Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs

Jan Nouza, Dana Nejedlova, Jindrich Zdansky, Jan Kolorenc

SpeechLab, Department of Electronics and Signal Processing
Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic
{jan.nouza, dana.nejedlova, jindrich.zdansky, jan.kolorenc}@vslib.cz

Abstract

This paper describes the first speech recognition system capable of transcribing a wide range of spoken broadcast programs in Czech language with the OOV rate being below 3 per cent. To achieve that level we had to a) create an optimized 200k word vocabulary with multiple text and pronunciation forms, b) extract an appropriate language model from a 300M word text corpus and c) develop an own decoder specially designed for the lexicon of that size. The system was tested on various types of broadcast programs with the following results: the Czech part of the European COST278 database of TV news (71.5 % accuracy rate on complete news streams, 82.7 % on their clean parts), radio news (80.2 %), read commentaries (78.6 %), broadcast debates (74.3 %) and recordings of the state presidents’ speeches (85.8 %).

1. Introduction

Automatic transcription of spoken broadcast programs is one of the most challenging tasks in the speech processing domain. Recently, several systems with that purpose have been developed for major languages. The results reported in papers [1-3] demonstrate that the machine transcription of broadcast news in English, Spanish or Japanese performs already quite well and relatively fast with vocabularies whose size may be limited to 60 thousand most frequent words. However, there are languages that need much larger vocabularies to cover the width of their lexical inventories. One of them is, for example, German, in which the lexicon must comprise up to 300k words to get the same text coverage as in English [4].

The transcription of spoken Czech is even a more complex task. Unlike German, where the lexicon size grows mainly due to the compound formation, the major difficulty of Czech (as well as Russian and other Slavic languages) consists in the existence of many inflected forms that can be derived from a single lexical lemma (up to 14 for a noun, 100 for an adjective and even more for a verb). These forms reflect the complex grammar rules that are governed by the principle of gender, number and case agreement between the interrelated parts of speech in a sentence [5]. The side effect of these inflected forms and their strong grammatical relation is a relatively free word order in a Czech utterance, i.e. something that goes against the assumptions of the standard N-gram language modeling method.

The first attempts to adopt the existing speech recognition platforms (the AT&T decoder and the SRILM toolkit) to the Czech broadcast news transcription task dates to 2001 [6]. It was clear, however, that the vocabulary size limits posed by those tools were critical for the task and they could not be efficiently eliminated by alternative approaches, like e.g. a morphological decomposition of words. In this paper we demonstrate that if we want to achieve significantly better results we must work with lexicons that contain several hundreds of thousands words and word forms. We further show how to optimize the lexicon building process, how to benefit from multiple text and pronunciation variants of lexical items, and in the end we briefly describe the design of the own decoder fitting well to the very large vocabulary task.

In the experimental part we present the results that were achieved in a wide range of applications. We do not focus just on the classical broadcast news task. Instead, we test the system also on other interesting jobs, like the transcription of radio debates, daily commentaries or broadcast speeches given by politicians.

2. Text corpus

As the research moves from domain oriented tasks to unconstrained speech recognition with virtually unlimited vocabulary, the importance of large text corpora increases dramatically. They are needed for compiling lexicons that will optimally represent the given language, for estimating parameters in probabilistic N-gram models, or for monitoring short and long-term socio-linguistic evolution.

For Czech language (which is spoken by some 10 million people) the National Czech Corpus has been compiled at the Charles University in Prague [7]. However, its primary orientation on linguistic research and its limited size makes it inappropriate for our purpose. Therefore, three years ago we started to build our own corpus of Czech texts.

2.1. 300M word corpus and its characteristics

The corpus is compiled from the texts that have been available in electronic form. Recently its size is about 1.9 GB and after some cleaning it contains 290 million words. The majority of the texts (275M words) come from Czech newspapers published in period 1990 to 2004, which were obtainable either on commercial CDs or on internet. The remaining part is constituted from other electronic sources, mainly novels and professional books available on web. Unfortunately, only a 12 MB portion represents the transcription of spoken language – it is mainly TV and radio news.

The analysis of the corpus discovered 1,910,641 distinct items in the texts. From these, 788,251 passed successfully the spell-checker built in the Czech version of the MS Word editor. All these passing words were general lexical items, not proper names. The proper names (i.e. those with the initial capital letter) made a group of other 400K words. The remaining part of almost 700M tokens was not analyzed in detail. We just assume that it consists mainly of typing errors, out-of-standard colloquial words and items coming from foreign languages. The analysis showed that we must expect
at least 1.3M words in standard written Czech. It should be also mentioned that 334,767 items from the group of the 788,251 general words occurred less than 4 times in the corpus. These “histogram-tail” words can hardly be denoted as unusual or extra rare. In many cases they were just the less frequent inflected forms of the very common words.

2.2. Corpus preprocessing

If the corpus is to be used for spoken language processing, it must be first preprocessed to fit better to the target application. In our case we had to perform the following steps:

- **Replacement of digits by their spelling.** This task is very difficult in inflected languages because the names of the digits can take multiple inflected forms according to the context. The most challenging subgroup is the ordinals. (In text they are followed by a period). They behave as adjectives and can take up to 30 different forms. For the translation into the correct spelled form we used the information about the morphology classes of the adjacent words – for details see part 3.2. Though, there were many ambiguous situations where the translation could not run automatically and had to be omitted.

- **Expansion of abbreviations.** While some abbreviations (like Czech equivalent of ‘etc’) could be expanded easily, other (like e.g. ‘EU’) had to be translated again in context dependent way. Since some abbreviations and acronyms may be uttered in the short as well as in the expanded way we used a random generator to get the right proportion of both variants.

- **Orthography normalization.** Since many words, mainly those of foreign origin, may have alternative spelling, we have made the unification of the orthography towards the most frequent variants. This also helped to make the lexicons slightly smaller and more compact.

- **Case normalization.** For the purpose of the N-gram modeling and lexicon building all words were converted to lower case. During transcription, the proper case for the recognized words is determined in a post-processing stage, which is based on a special bigram model.

3. Lexicon building

Based on the preprocessed corpus the word frequency list was compiled. Usually, the finite-size vocabulary is made in the way that the top L words are chosen. However, this approach does not take into account the context of the selected words. It may happen that some of them occur frequently but not as part of regular sentences. (This happens especially in newspaper text, e.g. in titles, short comments, editor’s or speaker’s names, etc).

3.1. Scaled lexicons

In our case we have compiled the parent (the largest) lexicon and its down-scaled versions with respect to the contextual frequency. Each word was ranked according to the average number of occurrences at both sides of the bigrams. After that we selected those words whose rank was larger than the prescribed threshold. In this way we have created 7 lexicons that are listed with their parameters in Table 1. The smaller lexicons are always subsets of the larger ones.

<table>
<thead>
<tr>
<th>Lexicon name</th>
<th>Size in words</th>
<th>Minimal word rank</th>
<th>OOV rate on text [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex31k</td>
<td>30957</td>
<td>600</td>
<td>14.52</td>
</tr>
<tr>
<td>Lex60k</td>
<td>66243</td>
<td>200</td>
<td>8.94</td>
</tr>
<tr>
<td>Lex102k</td>
<td>101536</td>
<td>100</td>
<td>6.71</td>
</tr>
<tr>
<td>Lex124k</td>
<td>124297</td>
<td>70</td>
<td>5.87</td>
</tr>
<tr>
<td>Lex148k</td>
<td>148011</td>
<td>50</td>
<td>5.28</td>
</tr>
<tr>
<td>Lex175k</td>
<td>175121</td>
<td>35</td>
<td>4.81</td>
</tr>
<tr>
<td>Lex200k</td>
<td>199138</td>
<td>30</td>
<td>4.55</td>
</tr>
</tbody>
</table>

3.2. Lexicon structure

For the text processing and speech decoding purposes we have proposed the lexicon structure that is shown in Table 2. Each item is represented by its standard orthography optionally followed by other alternative forms. This helps not only in the normalization of the training and testing corpora but also in the generation of pronunciation variants. These are listed in the third column. (For the transcription symbols see [8].) The last column contains the names of the morphology classes to which the word belongs [9].

3.3. Pronunciation modeling

Czech has rather simple and straightforward rules for generating phonetic transcription. About 100 major rules are mentioned in literature. Using the Grammatical Evolution approach [10] we have derived 154 other minor rules. Yet, we had to detect the words of the foreign origin and write their pronunciation manually. Further, it was necessary to add alternative phonetic forms to the items whose pronunciation varies with context in continuous speech. This happens mainly to the words ending with voice/unvoiced pair consonants. These were given alternative transcription if they were ranked among the 5000 most frequent words or if shorter than 6 phonemes. In this way, the largest Lex200k lexicon was equipped with 2358 additional phonetic forms.

3.4. Merging and splitting words

Like other authors we have also compiled a list of the most frequent collocations (word strings often appearing together). The list includes names (such as ‘New York’), 2-word terms (‘de facto’, ‘au pair’) or phrases (Czech equivalent to ‘it is’, etc). From the list we took those 314 collocations that were shorter than 3 syllables and added them to the lexicon.

Word compounding is not typical in Czech. Though, there is one exception: the adjectives having form like 15-year old or 27-kilometer long, etc. They are one-word terms, whose number grows very high - being the product of all possible numerals and all inflected forms of adjectives. We have created a scheme that decomposes these words, both in the
corpus and in the lexicon. A special post-processing routine merges them back after speech decoding is done.

4. Speech recognition

During the last five years we have been building our own speech recognition system that meets the specific needs of the Czech language. Its recent version is capable of operating with a very large vocabulary (above 100k words, with a temporal limit set to 300k). It may process continuous utterances whose length goes up to 2 minutes and can be directly coupled to the broadcast stream segmentation module described in [11].

4.1. Acoustic modeling

The signal processing module takes 8 kHz sampled data and transforms it into 39-dimensional MFCC vectors every 10 ms. The Cepstral Mean Subtraction routine is applied when the end of the segment is reached. After that the actual decoding procedure can start.

The acoustic signal is assumed to be made of 48 basic elements: 41 phonemes and 7 types of noise. They are listed in Table 3. Notice that in Czech we have to distinguish between short and long vowels because they discriminate many minimal word pairs. The schwa vowel is not typical for Czech but it had to be added to the list because it may occur when people spell letters or abbreviations. The set of the noises have been optimized on several speech databases. The introduction of the glottal stop showed to be very important and contributed to the significantly improved classification of short words starting with a vowel.

Each of the 48 units has its own HMM with the classic three-state left-right structure. Instead of often used triphones we use context-independent monophones with large number of mixtures. Their number was determined automatically during the HMM training process and most states are represented by 100 gaussians or so. The monophones are not only easier for manipulation and computation but they proved to be more robust against acoustic and pronunciation variations in speech recognition that operates under speaker and environment changing conditions. The HMMs have been trained on 45 hours of annotated speech database that was a mix of microphone and broadcast signal.

4.2. Language modeling and decoding

The language model is based on bigrams. About 35 million distinct word-pairs were found in the 300M word corpus. Such a low coverage rate calls for an efficient smoothing scheme. We adopt the well-known Witten-Bell smoothing technique with the modification described in [9]. It takes into account the fact that some of the word classes have zero bigrams. (For example, the prepositions cannot be followed by inflected word forms other than those corresponding to the preposition’s valence. The conditioned smoothing applies also to the split parts of the compound words mentioned in section 3.4 in order to prevent the merging of wrong word pairs. So far, the trigrams were tested only for smaller lexicons (up to 60k words). They yielded only a negligible improvement (less than 1 %) paid by significantly increased computation and memory load.

The decoder is based on the time synchronous one-pass Viterbi search. Its parameters, like the number of the word-end hypotheses, pruning and LM weighting factor or word insertion penalty, had been optimized earlier on the development data and remained constant for all the tests mentioned below. The decoder consumes about 200 MB operational memory for the 200k vocabulary, from which the major portion is occupied by the bigram values.

4.3. Post processing and evaluation

The standard output of the decoder is a string of recognized words printed in lower case. Optionally, a post-processing scheme is applied that capitalizes the initial letter of the utterance, makes the same with the potential proper names (using their list and corresponding bigrams) and adds the punctuation (periods and commas) at least to the points where their appearance is very probable.

The evaluation, however, was done in a case-insensitive and punctuation-less way. All tests were evaluated using the accuracy rate that takes into account substitutions, deletions and insertions. Recognition time was quantified by the real-time factor measured on a PC with a 2.6 GHz Intel processor.

5. Experiments and results

The goal of the experimental part was to learn the ability of the system to transcribe various types of broadcast programs spoken in Czech. It should be stressed that no effort was made to adapt the system and its parameters to any of the tested job. The acoustic models, the lexicons and the LM remained same in all the tasks. In contrast to many other authors, our evaluation policy does not consist in comparing ‘bad’ baseline results with the better ones after some refinement. Instead, we first present the best results of the fully optimized system and only if it is relevant we compare them to those achieved under some suboptimal conditions.

5.1. TV news transcription

Most experiments were performed on the Czech part of the multi-lingual TV news database that is available for public benchmarking [12]. Its test portion contains three complete TV shows from three different Czech stations (1 hour in total, 8451 words). Only the opening and closing jingles were removed, the rest was segmented according to the speaker changes. Therefore some utterances were quite long (max. 181 words). The speech read by professional speakers in studio made only 44 % of the data, the majority of the audio signal came from more or less severe conditions.

In Table 4 we present the total results achieved for the lexicons of different size. Compared are the values of the transcription accuracy, OOV rate and recognition time. We can see that the increase of the lexicon size consistently contributes to the reduction of the OOV rate and to the improvement in the accuracy. From that trend we can assume that larger lexicons would make the results even better.
In Table 5 we show the results as function of different focus conditions ([12]). Notice that the read speech (F0) is recognized quite well, while spontaneous speech (F1), speech with music background (F3) and speech on street (F4) degrade the overall results in significant way.

Table 5: Results under different conditions (Lex200k)

<table>
<thead>
<tr>
<th>Focus conditions</th>
<th>F0</th>
<th>F1</th>
<th>F3</th>
<th>F4</th>
<th>Fx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc. [%]</td>
<td>82.72</td>
<td>56.33</td>
<td>61.75</td>
<td>58.01</td>
<td>75.00</td>
</tr>
</tbody>
</table>

In Table 6 we try to compare the results of the optimized system with those tests where one of the components was removed or reduced. We can see that the introduction of the most frequent collocations and alternative pronunciations has surprisingly small effect. The consideration of the multiple lexical forms is more important because it eliminates possible mismatch between alternative spellings of the semantically same words. The size of the corpus for the LM training still play the relevant role as it can be seen from the experiment where only 80% randomly chosen sentences were used.

Table 6: Results for some suboptimal conditions (Lex200k)

<table>
<thead>
<tr>
<th>Suboptimal modifications</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized system</td>
<td>71.53</td>
</tr>
<tr>
<td>Lexicon without collocations</td>
<td>71.11</td>
</tr>
<tr>
<td>Lexicon without alternative pronunciation</td>
<td>70.85</td>
</tr>
<tr>
<td>Lexicon without multiple lexical forms</td>
<td>70.05</td>
</tr>
<tr>
<td>LM based on reduced corpus (80% of the orig.)</td>
<td>69.89</td>
</tr>
</tbody>
</table>

5.2. Other transcription tasks

Radio news: Here, the test data was made of 1 hour of radio evening news (national station CR1 - Radiozurnal). The results are better than those obtained for the TV news, which reflects the fact that public radio usually has better acoustic results are better than those obtained for the TV news, which even evening news (national station CR1 - Radiozurnal). The

Table 7: Results from other broadcast programs (Lex200k)

<table>
<thead>
<tr>
<th>Transcription task</th>
<th># words</th>
<th>PPX</th>
<th>Acc [%]</th>
<th>OOV [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV news</td>
<td>8451</td>
<td>517</td>
<td>71.53</td>
<td>1.38</td>
</tr>
<tr>
<td>Radio news</td>
<td>7857</td>
<td>522</td>
<td>80.19</td>
<td>1.76</td>
</tr>
<tr>
<td>Radio commentaries</td>
<td>19109</td>
<td>544</td>
<td>78.63</td>
<td>2.62</td>
</tr>
<tr>
<td>Radio debates</td>
<td>16510</td>
<td>686</td>
<td>74.30</td>
<td>0.99</td>
</tr>
<tr>
<td>Political speeches</td>
<td>2712</td>
<td>506</td>
<td>85.77</td>
<td>1.25</td>
</tr>
</tbody>
</table>

6. Conclusions

The system presented in the paper is the first very large vocabulary speech recognition system completely built in Czech Republic. It is primarily designed for the off-line transcription of Czech broadcast programs. With the largest 200k vocabulary it is capable of processing spoken Czech with 70 to 85% accuracy rate depending on the acoustic conditions and speaking style. The transcription is available in time which is approx. 6 times RT. The optimized 200k lexicon ensures OOV < 3% in a wide range of broadcast programs for such highly inflected language like Czech.

7. Acknowledgements

The work was supported by the Czech Grant Agency in project no. 102/02/0124.

8. References