Automatic Transcription of Czech, Russian, and Slovak Spontaneous Speech in the MALACH Project

Josef Psutka[1], *Pavel Ircing*[1], *J.V. Psutka*[1], *Jan Hajič*[2], *William J. Byrne*[3], *Jiří Mírovský*[2],

[1]University of West Bohemia, Department of Cybernetics, Univerzitní 8,
306 14 Plzeň, Czech Republic
e-mail:{psutka, ircing, psutka_j}@kky.zcu.cz

[2] Charles University, Center for Computational Linguistic, Praha, Czech Republic
e-mail:{hajic, mirovsky}@ufal.mff.cuni.cz

[3] Johns Hopkins University, Center for Language and Speech Processing, Baltimore, MD
e-mail: wjb31@cam.ac.uk

10.21437/Interspeech.2005-489

Abstract

This paper describes the 3.5-years effort put into building LVCSR systems for recognition of spontaneous speech of Czech, Russian, and Slovak witnesses of the Holocaust in the MALACH project. For processing of colloquial, highly emotional and heavily accented speech of elderly people containing many non-speech events we have developed techniques that very effectively handle both non-speech events and colloquial and accented variants of uttered words. Manual transcripts as one of the main sources for language modeling were automatically „normalized" using standardized lexicon, which brought about 2 to 3% reduction of the word error rate (*WER*). The subsequent interpolation of such LMs with models built from an additional collection (consisting of topically selected sentences from general text corpora) resulted into an additional improvement of performance of up to 3 % .

1. Introduction

In our previous papers we described the building of the baseline Czech LVCSR system [1] and the development of speech and language sources for the other two languages: Russian and Slovak [2]. The background of the whole project including its English part, which is being investigated at IBM, was introduced in [3].

This paper brings brand new results that we obtained on the baseline Russian and Slovak LVCSR systems and compares them with the latest performance of the Czech system. In order to effectively handle the large number of non-speech events we developed a technique for acoustic modeling that uses a special model trained on a very large number of non-speech sounds being manually annotated in training data. Colloquial and accented speech of survivors is treated using training and decoding procedure that considers colloquial and accented words to be pronunciation variants of grammatically correct words. The procedure involves a manual standardization of the lexicon and a consequent automatic „normalization" of manual transcripts. The normalization of manual transcripts not only made the parameters of the estimated language model more robust but also brought this main and most useful source for language modeling much closer to standard text sources (e.g. newspaper articles). This resulted into improved performance of the language model constructed as an interpolation of the model built from the manual transcripts and the model estimated using the collection of topically selected sentences from the general text corpus.

2. Characteristics of speech corpora

Testimonies of holocaust survivors are deposited at the VHF digital library as video interviews. The speech of each interview participant – the interviewer and interviewee – was usually recorded in quiet rooms via lapel microphones that recorded speech on separate channels. The speech quality in individual interviews is however very poor from the ASR point of view, as it contains whispered or emotional speech with many disfluences and non-speech events as crying, laughter etc.

The quality and fluency of speech was often affected by the age of speakers (the average age of all speakers was about 75 years), by various accents (mainly in spoken Russian), by using many colloquial (non-grammatical) words (in Czech) and also by a long term mutual influence of Czech and Slovak due to the common state of both nations. Unlike English, where the accent is because of the native tongues of the speakers, much of the Russian accents are due to regional differences in spoken Russian (regional variants of pronunciation). We found out that this accent is usually caused by the territory where the survivors are now living and where they were interviewed. Studying the demographic information provided by the VHF we learned that from about 7 thousand of Russian testimonies stored in the VHFs digital archives nearly one half (3,500) were provided in Ukraine, about 1,500 in Israel, 900 in U.S.A., and only 700 in Russia. The native Russians living outside Russia often adopted local non-Russian words and used them in their personal vocabulary.

At the VHF digital library the testimonies are divided into half-hour parts stored as MPEG-1 video files. We extracted the audio stream at 128kb/sec in 16-bit resolution and 44 kHz sampling rate. For all three languages – Czech, Russian, and Slovak – we decided to randomly select and manually transcribe 4 hundred 15-minute speech segments of individual speakers (for training purposes) and whole testimonies of 10 different survivors (about 20 hours of speech) for tests. Unfortunately only about 346 Czech testimonies were digitized at the VHF so we had to content ourselves with only 336 training and 10 test testimonies for building Czech ASR.

3. Czech, Russian, and Slovak phonetics

Although Czech and Slovak people for example understand each other relatively well and one can think that all Slavic languages are very similar it can be demonstrated that at least phonetics of Czech, Russian and Slovak are quite different.

3.1. Phonetic inventories

A phonetic inventory of Slovak contains 52 phonemes while Russian and Czech incorporate only 43 and 42, respectively. Table 1 shows numbers of vowels, consonants and diphthongs for all three languages. There are many phonemes belonging to the given language which don't have their counterparts in the second two languages, e.g. palatalized plosives in Russian (набирать, лагерь, банкир, опять etc.), diphthongs [i^e] (mier) or [u^o] (kôň) in Slovak etc. An absence of long vowels in Russian is also very interesting. As well the phoneme „h" common in Czech and Slovak is not a standard member of the Russian phonetic alphabet. Native Russians living in Russia usually replace it in words of foreign origin or in personal and geographical names by the phoneme „g". The native Russians living in the Ukraine territory as well as in Israel or USA learned to pronounce „h" and they use this phoneme frequently in words in which it is currently used in the local languages (for example in geographical names, personal names atc.). Because many survivors who yielded their testimonies lived out of Russian territory it was a reason why we had to add the phone „h" to the standard Russian phonetic inventory.

Table 1: No. of phonemes in Czech, Russian, and Slovak.

	Vowels	Consonants	Diphthongs
Czech	10	29	3
Russian	6	37	0
Slovak	11	37	4

3.2. Rule-based phonetic transduction

For all three languages we developed rule-based phonetic transduction, which are used to automatically transform the majority of the words in the transcriptions to their phonetic forms. Many words have two or more possible (correct) pronunciation variants. For example pronunciation of many Slavic words ending with some voiced pair consonants can be influenced by cross-word (voiced) assimilation therefore all these possible phonetic variants are also put into lexicon.

There are also many words, which are treated as exceptions to the pronunciation rules and those words must be transcribed manually [2]. The majority of exceptions to the Russian phonetic rules can be found among words containing the character -o-. If the position of this character in the word is before the stress then „o" is actually read as „a" (e.g. "Москва", Engl. „Moscow", is uttered as [m a s k v a] because the stress is on the "a"). Most exceptions to the rules of phonetic transcription in Czech and Slovak are connected with words containing prealveolar stops in sequences of characters: -ti-, -di-, and -ni-. These sequences are uttered as alveopalatal stops – [tj i], [dj i], and [nj i] in words of Czech origin whereas as prealveolar stops (e.g. "automatizace") in words of non-Czech origin.

4. Annotation of spontaneous speech

4.1. Annotation rules

Audio files were divided into segments corresponding roughly to a sentences and annotated using a annotation software Transcriber 1.4.1. (http://www.ldc.upenn.edu/). For processing of Russian testimonies this tool was adapted so that transcriptions in the Cyrillic alphabet could be encoded. The annotation rules were the same for all three languages. Fifteen types of non-speech events were used during annotation (<click>, <mauth>, <cough>, <laugh>, <uh>, <um>, <unt>, <hm>, <unintelligible>, <breath>, <inhale>, <silence>, <noise>, <noise_begin>, and <noise_end>). Human annotators worked at a rate from twelve (for Slovak) to twenty times (for Russian) real time. Transcription inspection and verification requires addition effort at least two or three times real time.

4.2. Annotation of colloquial & accented speech

All manual annotations were performed in the orthographic form of the words. This means that the eventual colloquial words were neither transformed to standard (formal, non-colloquial) forms nor written phonetically. Czech colloquial words are usually not considered to be phonetic variants of standard Czech words in that they can be properly written in their colloquial orthographic form. For example, the standard Czech word „oběd" (Engl. „lunch") has pronunciations [o b j e t] and [o b j e d]. If we wish to write this word phonetically, then we obtain „objet", but this form is used neither in standard nor in colloquial Czech. But, there does exist the standard Czech word „objet" (Engl. „to go round"). Similarly, the word „oběd" has also a colloquial variant, „voběd" with the two pronunciations [v o b j e t] and [v o b j e d].

In spontaneous Russian we observed problems with regional variants of pronunciation of many words. The main differences appear in different pronunciation of one or more characters in the word in comparison with standard Russian. For example the Russian word "когда" has the standard pronunciation according to the phonetic transcription [k a g d a] but many times this word was pronounced as [k o g d a]. The native Russian transcribers assessed these words not as colloquial words and/or only accented speech but rather as a speech of Russians whose pronunciation is partly modified by a non-Russian environment (Ukraine, Israel, etc.) where may have lived for a long time. These instances marked by placing the incorrectly pronounced portion of words between asterisks, as in "к*о*гда", and the region in question was transcribed phonetically.

Many pronunciation problems encountered in speech of Slovak survivors could be explained by a long term mutual influence of Czech and Slovak due to the common state of both nations (e.g. Czech endings in Slovak words, Czech pronunciation of a part of Slovak words etc.). All these disfluencies could be annotated in the orthographic form of the words (similar as in the Czech part).

4.3. Annotation statistics

Analyzing manual annotations enabled us to compute some interesting statistics, that can compare speech characteristics of groups of Czech, Russian, and Slovak speaking survivors. The speaking rate, for example, measured as the number of words uttered per minute, varies greatly depending on the speaker, changing from 60 to 180 with the average introduced for individual languages in Table 1. The next two statistics (non-speech sounds rate and #tokens/#non-speech sounds) have a bit attitudinal nature characterizing both the quality of speech and sensitivity of hearing of human annotators.

Tab 2: Speech properties in the corpora

	Czech	Russian	Slovak
Rate of speech [words / min]	122	132	125
Non-speech sounds rate [# / min]	22.3	23.4	16.9
# tokens/# non-speech Sounds	5.5	5.7	7.4

Figure 1 demonstrates other interesting but for Slavic languages very typical phenomenon. The problem is usually an insufficient coverage of test utterances by the recognition lexicon, which corresponds to a high level of OOV_rate. Behaviour of all three languages was in our task very similar. The extent of lexicons created from 600k tokens (running words) moves around 43k words however these lexicons provided OOV_rates only no lower than from 5% .

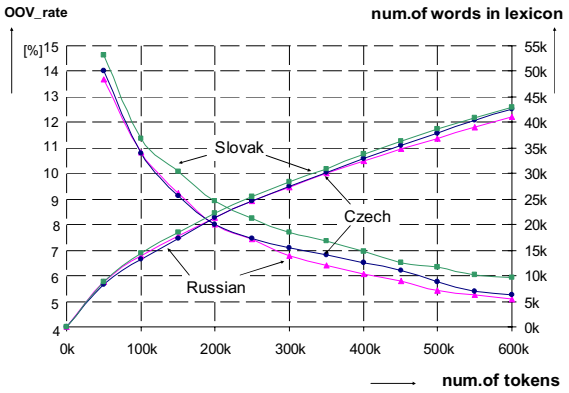

Figure 1: Word statistics for increasing size of transcripts

5. Treatment of colloquial and accented speech

We have found out that the different problems present in the processed languages (frequent occurrence of colloquial words in Czech, accented speech in Russian) can be successfully mitigated using a common approach – some form of the lexicon standardization.

The idea originated during the development of the Czech ASR system. It is well-known that the present-day Czech exhibits a substantial difference between standard and colloquial form of the language. Whereas standard Czech is used in most of Czech written materials as well as in official public speeches, colloquial Czech is widely used in spontaneous everyday communication. The difference between just pronunciation variants (as found in English and many other languages) and the Czech case is that Czech spelling rules are phonetically based. Therefore, the colloquial Czech words have well-defined, but different spelling than their standard variants.

The result of this phenomenon is an occurrence of many spelling variants of a single word in the lexicon. Since a word is defined purely by its spelling in the ASR domain, the already sparse language model training data become even sparser. One could argue that we should foresee this problem and transcribe the speech using the spelling of the standard variant directly. However, such approach would rule out the usage of our phonetic transducer (see Section 3.2), since it relies on accurate (phonetically based) transcription. Thus we have decided to preserve the spelling of the colloquial variants and append additional column with the spelling of the corresponding standard variant to the lexicon. Such approach allows us to "normalize" the transcripts and consequently use the resulting parallel corpora (original and normalized) for counting of the relative frequencies of the individual colloquial variants. Here we present an excerpt from the standardized lexicon (columns contain standard form, colloquial form, phonetic baseform of the colloquial form and the relative frequency of the colloquial form, respectively):

ODJET	ODEJET	o d e j e t	0.0161
ODJET	ODJEC	o d j e c	0.0161
ODJET	ODJECT	o d j e c t	0.0483
ODJET	ODJET	o d j e t	0.2741
ODJET	VODJECT	v o d j e c t	0.0169
ODJET	VODJET	v o d j e t	0.0967

The column with standard word forms then constitutes the lexicon of the language model (which is of course estimated using the normalized transcripts) and of the decoder. Colloquial word forms are treated as pronunciation variants, with a weight corresponding to the aforementioned relative frequencies. The usage of standard word forms makes the language model more robust, while the presence of colloquial spelling variants allows, besides the automatic generation of the phonetic baseforms, more accurate acoustic modeling. More detailed analysis together with several experimental results can be found in [4].

Although other language in question, Russian, does not exhibit a significant usage of colloquial words, the proposed technique of lexicon standardization can be exploited for treating the non-standard and/or accented pronunciations of the words in the Russian corpus. Such unusually pronounced words were marked by asterisks so that they could be excluded from the automatic phonetic transcription; the corresponding phonetic baseforms were generated manually. However, from the language model point of view, all the pronunciation variants of a word should be represented by a single type. Thus we again constructed a multi-column lexicon with the same layout as for Czech, only the second column now represents the spelling of the non-standard pronunciation variant instead of the colloquial word form. Here is the snippet:

москва	москва	m a s k v a	0.8000
москва	м*о*сква	m o s k va	0.2000
вдвоём	вдвоём	v d v a j o m	0.9394
вдвоём	вдв*о*ём	v d v o o m	0.0303
вдвоём	*двоём*	d v a j o m sp	0.0303

The way of treating the lexicon in the training and the decoding process is the same as in the Czech system.

When building the Slovak ASR system, we decided to create the same type of the lexicon purely in order to corroborate our original hypothesis that there are no problems with non-standard word forms in Slovak. In theory, the Slovak language was not supposed to exhibit neither the massive usage of colloquial words nor the frequent occurrence of non-standard pronunciations. However, after comparing the number of distinct types in the first and the second column in the lexicon, we obtained approximately the same ratio as for the Czech and Russian. We discussed this fact with several native Slovak speakers who performed the data preparation and they said that they observed many words influenced by the Czech language (mainly in interviews from speakers living in the western part of Slovakia) or by various dialects. Also speakers from the Hungarian minority, even though they mastered the Slovak grammar, have problems with pronouncing some of the Slovak phonemes. Consequently, the standardization of the lexicon had also a very positive effect on the performance of the Slovak ASR system.

The following table shows some lexicon and corpora statistics for all 3 languages. The first row depicts the average number of baseforms per one word type in the lexicon created from the original transcripts – multiple baseforms stem mainly from the different possible manifestations of the assimilation phenomenon. The second row shows the average number of

baseforms per one word type in the standardized lexicon – here you can see the influence of colloquial, accented or other non-standard words. Finally, the last two rows illustrate the percentage of such non-standard words in the lexicon and the transcripts, respectively. A reader should notice that whereas both the average number of baseforms and the percentage of non-standard words in the lexicon are comparable for all languages in question, the relative frequency of Czech non-standard words in the text is much higher than for the other two languages; this fact confirms our earlier statements regarding the frequent usage of colloquial words in the spontaneous Czech speech.

Tab 3: Lexicon and corpora statistics

	Czech	**Russian**	**Slovak**
# of phonetic variants per 1 word	1.20	1.15	1.14
# of pronounciation variants per 1 word (after standard.)	1.31	1.28	1.26
rel # of colloquial & accented words in lexicon [%]	9.12	8.97	8.99
rel # of colloquial & accented words in transcripts [%]	8.55	4.69	2.14

6. Recognition experiments

6.1. Acoustic modeling

The acoustic training portion consisted of 84 hours of Czech speech and of 100 hours of Russian and Slovak. The data was parameterized as 17 dimensional PLP cepstral features including their delta and delta-delta derivatives (3x17=51 dimensional feature vectors). These features were computed at rate of 100 frames per second. Cepstral mean subtraction was applied per utterance. The resulting triphone-based models were trained using HTK Toolkit and had approximately 6k states and more than 100k Gaussians (exactly 107k for Czech, 113k for Russian, and 126k for Slovak).

6.2. Handling of non-speech events

As was already mentioned, all non-speech events appearing in spontaneous speech of survivors were annotated very carefully. We used these annotated events (with the exception of events marked as <unintelligible> speech) to train a generalized model of silence, separately for each language. For these purposes we used following numbers of non-speech parts: 108k for Czech, 125k for Russian, and 99k for Slovak. The generalized model of silence is a 3-state HMM equipped with nearly 150 Gaussians and „catches" most of standard non-speech events appearing in running speech very well, which reduces the *WER* about 2 to 4% .

6.3. Language modeling

Experiments with 3 sets of language models were performed for each language on 500 sentences randomly selected from test data. All models are standard word-based bigrams with Katz's backing-off scheme, estimated using the SRILM toolkit. The first set of models (**ORIG**) was trained on the original transcripts, i.e. with lexicon and data containing also non-standard words. The second set of models (**STAN**) was estimated using the standardized version on the data, with weights assigned to the individual pronunciation variants in the lexicon. Finally the last set of models (**INTER**) consists of models constructed as an interpolation of the **STAN** model built from the manual transcripts and the model estimated using the collection of topically selected sentences from the general text corpus. The process of sentence selection was described thoroughly in [1]. Results obtained with the models trained solely on the selected sentences or even on the general newspaper corpus are not reported, since they are significantly worse then the baseline results achieved using the manual transcripts. The results in terms of *WER* are summarized in the following table, together with the lexicon size (*Lex.*) and training transcripts perplexity of the individual models (*PPL*), and the size of the training corpora – manual transcripts (**TRA**) and collection of the selected sentences (**SEL**).

Tab 4: Recognition results

		ORIG	**STAN**	**INTER**	**TRA**	**SEL**
Czech	*Lex.*	46k	42k	79k	606k	15.8 M
	PPL	126	120	153		
	WER	42.99	41.15	38.57		
Russian	*Lex.*	50k	45k	82k	643k	10.5 M
	PPL	125	122	152		
	WER	50.82	46.82	45.75		
Slovak	*Lex.*	50k	45k	83k	649k	13.5 M
	PPL	116	114	158		
	WER	40.69	38.09	34.49		

7. Conclusions

Our paper presents all stages of building the LVCSR systems for transcribing Czech, Russian and Slovak testimonies in the MALACH project. We have developed techniques tailored to the spontaneous speech that have brought consistent improvement in all of the processed languages. Whereas Czech and Slovak results are fully comparable with the results achieved for English [3], the overall performance of the Russian system is somewhat lower. The reason might lie in the massive presence of the accented speech in the corpus and the conjunctive inadequacy of the defined phonetic alphabet and/or transcription process. This hypothesis is currently a subject of intensive research.

8. Acknowledgements

This work has been funded by the NSF (U.S.A.) under the Inforamtion Technology Research (ITR) program NSF IIS Award No.0122466, and by the Ministry of Education of the Czech Republic projects LC536 and 1P05ME786.

9. References

[1] Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J., and Gustman, S. "Large Vocabulary ASR for Spontaneous Czech in the MALACH Project", in: *Eurospeech'2003*, *Geneva, Switzerland, pp. 1821-1824.*

[2] Psutka, J., Hajič, J., Byrne, W. "The Development of ASR for Slavic Languages in the MALACH Project", in: *ICASSP'2004*, *Montreal, Canada, pp.III-749 – III-752.*

[3] Byrne, W., et. al. "Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives", *IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, July 2004, pp.420-435.*

[4] Psutka, J., Ircing, P., Hajič, J., Radová, V., Psutka, J.V., Byrne, W., Gustman, S. "Issues in annotation of the Czech spontaneous speech corpus in the MALACH project", in: *LREC 2004, Lisboa, Portugal*.