Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments*

Josef Psutka[1], Pavel Ircing[1], Josef V. Psutka[1], Vlasta Radová[1], William Byrne[2], Jan Hajič[3], Samuel Gustman[4], and Bhuvana Ramabhadran[5]

[1] University of West Bohemia, Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{psutka, ircing, psutka_j, radova}@kky.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
[5] IBM T.J. Watson Research Laboratory, Human Language Technologies Group
Yorktown Heights, NY
bhuvana@us.ibm.com

Abstract. In this paper we describe the initial stages of the ASR component of the MALACH (Multilingual Access to Large Spoken Archives) project. This project will attempt to provide improved access to the large multilingual spoken archives collected by the Survivors of the Shoah Visual History Foundation (VHF) by advancing the state of the art in automated speech recognition. In order to train the ASR system, it is necessary to manually transcribe a large amount of speech data, identify the appropriate vocabulary, and obtain relevant text for language modeling. We give a detailed description of the speech annotation process; show the specific properties of the spontaneous speech contained in the archives; and present a baseline speech recognition results.

1 Introduction

After filming Schindler's List, Steven Spielberg established the Survivors of the Shoah Visual History Foundation (VHF) to develop archives and teaching materials based on the videotaped testimonies given by survivors of the Holocaust

* This work has been founded 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, project No. MSM234200004 and No. LN00A063

in order to preserve their memory and establish a basis for tolerance education around the world for generations to come.

Today, the VHF has gathered almost 52,000 testimonies (116,000 hours of video) in 32 languages to form a 180 terabyte digital library of MPEG-1 video. Several years ago the VHF began the task of manual cataloging of the archives in order to facilitate content-based searching. 4,000 testimonies given in English (about 8% of the entire archive) have been manually cataloged with the segment level description [1], using a domain-specific thesaurus containing 21 thousand places and concepts. Names of people have been cataloged separately (about 280,000 different items). work in addition with multilingual materials, the automation of the cataloging process is absolutely necessary if effective access to archives of such scale is required.

The MALACH (**M**ultilingual **A**ccess to **L**arge Spoken **A**r**CH**ives) project [2] will attempt to provide improved access to this large multilingual spoken archive by advancing the state of the art in automated speech recognition. An aim of the initial phase of the project will be to develop ASR for English and Czech testimonies with subsequent extensions to French, Spanish and several Eastern European languages. This paper describes the initial work concerning the Czech part of the project.

2 Recording Conditions and Speech Collections

Testimonies were delivered for further processing divided into half-hour segments stored as MPEG-1 video files. The average duration of a testimony in the collection of the first portion of 180 Czech testimonies delivered and processed at UWB was two hours. 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 collected on 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 provided 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 these annotated testimonies consists of about 45 hours of speech.

The speech quality in individual interviews is often very poor from an ASR point of view, as it contains whispered or emotional speech with many disfluencies and non-speech events as crying, laughter etc. The speaking rate (measured as the number of words uttered per minute) varies greatly depending on the speaker, changing from 64 to 173 with the average of 113 [words/minute].

3 Speech Annotation Conventions

The audio files were divided into segments and annotated using the special annotation software Transcriber 1.4.1 which is a tool for assisting the creation of speech corpora. It makes it possible to manually segment, label and transcribe speech signals for later use in automatic speech processing. Transcriber is freely available from the Linguistic Data Consortium (LDC) web site http://www.ldc.upenn.edu/.

The rules for the annotation were as follows:

- Audio files are divided into segments; each segment corresponds roughly to a sentence.
- The beginning of a segment is marked by `<b ti>`, where the `ti` gives the time when the segment begins. The time is given in seconds.
- The instant when a speaker turn occurs is marked by `<t ti> <<spk#, n, g>>`. The `ti` is again in seconds, `spk#` is a speaker ID according to the following table:

  `spk1` ..... interviewer
  `spk2` ..... interviewee
  `spk3` ..... another person

  `n` is the name and surname of the speaker (if known), and `g` is a letter marking the gender of the speaker:

  `m` ..... male
  `f` ..... female

- The situation, when the speakers spoke over each other is marked as follows:

  `<t ti> <<spk_1, n_1, g_1 + spk_2, n_2, g_2>>`
  `SPEAKER1:` transcription of what the speaker `spk_1` said
  `SPEAKER2:` transcription of what the speaker `spk_2` said

  If the speech from one or both speakers is completely unintelligible, it is marked as `<unintelligible>`.
- Everything said is transcribed as words, no numbers are used.
- Sentences begin with a low-case letter. Only proper names and acronyms like IBM, NATO are capitalized. If a word is spelled out, the letters are capitalized and a space is put between them.
- No punctuation is used in the transcription.
- If someone stammered and said "thir thirty", the corresponding transcription is `thir- thirty`. Note that the "`-`" has to be followed by a blank space. The "`-`" is also used in the case when a word is spoken incompletely due to some recording error. In such a case the "`-`" has to be preceded or followed by a blank space, depending on whether only the end or the beginning of the word was spoken. If the "`-`" is neither preceded nor followed by any blank space it is regarded as a part of a word.

– Sometimes a speaker uttered a word or a part of a sentence in a language other than Czech. Such parts are enclosed in `[]`.
– If human transcribers are unsure about a portion of the transcription, they enclose it in parentheses. For example, if they think a speaker said "looks like this", but are unsure, they should transcribe it as `(looks like this)`. If something is completely unintelligible, the transcription is `<unintelligible>`.
– Non-speech sounds like tongue clicks, coughing, laughter, breath noise, inhaling, and lip smacks are transcribed as `<click>`, `<cough>`, `<laugh>`, `<breath>`, `<inhale>`, and `<mouth>`, respectively.
– Background noise is marked according to the following rules: if no word overlaps with the background noise the mark `<noise>` is used; if a word or a part of an utterance overlaps with the noise, the mark `<noise_begin>` is used before the first affected word and the mark `<noise_end>` is used after the last affected word.
– Other disfluencies in the speech are marked as: `<UH>`, `<UM>`, `<UH-HUH>`, or `<UH-HUM>`.
– Distinct pauses and gaps in speech are marked with `<silence>`.

The complete list of all non-speech sounds used during the annotation is given in Tab. 1. An example of the annotated file is shown in Fig. 1.

Table 1. Complete List of Non-Speech Sounds

Non-speech sound	Transcription
Tongue click	`<click>`
Lip smack	`<mouth>`
Coughing	`<cough>`
Laughter	`<laugh>`
Breath noise	`<breath>`
Inhaling	`<inhale>`
UH	`<UH>`
UM	`<UM>`
UH-HUH	`<UH-HUH>`
UH-HUM	`<UH-HUM>`
Unintelligible	`<unintelligible>`
Background noise	`<noise>`
Start of background noise	`<noise_begin>`
End of background noise	`<noise_end>`
Silence	`<silence>`

4 Text Corpus Characteristics and Lexical Statistics

This section describes some features of the text corpus created by the annotation of the speech files. Several interesting lexical statistics are also presented.

```
<t 26.800> <<spk2, f>>
<mouth><inhale> to vám neřeknu data já si absolutně nepamatuju
<t 31.747> <<spk1, f + spk2, f>>
SPEAKER1: aspoň roční období
SPEAKER2: <mouth><inhale>
<t 33.372> <<spk2, f>>
roční tož to mohlo být v třiaštyrc- dvaaštyrycet už třiaštyrycátém roce
<b 40.838>
<noise_begin> protože to byl čas vždycky ten odstup <inhale><noise_end>
<b 45.525>
<inhale> jak ty chlapy odvedly tak sme zůstali jenom s maminkama
<b 53.172>
<inhale> v tý [Modělevi] já sem <inhale> utíkala z teho <noise> lágru
```

Fig. 1. A part of an annotated file

Table 2 shows ten most frequent words from the Czech transcriptions and their relative occurrences (columns 1 and 2) after processing the 15 minute chunks of the first 180 testimonies. Relative occurrences of those words in the Czech TV&Radio Broadcast News corpus (UWB_B02) [3] and the Lidové Noviny corpus (LN), together with their position in the sorted frequency list, are in the columns 3 and 4, respectively.

Table 2. Ten most frequent words and their relative occurrences

Word	Shoah	UWB_B02		LN	
a	0.044	0.021	(2)	0.025	(1)
to	0.034	0.007	(9)	0.006	(12)
se	0.022	0.018	(3)	0.017	(3)
sem	0.020	0.000	(3021)	0.000	(1326)
že	0.019	0.010	(5)	0.008	(6)
sme	0.018	-	(-)	0.000	(18432)
tam	0.017	0.001	(156)	0.000	(174)
tak	0.017	0.003	(23)	0.002	(39)
v	0.016	0.022	(1)	0.022	(2)
na	0.013	0.017	(4)	0.015	(4)

It can be seen that while the values in the columns 3 and 4 are very similar to each other, relative occurrences in this corpus are quite different. These differences are caused by the fact that the UWB_B02 and the LN corpora contain standard Czech from broadcast news and the newspaper articles whereas, this corpus consists of a transcribed spontaneous speech and therefore contains a large number of colloquial words.

A good example of the influence of colloquial Czech on the lexical statistics is the word *sem*. While in standard Czech this word means *here*, in colloquial Czech it is also used instead of the correct form *jsem ((I) am)* which naturally occurs quite frequently. Other differences between standard and colloquial Czech are very common. Some differences can even be formalized:

- Words that begin with *o* in standard Czech are prefixed by *v* in colloquial Czech (*okno* → *vokno*)
- *ý* changes into *ej* (*modrý* → *modrej*, *výr* → *vejr*)
- *é* inside words changes to *í* (*plést* → *plíst*)
- *é* in endings changes to *ý* (*nové* → *nový*)

The rules above hold for geographical names as well. These differences will cause serious problems in language modeling and also morphological and syntactic analysis, since the text data collected so far is made up mostly of standard Czech. The available morphological analyzers, taggers and parsers were developed for the standard form of the language as well.

Personal names, geographical names and foreign words also pose a challenge for language modeling. The obvious problems that arise due to the occurrence of new proper names are further compounded by the highly inflectional nature of the Czech language. The relative occurrences of these problematic words in a standard LVCSR dictionary and in the corpus are given in Table 3.

Table 3. Percentages of Problematic Word Classes

	Colloquial words	Personal names	Geographical names	Foreign words
Per_Vocab	8.27%	3.58%	4.76%	2.71%
Per_Corpus	6.55%	0.67%	1.63%	0.49%

In the table above, Per_Vocab denotes the percentage of words from the specified class as found in the LVCSR dictionary, while Per_Corpus expresses the percentage of tokens from each class as found in the corpus. The classes are described here in detail.

The class of **personal names** contains first names and last names, including dialectical variants of the first names. This class contains roughly an equal number of first and last names, however, it is to be expected that the number of the last names will grow far more rapidly than the number of the first names as the corpus increases. Thus we expect to be able to add the list of all first names in the language model dictionary, but the recognition of the last names will likely remain an issue.

The class of **geographical names** covers the names of countries, cities, rivers and other places, as well as the names of languages and nationalities, including the derived adjectives. About 1/3 of the class are words derived from the names of countries and/or nations.

The **foreign words** class contains mostly Slovak words (58% of all foreign words) and German words (19%). The remainder of the class is constituted by Russian words and words that are probably Hebrew or Yiddish. Some survivors also switched from Czech to Slovak during the interview.

5 Baseline Automatic Speech Recognition Results

The baseline ASR system was trained in order to check the correctness of the proposed annotation procedure and to prove the feasibility of the project task, that is, the automatic transcription of the survivor testimonies. The witnesses transcribed so far were divided into data used for the acoustic model training and for ASR performance testing.

5.1 Acoustic Models

The acoustic models were trained using the HTK, the hidden Markov model toolkit [4]. The models are based on a continuous density HMMs. The speech features parameterization employed in training are the PLP coefficients, including both delta and delta-delta sub-features. Neither speaker adaptation nor noise subtraction methods were used.

5.2 Language Models

Three language models were used in our basic experiments. All of them are standard word n-gram models with Katz's discounting and they were estimated using the SRILM toolkit [5]. They differ in their vocabulary and/or the training data used to estimate them.

The first model (*Shoah_closed*) uses the vocabulary from both training and test portion of the annotated data. Thus the vocabulary is artificially closed on the test set. However, only the training part of the corpus is used for the estimation of the language model parameters. This model was applied mainly because we wanted to check the correctness of the estimated acoustic models.

The second model (*Shoah_open*) is trained on the same data as the first model, but it employs the vocabulary resulting from the training data only and therefore it represents a fair language model (it does not employ any a priori knowledge about the test data).

Finally the third model (*LN_open*) uses both the vocabulary and the training data from the Lidové Noviny (LN) corpus.

5.3 ASR Results

Recognition experiments were carried out using the AT&T decoder [6] on 90 minutes of test data (from 5 male + 5 female randomly selected speakers). Initial ASR results are summarized in Table 4.

Table 4. Baseline ASR results

Language Model	Vocabulary Size	OOV rate	Recognition Accuracy		
			Zerogram	Unigram	Bigram
Shoah_closed	24k	0%	21.64%	43.56%	49.04%
Shoah_open	23k	8.19%	18.92%	37.50%	42.08%
LN_open	60k	9.66%	13.84%	26.39%	34.00%

Please note that the *Shoah_open*/bigram performance is currently higher than that of the *LN_open* model. This is mainly due to the Shoah and LN corpora differences described in Section 4. Nevertheless, the LN corpus is a very valuable resource and will be used for the language modeling purposes in the future ASR experiments. However, some special approach will be necessary - for example, we will probably have to exploit the rules describing the standard-colloquial word changes (see Section 4).

For comparison, current Czech ASR results for the Broadcast News task are at the 65% accuracy level for the 60k vocabulary and the bigram language model and at the 70% level accuracy for the trigram model with the same vocabulary. It shows that the survivor testimonies are really difficult to transcribe.

6 Conclusion

We have described the initial Czech language ASR development efforts in the MALACH project. We have developed a well-defined annotation procedure and have transcribed a enough speech to begin ASR development. We have observed that the language as used by the survivors differs substantially from standard Czech as contained in available text corpora and thus the language modeling in the future Czech MALACH ASR system will require specialized approaches. Finally, we have presented a baseline speech recognition results showing the difficulty that we face in developing ASR for this corpus.

References

1. `http://www.clsp.jhu.edu/research/malach`
2. S. Gustman, D. Soergel, D. Oard, W. Byrne, M. Picheny, B. Ramabhadran, D. Greenberg: Supporting Access to Large Digital Oral History Archives. JCDL'02, Portland, Oregon, USA.
3. J. Psutka, V. Radová, L. Müller, J. Matoušek, P. Ircing, D. Graff: Large Broadcast News and Read Speech Corpora of Spoken Czech. Eurospeech 2001, Aalborg, Denmark, 2001.
4. S. Young et al.: The HTK Book. Entropic Inc. 1999
5. A. Stolcke: SRILM - The SRI Language Modeling Toolkit. `http://www.speech.sri.com/projects/srilm/`
6. M. Mohri, M. Riley, F. C. N. Pereira: Weighted Finite-State Transducers in Speech Recognition. International Workshop on Automatic Speech Recognition: Challenges for the Next Millennium. 2000