# Building LVCSR System for Transcription of Spontaneously Pronounced Russian Witnesses in the MALACH Project: Initial Steps and First Results*

Josef Psutka[1], Ilja Iljuchin[1], Pavel Ircing[1], Josef V. Psutka[1], Václav Trejbal[1], William Byrne[2], Jan Hajič[3], and Samuel Gustman[4]

[1] University of West Bohemia, Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{psutka, illusion, ircing, psutka_j}@kky.zcu.cz, trejbal@ksj.zcu.cz
[2] Johns Hopkins University, Center for Language and Speech Processing
309 Barton Hall, 3400 N. Charles St., Baltimore, MD 21218
byrne@jhu.edu
[3] Charles University, Institute of Formal and Applied Linguistic
Malostranské náměstí 25, 118 00 Praha, Czech Republic
hajic@ufal.mff.cuni.cz
[4] Survivors of the Shoah Visual History Foundation
P.O. Box 3168, Los Angeles, CA 90078-3168
sam@vhf.org

**Abstract.** The MALACH project [1] uses the world's largest digital archives of video oral histories collected by the Survivors of the Shoah Visual History Foundation (VHF) and attempts to access such archives by advancing the state-of-the-art in Automated Speech Recognition (ASR) and Information Retrieval (IR). This paper discusses the initial steps and the first results in building large vocabulary continuous speech recognition (LVCSR) system for transcription of Russian witnesses. Russian as the third language processed in the MALACH project (after English [2] and Czech [3]) brought new problems especially in the phonetic area. Although the most of the Russian testimonies were provided by native Russian survivors we have encountered many different accents in their speech caused by a territory where the survivors are living.

## 1 Introduction

Russian as the third language processed now in the project MALACH has addressed new problems that were not encountered in the speech of survivors who yielded their witnesses in English or in Czech. Current difficulties in automatic recognition of English testimonies seem to arise from the strong accents of the

* Support for this work was provided by NSF (U.S.A.) under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466 and by the Ministry of Education of the Czech Republic, projects No. MSM234200004 and No. LN00A063

speakers. Many of the survivors have not spoken English from the childhood and their speech is often accompanied by the mild or stronger accent. In contrast, the Czech testimonies contain relatively little non-native speech. However, we have encountered another problem, which is a high degree of inflection, a high degree of derivations and frequent using colloquial words especially in spontaneous spoken Czech. We therefore have to solve an unexpected problem of developing language models for spontaneous spoken Czech and owing to a flexible nature of the Czech language also difficulties with the high level of Out Of Vocabulary (OOV) words, which is not a problem in English.

Initial experience with processing of Russian testimonies suggests that besides the flexible nature of Russian language (Russian together with Czech and some other Central and East European languages belong to a family of Slavic languages which share this phenomenon) the dominant problems will be with the accented speech. But unlike English, where the accent is due to the native tongue of the speakers, much of the Russian accents are due to regional differences in spoken Russian. We found out that this accent is usually caused by the territory where the survivors are now living and where they yielded their interview. Studying web pages of the VHF we found out that from about 7 thousand of Russian testimonies stored in the VHF digital archives nearly one half (3500) was provided in Ukraine, about 1500 in Israel, 900 in the USA, and only one tenth (700) in Russia. The native Russians living outside Russia often adopt local non-Russian words and use them in their personal vocabulary. Perhaps this is the reason of using the non-Russian phoneme "h" in many words appearing in the Russian testimonies.

## 2 Speech collection and annotation

### 2.1 Collection conditions

Similarly as in the Czech part of the MALACH project also Russian testimonies were delivered from the VHF for further processing divided into half-hour segments stored as MPEG-1 video files. The first portion of Russian testimonies delivered and processed at University of West Bohemia (UWB) contained about eighty witnesses. The audio stream was extracted at 128kb/sec in stereo, at 16 bit resolution and 44kHz sampling rate. The speech of each interview participant - the interviewer and interviewee - was recorded via lapel microphones connected to separate channels. For annotation we chose the second part (speech contained on the second video tape) of each testimony. These segments usually do not contain any personal data of the people who yielded their testimonies and are suitable for annotation. Selected parts were burned (only the channel containing voice of the survivor) on CD ROMs and were given to annotators for processing. Annotators processed the first 15-minute segments of these parts. The initial portion of hereby prepared speech corpus consists of about 20 hours of speech.

### 2.2 Russian phonetic alphabet

For automatic recognition purposes of Russian testimonies we defined the Russian phonetic alphabet, which contains 43 phonemes. In Table 1 you can see the phoneme inventory that was specified.

**Table 1.** Russian phonetic alphabet used in the MALACH project

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a | m a m a | мама | mummy | M | k r o M e | кроме | excepting |
| b | a b y C n a j | обычной | usual | n | m a w y n a | машина | engine |
| B | n a B i r a T | набирать | collect | N | N e | не | no |
| c | b a L N i c a | больница | hospital | o | N i k t o | никто | nobody |
| C | V e C e r | вечер | evening | p | o p y t | опыт | experience |
| d | t u d a | туда | there | P | a P a T | опять | again |
| D | N e D e L a | неделя | week | r | P e r v a j | первой | first |
| e | v C e r a | вчера | yesterday | R | p R i M e r | пример | example |
| f | T e L e f o n | телефон | telephone | s | r u s K i j | русский | Russian |
| F | k o F e | кофе | coffee | S | s p a S i b a | спасибо | thanks |
| g | k a g d a | когда | when | w | s t r a w n y j | страшный | horrible |
| G | l a G e R | лагерь | camp | W | b u d u W i j | будущий | future |
| h | S i n a h o h a | синагога | synagogue | t | z a W i t a | защита | defense |
| x | v x o t | вход | entry | T | i n T e R e s | интерес | interest |
| X | T i X i j | тихий | silent | u | k a N i k u l y | каникулы | vacation |
| i | p u T i n | Путин | Putin | v | s v a b o d a | свобода | freedom |
| j | r u s K i j | русский | Russian | V | s V a S | связь | union |
| k | s a b a k a | собака | dog | z | V e Z D e | везде | everywhere |
| K | b a n K i r | банкир | banker | Z | o Z e r a | озеро | lake |
| l | o k a l a | около | around | J | t a k J e | также | also |
| L | d a L e k o | далеко | far | y | T o p l y j | тёплый | warm |
| m | d o m a | дома | at home | | | | |

The first column indicates characters that were used as the symbols of Russian phonemes. The second column shows an example of phonetic transcription of words with orthographic form given in the third column. The Russian alphabet consists of six vowels ([a], [e], [i], [o], [u], [y]) and 36 consonants. Most consonants come in pairs, "hard" (non-palatalized) and "soft" (palatalized). Among plosives we put the non-palatalized [b], [d], [g], [k], [p], and [t] and their palatalized counterparts [B], [D], [G], [K], [P], and [T]. Also fricatives ([f]/[F], [x]/[X], [s]/[S], [w]/[W], [v]/[V], and [z]/[Z]), sonants ([l]/[L], [m]/[M], [n]/[N], and [r]/[R]), and affricates ([c]/[C]) appear in non-palatalized/palatalized pairs. There are 2 non-pair consonants: fricative [J] and sonant [j].

Native Russians living in Russia don't use the phoneme "h" and usually replace it in foreign words, personal and geographical names by the phoneme "g" (in our Russian alphabet [g] or [G] - for instance in the name "Harry") or "ch" (in our alphabet [x] or [X] - for instance in the surname "Hussain"). The native Russians living in the Ukraine territory as well as in Israel or the USA learned

to pronounce "h" and they use this phoneme frequently in words in which it is actually used in local languages (for example in geographical names, personal names etc). In our Russian phonetic alphabet this extra phoneme was expressed as *&*.

### 2.3 Annotation conventions

The audio files were annotated using the special annotation software tool Transcriber 1.4.1 [5], which was adapted for processing of spoken Russian language so that transcriptions in the Cyrillic alphabet could be encoded. A macro can be assigned to each non-speech event used in the annotation process, so that we can insert a non-speech event into the text very quickly.

The rules for annotation process were the same as those used for processing of the Czech testimonies [4]. Fifteen types of non-speech events were used during annotation (Tongue click, Lip smack, Coughing, Laughter, Breath noise, Inhaling, UH, UM, UH-HUH, UH-HUM, Unintelligible, Background noise, Start and End of background noise, Silence). The first portion of 80 testimonies was processed by a group of five native Russian human annotators. They needed about 18 hours to transcribe an hour of speech using Transcriber. The difficulty lies in understanding the unfamiliar names, places, coarticulations related to age, and heavily accented speech.

## 3 Text Corpus Characteristics and Lexical Statistics

Slavic languages such as Czech or Russian are characterized by high degree of inflection, rich derivations and relatively free word order. Many inflections and derivations are rarely or never observed even in a large corpus, which leads to high OOV rates relative to other language families. Also personal and geographical names, foreign words and colloquial words appearing frequently in human transcripts pose a challenge for language modeling. The relative occurrences of these problematic word groups are given in Table 2 ($Per\_Vocab$ denotes the percentage of words from the specified class as were found in the Shoah dictionary while $Per\_Corpus$ denotes the percentage of tokens from each class as were found in the Shoah manual transcripts). To compare differences between Russian and Czech the numbers in brackets indicate characteristics obtained for the Czech Shoah corpus [3]. All statistics were computed from the manual transcripts containing more than 17k different words and 148k tokens (running words) that were at our disposal for building baseline ASR system.

Analyzing the results given in Table 2 we see that Russian Shoah transcripts contain distinctly less colloquial words than the Czech ones. This strengthens our belief in much more greater chance to use besides manual transcripts even many further sources for language modeling (compared with the Czech part of the MALACH project).

On the other hand, many running words (nearly 2.5%) uttered in Russian testimonies were incorrectly pronounced. The native Russian qualified these words

**Table 2.** Percentages of problematic word classes

| | Colloquial Words | Personal Names | Geographical Names | Foreign Words |
|---|---|---|---|---|
| *Per_Vocab* | 1.4%(8.9%) | 3.5%(5.0%) | 5.5%(4.7%) | 0.6%(4.2%) |
| *Per_Corpus* | 1.4%(6.8%) | 0.7%(0.7%) | 1.8%(1.6%) | 0.3%(0.5%) |

not as colloquial words and/or only accented speech but rather as a speech of Russians whose pronunciation of many words is partly modified by a non-Russian environment (Ukraine, Israel, etc.) where they have been living for a long time.

## 4 Experiments with the Russian Baseline ASR System

### 4.1 Front-end and acoustic modeling

The acoustic training set consists of approximately 20 hours of speech. The front-end uses 27 PLP filters distributed in the frequency axis up to 25,77 Barks (22 kHz) according to the critical-band theory. 13 PLP cepstral coefficients including their first and second derivatives were computed to yield 39 dimensional vector per frame, at a rate 100 frame per second. A cepstral mean subtraction and amplitude normalization were applied to ensure suitable feature normalization. Neither speaker adaptation nor additive noise subtraction techniques were used. The acoustic model was trained using 159 HMM states (the baseline ASR system was built as a monophone-based system) and 7.6k gaussians.

### 4.2 Language modeling

Only the transcriptions of the acoustic training data were at our disposal for the baseline ASR experiments. We estimated a standard bigram language model with Katz's discounting using these training data.

### 4.3 Recognition experiments

Recognition experiments were performed using 100 sentences randomly selected from the speakers whose testimonies were used neither for the acoustic model nor the language model training. Using the acoustic and the language model described above we have achieved recognition accuracy 33.93%. It is lower than the accuracy observed in the Czech baseline system (42.08%) but reader should bear in mind that we had more training data at our disposal (see [4]). Moreover, our team has years of experience in Czech LVCSR and therefore both the Czech phonetic transcription and individual ASR component settings have been tested for a long time. On the other hand, the MALACH project is one of our first task of transcribing spoken Russian.

## 5 Conclusion

The goal of our work was to verify whether the manual transcription process, Russian phonetic alphabet and rules for phonetic transcription of words were well designed or should be somewhat corrected. Comparing the first results with those obtained in the last year for spoken Czech we can say that our system is designed correctly. However, some inspection of the transcription rules is advisable.

To improve a performance of the LVCSR system for recognition of Russian witnesses in the project MALACH we suppose to build a triphone-based acoustic models trained from a larger portion of transcribed speech. Also a more robust language model should be designed using more text data. As the spontaneous Russian is not flooded by colloquial words such as the spontaneous Czech we believe that the LM built from topic-oriented newspaper articles or novels together with a large portion of manual transcripts could improve the performance of the recognition task.

## References


1. `http://www.clsp.jhu.edu/research/malach`
2. B. Ramabhadran, J. Huang, M. Picheny: Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH Project. Proceedings of ICASSP 2003, Hong Kong, 2003.
3. J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovský, S. Gustman: Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. Submitted to Eurospeech 2003.
4. J. Psutka, P. Ircing. J. V. Psutka, V. Radová, W. Byrne, J. Hajič, S. Gustman, B. Ramabhadran: Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments. Proceedings of TSD 2002, Brno, 2002.
5. `http://ldc.upenn.edu/mirror/Transcriber/`
6. S. Young et al.: The HTK Book. Entropic Inc., Cambridge, 1999.
7. A. Stolcke: SRILM - an Extensible Language Modeling Toolkit. Proceedings of ICSLP 2002, Denver, 2002.
8. M. Mohri, F. Pereira, M. Riley: Weighted Finite-State Transducers in Speech Recognition. Proceedings of ASR2000, International Workshop on Automatic Speech Recognition: Challenges for the Next Millennium, Paris,2000.