Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak)

Jan Nouza, Jindrich Zdansky, Petr Cerva, and Jan Silovsky

Institute of Information Technology and Electronics, Faculty of Mechatronics,
Technical University of Liberec,
Studentska 2, CZ 46117 Liberec, Czech Republic
{jan.nouza,jindrich.zdansky,petr.cerva,jan.silovsky}@tul.cz

Abstract. Slavic languages pose a big challenge for researchers dealing with speech technology. They exhibit a large degree of inflection, namely declension of nouns, pronouns and adjectives, and conjugation of verbs. This has a large impact on the size of lexical inventories in these languages, and significantly complicates the design of text-to-speech and, in particular, speech-to-text systems. In the paper, we demonstrate some of the typical features of the Slavic languages and show how they can be handled in the development of practical speech processing systems. We present our solutions we applied in the design of voice dictation and broadcast speech transcription systems developed for Czech. Furthermore, we demonstrate how these systems can be converted to another similar Slavic language, in our case Slovak. All the presented systems operate in real time with very large vocabularies (350K words in Czech, 170K words in Slovak) and some of them have been already deployed in practice.

Keywords: Speech recognition, voice dictation, spoken document transcription, Slavic languages, inflective languages.

1 Introduction

During the last decade, speech technology has become a well established platform for human-computer interaction. It has been successfully deployed in voice dictation programs (IBM ViaVoice being one of the first in 1997, [1]), spoken document transcription systems [2], dialogue based information services [3], mobile device interfaces [4], or assistive tools for handicapped people [5]. The earliest systems were produced for major languages, like English, German, French or Japanese. Later some other tongues spoken in Europe and Asia have been covered, too.

Yet, there are groups of languages that still wait for more intensive deployment of modern voice technologies, and the family of Slavic languages is one of them.

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNCS 5967, pp. 225–241, 2010.

Research focused on developing functional text-to-speech (TTS) and speech-to-text (STT) services has existed in many Slavic countries, however, it is mainly the former that have been already applied in broader scale while the latter - automatic speech recognition (ASR) systems - have not reached such a mature level, so far. One of the reasons is the complex nature of Slavic languages. They are inflective, which means that words can get many different forms according to context. Moreover, they have very complex grammar, they allow for rather free word order in sentences, and last but not least, a large set of acoustically and phonetically similar prefixes and suffixes makes the traditional speech recognition algorithms less successful.

In this paper, we want to explain where the main problems and challenges are and how they can be solved. Most of the features that are typical for Slavic languages are demonstrated on Czech because it is the language we have been working on for more than 15 years. For Czech we can also present the solutions that proved to be successful and led to the development of practical systems, both commercial and non-commercial ones. In the last part of the paper we show how the experience gained during the long-term research can be re-used for another similar language, in our case Slovak.

2 Slavic Languages

Slavic (sometimes called also Slavonic) languages are spoken by almost 300 million people living in central, southern and eastern Europe, and in Asian part of Russia. This family of languages consists of three main branches, with the following main representatives:

- West Slavic - Polish, Czech and Slovak,
- South Slavic - Serbian, Croatian, Bulgarian, Slovene and Macedonian,
- East Slavic - Russian, Ukrainian and Belarusian.

The name of the branches highly correlates with geographic distribution of the countries where these languages are used as it is shown in Fig. 1. All the West Slavic languages as well as Slovene and Croatian use Latin alphabet while the other South Slavic and all East Slavic nations use Cyrillic characters.

3 Major Challenges in Machine Processing of Slavic Languages

3.1 Inflective Nature of Slavic Languages

All Slavic languages exhibit very large degree of inflection. This means that the vast majority of lexical items (except of adverbs, prepositions and conjunctions) modify its basic form (lemma) according to grammatical, morphological and contextual relations. Nouns, pronouns, adjectives and numerals change their orthographic and phonetic forms with respect to grammatical case, number and gender. For examples, see Tables 1 and 2. Verbs are subjects to conjugation controlled by grammatical categories, like person, number, gender, tense, aspect, etc. An example is in Table 3.

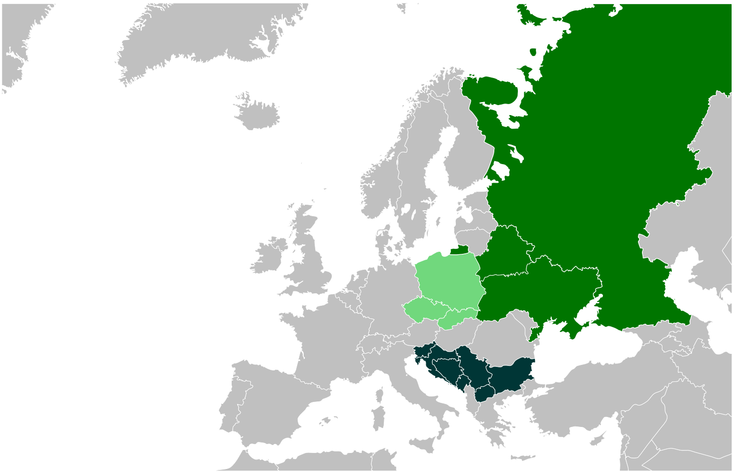


Fig. 1. Three groups of Slavic languages spoken in Europe (source Wikipedia)

Table 1. Comparison of an adjective *(nice)* in English and its inflected equivalents in Czech

English (1)	**Czech (12)**
nice	krásný, krásného, krásnému, krásném, krásným, krásná, krásné, krásnou, krásní, krásných, krásnými, krásnýma

3.2 Rich Morphology Results in Extremely Large Lexicons

Three examples presented in Tables 1, 2 and 3 show just the most straightforward type of inflection when a lemma is modified by endings attached to a word stem. However, morphology of Slavic languages is even more complex. New words with more or less similar meaning can be created by adding (single or multiple) prefixes, suffixes and endings to a stem, or also by modifying the stem itself. An example of some of these word production patterns, from simple to very complex ones, is shown in Table 4. One can notice, that in Czech (and in the other Slavic languages as well) even a negative form of a verb is a new lexical item (e.g. ”neučit”) unlike in English were the negative form is expressed by separate particle ”not”. In a similar way, the grading of adjectives and adverbs is done entirely by adding special suffixes and prefixes, not by separate words, like ”more” and ”most” in English. Last but not least, Slavic languages often use special words for male and female surnames (”Navratil” vs. ”Navratilova”) and names of professions (as shown in Table 2).

All these specific features result in very large number of word-forms. This number often exceeds one million distinct items, which have to be covered and managed by automatic speech processing systems. This is a big difference if we compare it to ASR systems designed e.g. for English where the inventory of 50

Table 2. Comparison of a noun *(student)* in English and its inflected equivalents in Czech (notice distinct forms used for *male student* and *female student* in Czech)

English (2)	**Czech (20)**
student, students	**Masculine:** student, studenta, studentu, studentovi, studente, studentem, studenti, studentů, studentům, studenty, studentech **Feminine:** studentka, studentky, studentce, studentku, studentkou, studentek, studentkám, studentkách, studentkami

Table 3. Comparison of a verb *(break)* in English and its inflected equivalents in Czech

English (5)	**Czech (19)**
break, breaks, broke, broken, breaking	zlomit, zlomím, zlomíš, zlomí, zlomíme, zlomíte, zlom, zlomme, zlomte, zlomil, zlomila, zlomilo, zlomili, zlomily, zlomen, zlomena, zlomeno, zlomeny, zlomeni

Table 4. Several examples of word production patterns applied to Czech word *učit (teach)*

Pattern type	**Czech word**	**English equivalent *(approximated)***
basic form (verb)	učit	teach
stem + suffix	učit-*el*	teacher
stem + sufix + ending	učit-*el-ovi*	to teacher
prefix + stem	*do*-učit	to have taught
(negative) prefix + stem	*ne*-učit	not to teach
prefix + prefix+ stem + sufix+ sufix + ending	*ne-po*-učit-*el-ný-m*	unteachable

thousands most frequent words yields the coverage rate about 99 %. In general, Slavic languages require ASR vocabularies that are 10 to 20 times larger.

3.3 Free Word Order, But Strong Grammatical Agreement

Another serious complication for automatic speech and text processing is relatively free word order in sentences. A subject of a sentence can occur at an almost arbitrary place, i.e. at the beginning, in the middle as well as at the end of the sentence. The same applies also to a verb or to an adjective. This freedom is possible due to the previously mentioned rich morphology, as the role of the word in the sentence is determined by its inflected form. On the other side, this phenomenon significantly complicates automatic decoding of speech because the decoder cannot rely on any standard word sequence.

Table 5. Fixed word order in English vs. free word order in Czech. All the shown Czech word sequences have the same meaning as the English one.

English	**Czech**
I love Peter.	Mám rád Petra. Rád mám Petra. Petra mám rád.

Table 6. Grammatical agreement in Czech demonstrated on a sentence translated from English (The first translation corresponds to *male students*, the second one to *female ones.*)

English	**Czech**
Two young Czech students were successful in the competition.	**Masculine:** Dv**a** mlad**í** če**ští** student**i** byl**i** úspěšn**í** v soutěži. **Feminine:** Dv**ě** mlad**é** če**ské** student**ky** byl**y** úspěšn**é** v soutěži.

In contrast to the free word order, there is strong grammatical agreement between parts of a sentence in Slavic languages [6]. The inflected form of the sentence subject must agree in gender, person, case and number with the verb and with all the related adjectives, pronouns and numerals. Unlike in English, the agreement is much stronger, as shown in the example in Table 6 where Czech translation of an English sentence significantly differs if the subject, *students*, is of masculine or feminine gender.

4 Solutions Applied to Speech Recognition of Czech

In this section we describe several alternative solutions we investigated during the long term work on ASR systems for Czech language. After explaining them, we present and evaluate some of the developed systems.

4.1 Lexicon

When preparing lexicons for inflected languages, we have to solve one crucial problem: Which words and how many of them should we include in the lexicon. We know that we cannot take all the existing ones, because a) inflected languages allow for producing virtually unlimited number of words, b) even 1 or 2 million words is still a too large inventory to be processed in real time by recent computers. Hence, we have to find a good balance between the lexicon size and the ASR system performance.

During the design of our systems we considered the following strategies:

Lexicon based on most frequent lemmas. The idea was to identify the most frequent basic forms of words (so called lemmas) and to use them for the derivation of all their possible inflected forms. In this approach we utilized the morphologic analyzer and generator developed at Charles University in Prague [6] and the details of the procedure are described in [7]. A set of 74,867 roots (stems + prefixes) together with 5,729 tails (suffixes + ending) produced 972,915 distinct word-forms. On an independent text corpus we found that the coverage rate achieved by this word set was 95.82 %. In other words 4.18 % words in this corpus were out of vocabulary (OOV).

Lexicon based on most frequent word-forms. In this approach we created the lexicon simply from those word-forms that occurred in the training text corpus at least N-times. In our case, the corpus size was 55 million words, N was equal to 3, and the resulting lexicon contained 644,635 items. In spite of its smaller size, the lexicon yielded larger coverage on the independent corpus - 97,07 % [7]. This and the previously mentioned experiments were performed in 2003. Later, we collected much larger training and testing corpora and evaluated the word coverage rate for both the printed (mixture of newspaper and novel text) and spoken documents. The results are shown in Fig. 2. The diagram tells us that if we want to achieve the OOV rate below 2 %, the lexicon should contain at least 300 thousand words.

Lexicon based on morphemes. The very high degree of inflection occurs also in some non-Indo-European languages, e.g. in Finnish, Hungarian, Estonian or Turkish. Linguists and speech researchers from these countries investigated another approach to the lexicon building task. It consisted in the decomposition of words into smaller, morphologically oriented units, morphemes. They should have served as the basic units in the first stage of speech decoding and next, in the second stage, for re-composition of the true words (see e.g. [9]). In this way, even a very large lexical inventory with several million words could be covered by several tens of thousands of morphemes. This approach was tested also for Czech [10], however, the results did not offer any relevant benefit for practical usage. The morpheme based lexicon had three serious drawbacks: 1) it required N-gram language models with $N > 4$ (which complicated fast implementation of the speech decoder), 2) it sometimes produced non-existing words, and 3) it

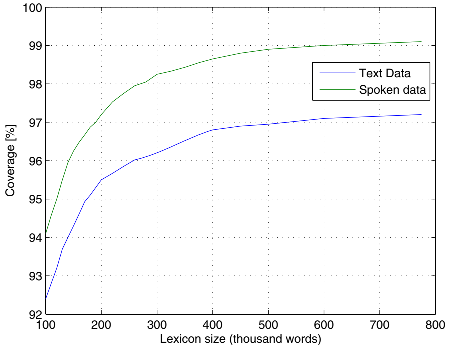


Fig. 2. Coverage rate (for spoken and printed Czech documents) as function of lexicon size (lexicon is made from words ordered by their frequency in the 360M-word corpus)

Table 7. Word error rate and out-of-vocabulary rate as function of lexicon size in Czech broadcast news transcription task

Lexicon size	**Min. word frequency**	**WER[%]**	**OOV[%]**
Lex64K	300	31.5	5.2
Lex100K	140	28.8	3.3
Lex149K	70	26.9	1.9
Lex195K	40	25.9	1.3
Lex257K	20	25.0	1.0
Lex310K	10	24.5	0.8
Lex310K + 1708 multi-word entries	10	23.2	0.8

was based on the assumption that the morphemes are unique and distinctive also from the phonetic point of view, which was not always true for languages like Czech.

Lexicon with added frequent multi-words. The ultimate goal in Slavic language ASR is to find the optimal balance between the lexicon size and ASR system performance. Therefore, we primarily search for the minimal lexicon size that meets the two main performance criteria - recognition accuracy and computation time. However, we found out that the accuracy could be further improved if the vocabulary is enhanced by adding a small number of the most frequently occurring multi-word sequences. This is done also in ASR systems designed for other languages where such collocations, like e.g. "New York" or "Rio de Janeiro" are handled in the same way as single words. We experimented with this strategy and defined several criteria to search automatically for the multi-word candidates [11]. As we show in Table 7, even a small number (1708) of very frequent multi-word entries improves the ASR performance by more than 1 %.

The impact of increasing the size of the lexicon and thus reducing the word error rate (WER) and out-of-vocabulary (OOV) rate is demonstrated in Table 7. The results come from the experiments conducted in a broadcast news transcription project [8].

4.2 Acoustic Modeling and Phonetic Issues

In contrast to English, Slavic languages have an advantage in much more straightforward correspondence between orthography and pronunciation. In general, grapheme-to-phoneme (G2P) transcription can be based on rules, with exceptions applying mainly to the words of foreign origin.

For Czech, we use the phonetic alphabet defined in [12]. It contains 41 phonemes. Non-speech sounds are represented by 7 types of noise. All these 48 elementary acoustic units are modeled by 3-state hidden Markov models (HMM) with distribution functions in form of mixture of Gaussians. During the long-term development process we conducted many experiments to find out the optimal settings of the acoustic model. Our conclusion was that context-independent HMMs (monophones) perform almost as well as the context-dependent ones (triphones), if we use a large number of Gaussian components (at least 100) in each model state. This rather surprising result was achieved under the following conditions: a) using gender-dependent models, b) each trained on approx. 40 hours of carefully annotated training recordings, and c) taking into account the language model. Therefore, in most our systems we employ monophone models (with 100 and more components) and we benefit from much faster and significantly simpler decoding procedure.

All our systems use 39 Mel-frequency based cepstral coefficients (MFCC) computed from 16 kHz sampled signal, parameterized every 10 ms within 25 ms long Hamming window.

Another important issue is accurate phonetic representation of lexicon items. Each word in the lexicon has at least one basic pronunciation form that was either derived automatically by employing a G2P transcriptor or (in case of foreign words) created manually. Many lexicon entries have multiple pronunciations. This applies namely to the words that start with a vowel (then an optional glottal stop is added), those that end with any of the pair consonant (then both voiced and unvoiced final consonants are included), or those that contain a cluster of consonants (in this case also a more colloquial pronunciation form is added). The recent version of the lexicon contains 1.2 pronunciations per word. It is shown (e.g. in [8]) that these additional phonetic variants contribute to almost 10 % relative reduction of the WER.

Another improvement in ASR performance can be achieved if a speaker-fitted acoustic model is used. This is possible namely in dictation applications where the same user is assumed. In other tasks where speakers often change, like in broadcast news transcription, a speaker recognition module can be applied. When speaker's identity is known, the ASR system uses his or her model. In general, a speaker adapted model reduces the WER by 20-25 % relatively.

4.3 Language Model

Nowadays, probabilistic language models (LM) in form of N-grams have become almost a standard in continuous speech recognition. Many ASR systems utilize bigrams (because of their easy implementation), some use also trigrams - either within the decoding procedure itself or for rescoring the hypotheses obtained with bigrams.

For Czech (and for the other Slavic languages as well), an N-gram LM constitutes at least three crucial problems. The first one is the size of the model. When the lexicon size is 10 times larger than in English, the N-gram matrix should be 10^N larger. Even in case of bigrams, it means 100 times more values to be estimated. And this is the second problem. To estimate that huge number of bigram values, one should have much more text data than it is used for an English LM (it means much more than 100 GB of texts). This is impossible simply because such huge corpora do not exist for most languages. The third problem is related to the free word order mentioned in section 3.3. Due to this freedom, an N-gram model built for Czech will never be as reliable as in the languages with a firmer sentence structure.

To cope with the above problems, we investigated several approaches.

Smoothed N-grams. Smoothing is very important because only a small portion of linguistically and semantically possible word sequences is seen in a LM training corpus. The unseen ones must be assigned a small, non-zero probability to give them a chance to be recognized. We experimented with several smoothing techniques, and found the Witten-Bell formula was optimal for the performance as well as well for the implementation of the decoder - for details, see [14] and [15].

Bigrams and multi-words. From the reasons mentioned above we use bigrams in our real-time systems. The bigrams are estimated on a large text corpus. Recently, the corpus contains almost 20 GB of data published in period 1990 - 2009. It is mainly electronic versions of Czech newspaper, further a small portion of professional texts from various fields, some novels available on internet, and transcriptions of broadcast programs (news, debates, talk-shows, parliament sessions, etc.) Even if we utilize the bigrams only, their context span is often longer than just two words. It is because some very frequent multi-word sequences (2 to 3 word long) are included in the lexicon as mentioned in section 4.1. Therefore, quite a lot of bigrams actually cover sequences of 3, 4, 5 and even 6 words. This contributes up to 10 % relative WER reduction [16].

Class based bigrams. In section 3.3 we demonstrated that Czech exhibit quite strong grammatical agreement between the parts of a sentence. Unfortunately, this agreement has very complex nature to be captured by rules that could be applied automatically by a computer. Moreover, some of the rules do not fit to the common left-to-right direction of speech processing. In [15] we proposed a simpler approach. We defined a limited number (less than 500) of classes based on linguistic and morphological categories like nouns, adjectives and pronouns (all in corresponding genders, cases and numbers), further verbs, adverbs, prepositions (classified according to their valences), conjunctions and others. For each lexicon item, we found a mapping between the word and one or more classes - see Table 8. After that we converted the LM training corpus into its class representation and computed class-based bigrams. Their values do not need to be smoothed because a) each class has a lot of its members occurring in the corpus and b) the bigram matrix is much smaller than the classic word-based one. In this way we received information that can be further employed [15], e.g. for conditional smoothing of word bigrams, for searching of possibly grammatically wrong sequences of words or for the fast identification of the words that play a key role in a sentence (a subject, a verb, etc.) Recently, this concept is still under investigation.

4.4 LVCSR Decoder

When developing the decoder for continuous speech recognition of Czech (or any other inflected language), the most challenging task is to manage very large

Table 8. Illustration of a lexicon enriched by additional linguistic class information (It contains examples of a conjunction, a preposition and 2 nouns of different gender, case and number.)

Word	Pronunciation	Linguistic class	Lemma (Basic form)
a	a, á	ConjA	a
⋮	⋮	⋮	⋮
s	s, z, sE, zE	Prep4, Prep7	s
studenta	studenta	NounMasc2S	student
studentek	studentek, studenteg	NounFem2P	studentka

lexicons - with respect to memory, computation speed and efficient usage of language model.

In our case, the decoder was designed for the lexicons that may contain up to 500 thousand recognition items. These items are the phonetic forms of the words and because we employ multiple pronunciations (as explained in section 4.2), the actual lexicon size can go up to 400 thousand lexical entries.

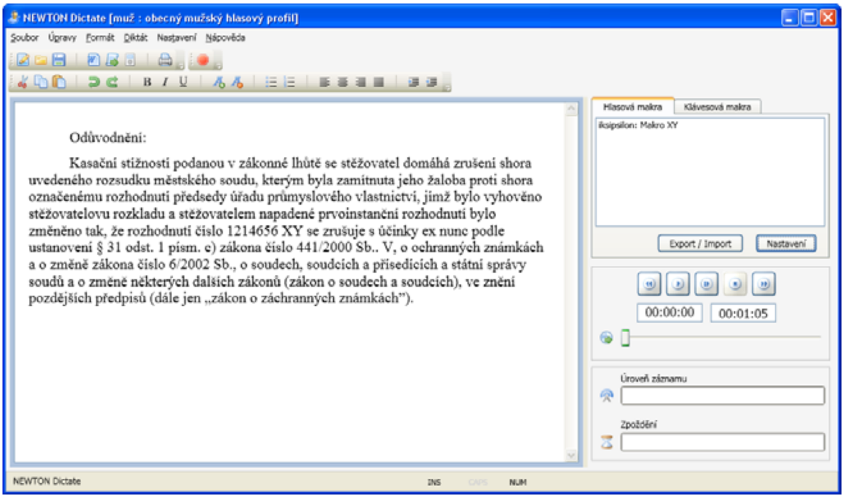
The recent version of the decoder has the following parameters: 1) it can process on-line as well as off-line input data, 2) it is able to manage large lexicons in real-time on currently available PCs and notebooks (a Dual Core CPU running at 2.5+ GHz is required), and 3) it can produce text output in continual way. These features allow for employing the same engine for various tasks: from voice dictation to transcription of on-line captured or previously recorded speech data.

The decoder has been extensively optimized for the lexicons of the desired size. Special effort has been made to speed up computation of Gaussian likelihoods and their efficient caching, to predict the most promising hypotheses at one side and to prune off the non-promising ones at several levels, or to handle bigram matrices effectively. Some of the techniques are mentioned in [14].

4.5 Practical Applications - Czech Voice Dictation Systems

During the last five years, we have developed two types of voice dictation programs for Czech language. Their prototypes were presented already in 2005 [17] but it took us three more years to make them available as commercial products.

The first program, called **MyDictate**, was developed for a special target group: people whose motor handicap prevents them from using computer keyboard and mouse. Therefore, the whole program environment had to be designed as hands-free. In order to simplify all voice typing, correcting, editing and formatting actions, the program uses a discrete speech input [18]. This means that words and phrases are dictated with short pauses between them. At one side, the dictation is slower and less natural, but on the other side, the user can correct - entirely by voice - misrecognized or mispronounced words immediately after they occur. Moreover, because the discrete speech decoder requires significantly less computation, the lexicon can be larger (currently 570K words) and the program runs even on a mid-level PC or notebook.



Odůvodnění:

Kasační stížnosti podanou v zákonné lhůtě se stěžovatel domáhá zrušení shora uvedeného rozsudku městského soudu, kterým byla zamítnuta jeho žaloba proti shora označenému rozhodnutí předsedy úřadu průmyslového vlastnictví, jimž bylo vyhověno stěžovatelovu rozkladu a stěžovatelem napadené prvoinstanční rozhodnutí bylo změněno tak, že rozhodnutí číslo 1214656 XY se zrušuje s účinky ex nunc podle ustanovení § 31 odst. 1 písm. c) zákona číslo 441/2000 Sb.. V, o ochranných známkách a o změně zákona číslo 6/2002 Sb., o soudech, soudcích a přísedících a státní správy soudů a o změně některých dalších zákonů (zákon o soudech a soudcích), ve znění pozdějších předpisů (dále jen „zákon o záchranných známkách”).

Fig. 3. A snapshot showing NewtonDictate with dictated juristic text

Table 9. Czech dictation in 4 different domains. (All the texts were read in quiet conditions)

Domain	#words in test	WER[%]
Newspaper articles	15492	4.53
Medical text - radiology	7304	2.26
Medical text - pathology	17107	3.16
Juristic text	6612	1.44

The second program is named **NewtonDictate** and it allows for fluent speech input. Its commercial version was developed in collaboration with Czech company Newton Technologies. The layout of the program is shown in Fig. 3. Currently, the software is distributed in three versions. One comes with a general purpose lexicon, the other two are domain specific. The first one contains 350K words and its language model has been trained mainly on newspaper text, which means that it can be used for most common situations. The second one is juristic and it was designed for judges and lawyers. Since summer 2008 it has been tested in Czech courts. The third domain is medicine and recently there exist two fields where the program is used: radiology and pathology.

The performance of the NewtonDictate program was measured on a large set of recordings provided by several tens of test subjects. The results from these evaluations are presented in Table 9.

4.6 Practical Applications - Broadcast Speech Transcription System

The second application of the voice technology is a broadcast speech transcription system. It is a complex modular platform that can be configured for several operational modes, such as a server producing sub-titles for TV programs, an off-line transcriptor for already recorded TV and radio shows, or a system that provides text data for broadcast monitoring, indexing and search services.

The system is composed of modules shown in Fig. 4. The first module processes acoustic track of the broadcast data and converts it into signal parameters (MFCC feature vectors). The second module searches for significant changes in audio characteristics and splits the stream into non-speech segments and segments spoken by a single person. (For the transcription of TV shows we can utilize also the video part of the signal [21].) The next module tries to determine the identity of the speakers. If he or she is in the database (of some 300 frequently occurring persons) and is correctly recognized and verified, his/her speaker adapted acoustic model is used in the speech recognition module. In other cases, at least the speaker's gender is identified and the proper gender dependent model is applied. The last module provides final text post-processing. This means conversions of number strings into corresponding digit strings, capitalization of proper names, text formatting, etc.

In Table 10 we show some relevant results from performance evaluation tests, conducted in 2007. In these tests, always the whole shows were processed and transcribed. In case of TV news, for example, it means that not only studio

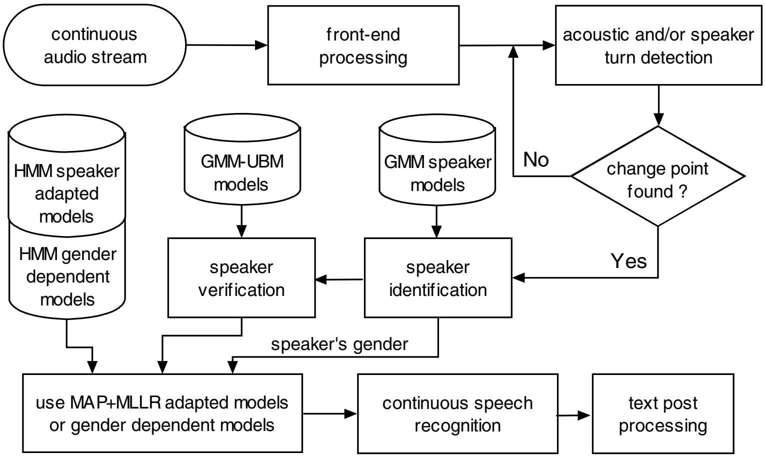


Fig. 4. Modular architecture of the broadcast speech transcription system

Table 10. Results from automatic transcription of several types of broadcast programs

Program Type	**Time**[min]	#**Words**	**WER**[%]	**OOV**[%]
TV News	62	9760	21.5	1.1
Weather news	10	1745	9.4	0.4
Talk shows	118	13624	33.8	0.6
Parliament sessions	52	6395	20.8	1.1

speech, but also news headlines with music played in background and shots recorded in noisy environments are included in the total WER values.

5 Voice Technology Transfer to Slovak Language

In the situation where we have programs, systems and modules that work for one language it is natural to ask if the same approach could be applied for another language. We have already had one positive experience with porting voice controlled software to other languages. It was program MyVoice that we had developed originally for Czech handicapped users and later modified for Slovak and also for Spanish [22]. In that small project we have verified procedures that seem to be applicable for larger systems, too. In the following sections we present our recent experience with transferring voice technology from Czech to Slovak, i.e. between two rather closely related Slavic languages.

5.1 Lexicon

Czech and Slovak belong to the same West-Slavic branch of languages. They are considered very similar and closely related because in the past they were official languages used within one state (former Czechoslovakia). In order to quantify their similarity on the lexical level we have analyzed a large amount of parallel corpora, in our case documents of the European Union published in both the languages. We have found that about 25 % of lexical items are same in Czech and Slovak. Among the remaining ones there is still a large group of words that are very similar (differing in 1 or 2 letters only).

For creating a representative lexicon of contemporary Slovak and for computing the corresponding language model we were provided by a corpus of newspaper articles and broadcast news transcriptions from the 2005 - 2007 period. Its

Table 11. Slovak-specific phonemes (with corresponding orthography) mapped onto acoustically closest Czech ones (SAMPA notation is used in this table)

SK letter(s)	SK phoneme	CZ phoneme(s)
ä	{	e
ľ	L	l
ĺ	l=:	l
ŕ	r=:	r
v	U_^	u
v	w	v
h	G	h
j	I_^	j
ô	u_^o	uo
ia	I_^a	ja
ie	I_^e	je
iu	I_^U\	ju

size was 1.9 GB. After cleaning it (which included also detecting and removing Czech texts frequently occurring in Slovak media), we compiled the first version of the Slovak lexicon. It was made of the 166,535 most frequent words and wordforms. To get pronunciations for them, we modified our grapheme-to-phoneme converter by including Slovak specific phonetic rules described in [23].

Because we knew that the available amount of Slovak acoustic data had not been large enough to train an independent Slovak acoustic model, we had to utilize (and later adapt) the existing Czech AM. Therefore, it was necessary to make conversion from Slovak phonetic inventory to the Czech one. This was done by mapping the Slovak-specific phones and diphthongs into their closest Czech counterparts, either single phonemes or phoneme strings. The mapping rules are summarized in Table 11.

5.2 Acoustic Model

The initial experiment in Slovak speech recognition was done with the Czech acoustic model. The results were surprisingly good - about 25 % WER on the broadcast news task [24]. Anyway, the next natural step was the adaptation of the model. We used 6 hours of annotated Slovak speech and added them to the 61-hour Czech training database. On that data we trained new male and female (i.e. gender dependent) acoustic models. They were used in the experiments whose results are summarized in Tables 12 and 13.

5.3 Language Model

The language model was trained on the 1.9 GB Slovak corpus mentioned in section 5.1. In the corpus we found 234 million occurrences of the 166K-word lexicon items. A bigram LM was computed from 32 million word-pairs seen in the corpus by applying the Witten-Bell smoothing technique. We employed the same tools and the same procedures that had been developed previously for Czech.

5.4 Prototypical Applications - Slovak Voice Dictation System

Because all the Slovak specific modules, i.e. the lexicon, the language model and the acoustic model are fully compatible with the Czech ones, we can use all the voice technology software we developed so far.

In the first series of experiments we tested dictation into the NewtonDictate program. Four Slovak native speakers (two men and two women) were asked to read several articles from Slovak newspapers. The articles were selected to cover various topics (mainly domestic and international news). The results from these initial tests are shown in Table 12. If we compare the values with those in Table 9 (on the first line), we see that the performance of the Slovak version is significantly poorer. There are several reasons for that: 1) the size of the Slovak lexicon is only a half of the Czech one, 2) also the corpus for Slovak language

Table 12. Initial tests of Slovak dictation with NewtonDictate software

Speakers	#words in test	WER[%]
Male speakers	3749	13.45
Female speakers	3811	12.10

model training was much smaller, 3) the Slovak lexicon and language model do not utilize multi-word entries, yet, and 4) the acoustic model for Slovak was trained on the data in which Czech speech prevailed in 10:1 ratio. In future, we will focus on eliminating all these negative factors.

5.5 Prototypical Applications - Slovak Broadcast New Transcription

The second application field where we tested the existing Slovak specific modules was automatic transcription of broadcast news. We used the same system as it is described in section 4.6 with one exception: We could not apply speaker recognition and speaker adaptation modules because we had not had the speaker database that was necessary for their proper function.

The experiments were performed on data collected during March and April 2008. We recorded and annotated a test set consisting of eight complete news shows from three major TV stations and one nation-wide radio. Like in the Czech experiments mentioned in section 4.6, all parts the 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. More details about the experiments can be found in [24]. For further improvement, the same issues mentioned in the previous section should be solved.

Table 13. Experiments with Slovak broadcast news transcription

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

6 Conclusions

Slavic languages certainly pose a big challenge for researchers dealing with voice technologies, particularly speech recognition. Their inflected nature and rich morphology result in extremely large vocabularies, whose size (up to several million different word-forms) exceeds the limits of today's common algorithms and available PCs. Therefore, it is necessary to search the methods that can cope with extremely large lexicons or that can choose optimal subsets of those huge lexical inventories. In language modeling, the problem is even more serious. If the classic N-gram approach is used, the LM matrices grow towards extreme sizes, but it also means that they require much more data for reliable estimation. In most languages that size of required amount of training data is not available. Hence, some other methods or modifications of the existing ones must be searched.

In spite of these problems, even on recent computers it is possible to implement solutions that are applicable in practice, e.g. in dictation programs or in systems for automatic transcription of spoken documents. We present the solutions that proved to work for speech recognition of Czech and that seem to be portable also for other Slavic languages. Examples of alternative approaches applied to other languages like, e.g., Slovene, Slovak or Polish can be found in [25,26,27].

Acknowledgement

This work was supported by grants no. 102/08/0707 by the Grant Agency of the Czech Republic and no. OC09066 by the Ministry of Education of the Czech Republic (within EU Action COST 2102).

References

1. http://www.research.ibm.com/hlt/html/body_history.html
2. Gauvain, J.L., Lamel, L., Adda, G., Jardino, M.: The LIMSI 1998 HUB-4E Transcription System. In: Proc. of the DARPA Broadcast News Workshop, Herndon, pp. 99–104 (1999)
3. Os, E., Boves, L., Lamel, L., Baggia, P.: Overview of the ARISE Project. In: Proceedings of Eurospeech 1999, Budapest, pp. 1527–1530 (1999)
4. Tan, Z.-H., Lindberg, B. (eds.): Automatic speech recognition on mobile devices and over communication networks. Springer, London (2008)
5. Tronconi, A., Billi, M.: New technologies for physically disabled individuals. European Transactions on Telecommunications (6), 633–640 (2008)
6. Hajic, J.: Disambiguation of Rich Inflection-Computational Morphology of Czech. Karolinum Charles University Press, Prague (2004)
7. Nejedlova, D., Nouza, J.: Building of a Vocabulary for the Automatic Voice-Dictation System. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 301–308. Springer, Heidelberg (2003)
8. 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: Proc. of Interspeech 2005, Lisbon (September 2005)
9. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J.: Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish. Computer Speech & Language 20(4), 515–541 (2006)
10. Byrne, W., Hajic, J., Ircing, P., Krbec, P., Psutka, J.: Morpheme Based Language Models for Speech Recognition of Czech. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2000. LNCS (LNAI), vol. 1902, pp. 139–162. Springer, Heidelberg (2000)

11. Kolorenc, J., Nouza, J., Cerva, P.: Multi-words in the Czech TV/radio News Transcription system. In: Proc. of Specom 2006 conference, St. Petersburg, pp. 70–74 (2006)
12. Nouza, J., Psutka, J., Uhlir, J.: Phonetic Alphabet for Speech Recognition of Czech. Radioengineering 6(4), 16–20 (1997)
13. Cerva, P., Nouza, J.: Supervised and unsupervised speaker adaptation in large vocabulary continuous speech recognition of Czech. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 203–210. Springer, Heidelberg (2005)
14. Nouza, J.: Strategies for developing a real-time continuous speech recognition system for czech language. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2002. LNCS (LNAI), vol. 2448, pp. 189–196. Springer, Heidelberg (2002)
15. Nouza, J., Drabkova, J.: Combining Lexical and Morphological knowledge in language model for Inflectional (Czech) Language. In: Proc. of 6th Int. Conference on Spoken Language Processing (ICSLP 2002), Denver, September 2002, pp. 705–708 (2002)
16. Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: Continual On-line Monitoring of Czech Spoken Broadcast Programs. In: Proc. of 7th International Conference on Spoken Language Processing (ICSLP 2006), Pittsburgh, September 2006, pp. 1650–1653 (2006)
17. Nouza, J.: Discrete and Fluent Voice Dictation in Czech Language. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 273–280. Springer, Heidelberg (2005)
18. Cerva, P., Nouza, J.: Design and Development of Voice Controlled Aids for Motor-Handicapped Persons. In: Proc. of Interspeech, Antwerp, pp. 2521–2524 (2007)
19. `http://www.v2t.cz/newton-media.php`
20. Nouza, J., Zdansky, J., Cerva, P., Kolorenc, J.: A system for information retrieval from large records of czech spoken data. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 401–408. Springer, Heidelberg (2006)
21. Chaloupka, J.: Visual Speech Segmentation and Speaker Recognition for Transcription of TV News. In: Proc. of Interspeech 2006, Denver, September 2006, pp. 1284–1287 (2006)
22. Callejas, Z., Nouza, J., Cerva, P., López-Cózar, R.: Cost-efficient cross-lingual adaptation of a speech recognition system. In: Advances in Intelligent and Soft Computing. Springer, Heidelberg (2009)
23. Ivanecky, J.: Automatic speech transcription and segmentation. PhD thesis, Kosice (December 2003) (in Slovak)
24. Nouza, J., Silovsky, J., Zdansky, J., Cerva, P., Kroul, M., Chaloupka, J.: Czech-to-Slovak Adapted Broadcast News Transcription System. In: Proc. of Interspeech 2008, Brisbane, September 2008, pp. 2683–2686 (2008)
25. Rotovnik, T., Sepesy Maucec, M., Kacic, Z.: Large vocabulary continuous speech recognition of an inflected language using stems and endings. Speech Communication 49(6), 437–452 (2007)
26. Pleva, M., Cizmar, A., Juhár, J., Ondas, J., Michal, M.: Towards Slovak Broadcast News Automatic Recording and Transcribing Service. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 158–168. Springer, Heidelberg (2008)
27. Korzinek, D., Brocki, L.: Grammar Based Automatic Speech Recognition System for the Polish Language. In: Recent Advances in Mechatronics. Springer, Heidelberg (2007)