that the difference in WER is about 5% in case of the advanced system, but only 1.3% for the comparable system.

Table 6: Results of comparative tests done with Czech data and two configurations of recent Czech BN system

<table>
<thead>
<tr>
<th>Czech system configuration</th>
<th>OOV [%]</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon 315K, SA+GD models</td>
<td>0.87</td>
<td>18.26</td>
</tr>
<tr>
<td>Lexicon 170K, only GD models</td>
<td>1.86</td>
<td>22.34</td>
</tr>
</tbody>
</table>

5.2. Evaluation on 2008 data

In order to get a more representative picture about the performance of the Slovak transcription system, we prepared a new collection of BN shows from all major TV stations and one nation-wide radio. During March and April 2008 two complete shows from each station were recorded and annotated. All parts that contain speech were included in the test set (even headlines with music played in background and shots taken in very noisy conditions). The test data had total duration of 128 minutes and contained 19,021 words in total. Results from the tests are summarized in Table 7.

Table 7: Evaluation tests on recently recorded BN shows

<table>
<thead>
<tr>
<th>Station</th>
<th>#words</th>
<th>OOV [%]</th>
<th>WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV - TA3</td>
<td>5,568</td>
<td>2.14</td>
<td>22.85</td>
</tr>
<tr>
<td>TV - STV1</td>
<td>5,318</td>
<td>2.37</td>
<td>25.58</td>
</tr>
<tr>
<td>TV - JOJ</td>
<td>5,117</td>
<td>3.89</td>
<td>27.81</td>
</tr>
<tr>
<td>Radio - Slovensko</td>
<td>3,018</td>
<td>2.02</td>
<td>16.76</td>
</tr>
<tr>
<td>Overall results</td>
<td>19,021</td>
<td>2.65</td>
<td>23.95</td>
</tr>
</tbody>
</table>

We can notice that there is a significant difference in results received for individual stations. This reflects the style of reporting that each broadcasting subjects prefer. Radio Slovensko and TV TA3 are news oriented stations that want to be considered as serious sources of information. The other two TV stations seem to put more focus on attractive presentation of news by often (mis)using music and noisy background and including fast sequences of very short shots.

All the recently acquired test data can serve for other researchers interested in Slovak language processing as they are publicly available at [9].

6. Discussion

The results from both types of experiments show that the BN transcription system adapted to Slovak lexicon, phonetics and language performs almost as well as the similarly configured original system developed for Czech. To reach significantly better performance, we will have to follow the same steps we employed in refining the Czech system, namely 1) increasing the lexicon at least to 300K words, 2) including frequent multi-word terms in the lexicon [11], 3) training speaker-specific acoustic models for frequently occurring speakers [12], and 4) increasing the size of text corpus to get a richer source for language modeling. Most of these steps can be accomplished when the recent version is applied in practical test service (planned for second half of 2008). The system will be used for automatic monitoring of Slovak broadcast media and at the same time it will collect data that are necessary for all the above mentioned refinement steps.

7. Conclusions

In this paper we describe a cost and time efficient adaptation of an existing BN system to another tongue from the same family of Slavic languages. All of them exhibit a high degree of inflection and hence they require very large vocabularies, tools for morphological analysis and a decoder that can cope with these specific features. Our modular system originally developed for Czech meets these requirements and can serve as a starting platform for other Slavic languages. In this pilot study we focused on Slovak as the language closest to Czech and proved applicability of the concept. In near future, we would like to test the same or similar approach for other (major and minor) Slavic languages, like Polish and Russian.

8. Acknowledgements

This work has been supported by GA CR (grant no. 102/08/0707) and GA AVCR (grant no. I QS108040569).

9. References