LARGE VOCABULARY ASR FOR SPONTANEOUS CZECH IN THE MALACH PROJECT

Josef Psutka[1], *Pavel Ircing*[1], *J.V. Psutka*[1], *Vlasta Radová*[1], *William J. Byrne*[2], *Jan Hajic*[3], *Jirí Mírovsky*[3], *and Samuel Gustman*[4]

[1] University of West Bohemia, Department of Cybernetics, Plzen, Czech Republic
{psutka,ircing,psutka_j,radova}@kky.zcu.cz
[2] Johns Hopkins University, Center for Language and Speech Processing, Baltimore, MD
byrne@jhu.edu
[3] Charles University, Center for Computational Linguistics, Praha, Czech Republic
e-mail: {hajic,mirovsky}@ufal.mff.cuni.cz
[4] Survivors of the Shoah Visual History Foundation, Los Angeles, CA 90078-3168
sam@vhf.org

ABSTRACT

This paper describes LVCSR research into the automatic transcription of spontaneous Czech speech in the MALACH (Multilingual Access to Large Spoken Archives) project. This project attempts to provide improved access to the large multilingual spoken archives collected by the Survivors of the Shoah Visual History Foundation (VHF) (www.vhf.org) by advancing the state of the art in automated speech recognition. We describe a baseline ASR system and discuss the problems in language modeling that arise from the nature of Czech as a highly inflectional language that also exhibits diglossia between its written and spontaneous forms. The difficulties of this task are compounded by heavily accented, emotional and disfluent speech along with frequent switching between languages. To overcome the limited amount of relevant language model data we use statistical techniques for selecting an appropriate training corpus from a large unstructured text collection resulting in significant reductions in word error rate.

1. INTRODUCTION

The goal of MALACH (www.clsp.jhu.edu/research/malach) is to use automatic speech recognition and information retrieval techniques to provide improved access to the large multilingual spoken archives created by the VHF. These archives contain approximately 52,000 interviews ("testimonies") in 32 languages of personal memories of survivors of the World War II Holocaust (116,000 hours of video). In addition to preserving this material for posterity, VHF's mission is to develop instructional resources for tolerance education. 4,000 English language testimonies (about 8% of the entire archive) have been manually cataloged at VHF to date. Cataloging is done by experts who manually assign segment-level descriptors to the video segments from a thesaurus of 21000 places and concepts created by VHF for this purpose. In addition to the assignment of these thesaurus terms, the catalogers also maintain a list of names of people mentioned in the testimonies (280,000 unique items thus far). The catalogers are able to work in near real-time. However, given the size and diversity of the archives, cataloging is still a daunting challenge. It is our goal to develop automatic speech recognition and retrieval techniques to improve cataloging efficiency and eventually to provide direct access into the archive itself.

All aspects of ASR are challenging within this corpus. The speakers are usually elderly, their speech is often heavily accented, and due to the nature of the stories they relate it is often very emotional. Previous work has focused on the basic approaches in acoustic modeling and adaptation [1]. In this paper, we address difficulties in ASR language modeling for this unusual domain. The stories recounted by the subjects are inherently personal and refer to events long past. As will be discussed, the bulk of text resources available in electronic form is likely to be from inappropriate domains (e.g. news) and in unsuitable style (formal vs. colloquial). This is in addition to the usual language modeling (LM) problems of Slavic languages such as Czech.

This paper describes baseline Czech ASR systems developed within MALACH. ASR performance suffers greatly due to a lack of appropriate LM training data. Techniques that automatically select in-domain LM training data from large, general text collections are used to reduce out-of-vocabulary (OOV) rates and to improve ASR error rates.

2. CHARACTERISTICS OF SPOKEN CZECH IN THE MALACH CORPUS

The testimonies of nearly six hundred survivors have been recorded in Czech, but only about 350 of these have been digitized to date. The interviewer and the interviewee have been recorded via lapel microphones on separate channels. Testimonies are stored in the VHF digital library, divided into 30-minute segments and stored as MPEG-1 video files. Audio was extracted at 128 kb/sec in 16-bit stereo and 44KHz sampling rate.

The average length of a Czech testimony is 1.9 hours. The testimonies were randomly divided into training and test sets. The training set was created by transcribing 15-minute segments, 30 minutes into each testimony (i.e. at the beginning of the second segment), getting past the biographical questions, initial awkwardness and into the

middle of the subjects' stories. The acoustic and language modeling test set consists of completely transcribed testimonies from 5 males and 5 females (Table 1).

The audio files were divided (roughly) into sentences. Transcription was done using the Transcriber 1.4.1 speech editing tool (http://www.ldc.upenn.edu), which was modified to incorporate Unicode. In addition to lexical transcription, the following non-speech sounds were marked: Tongue click, lip smack, cough, laughter, breath, inhalation, UH, UM, background noise, silence, and unintelligible. The rules for the entire annotation process have been published previously [2]. The annotators worked at a rate of fifteen times real time. Transcription inspection and verification requires additional effort at approximately twice real time.

2.1 Speech and Speaker Characteristics

The speech quality is often quite poor to use in building ASR systems. There is frequent whispered or emotional speech along with many disfluencies and non-speech events as crying, laughter, etc. Transcribers observed that the quality and fluency of speech was often affected by the age of speakers. The age of the oldest survivor was 94; the average age of all speakers was 75 years. The speaking rate was also quite variable, ranging from 64 to 173 words per minute, with an average rate of 113.

	Training (336)		Test (10)	
	Male	Female	Male	Female
Speakers	145	191	5	5
Hours transcribed	36.25	47.75	13.15	9.7

Table 1. Czech Speakers and Transcriptions

Many of the survivors were originally from territories[1] where Czech was not frequently used (Table 2). Their speech was heavily accented, and their word usage was also influenced by their place of origin.

	Bohemia	Slovakia	Carpathians	Other
Place of childhood	73.4%	13.0%	5.2%	8.4%

Table 2. Territories Where Survivors Spent Their Childhood

2.2 Colloquial Usage

Spontaneously spoken Czech contains words and usages not found either in standard written or in formal spoken Czech. The MALACH corpus, in particular, is rich in these spontaneous forms. Examples are given in Table 3, showing the differences between the colloquial usage that appears in the MALACH transcriptions and the formal versions that would appear if the same sentiment were to appear in news text or broadcast transcriptions. Many of these forms can be analyzed morphologically, however usage is variable. As a result, constructing a transducer to map formal text to its spontaneous form is problematic, because users are not consistent in their choice of spontaneous form. As a result, the best source of information on spontaneous usage is transcribed speech itself.

MALACH	Voni mi vopatøili èesky´ pas.
Formal	Oni mi opatøili èesky´ pas.
English	They provided a Czech passport for me.
MALACH	… bejvalej èesky´ rotmistr
Formal	… by´valy´ èesky´ romistr
English	… a former Czech sergeant

Table 3. Example Colloquial and Formal Variations

2.3 Lexical Statistics

Slavic languages such as Czech are characterized by a high degree of inflection, rich derivations (prefixes and suffixes) and relatively free word order. Many inflections and derivations are rarely or never observed even in a large corpus, which leads to large OOV rates relative to other language families. Each OOV word causes necessarily an ASR error, and it is quite possible that these OOV words are the very terms that will be needed by subsequent processing steps. Free word order negatively affects n-gram reliability. Coupled with the usual stylistic differences between spontaneous speech and written text, the rich morphology and the presence of colloquial language make the need for relevant language model data even more of an issue.

The acoustic training set contained 43,702 different words and 565,517 tokens (running words). Pronunciations were generated to create a lexicon for the corpus. The distribution of words is considerably different in this corpus than in broadcast news or newspaper articles [2,3]. The transcriptions of the testimonies contain a large number of colloquial words, personal and geographical names, and foreign words (Table 4) that are underrepresented in a typical collection of Czech text.

Colloquial Words	Personal Names	Place Names	Foreign Words
8.9% / 6.8%	5.0% / 0.7%	4.7% / 1.6%	4.2% / 0.5%

Table 4. Problematic Word Classes. Frequencies of word classes by words (vocabulary types) / tokens.

Personal names (5.0% of the vocabulary) contain first names and last names, including dialectal variants of 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 first names as the size of the corpus increases. **Geographical (place) names** (4.7% of words) cover the names of countries, cities, rivers and other places, as well as names of languages and nationalities. **Foreign words** class (4.2% of words) contains mostly Slovak and German words (also English, Russian, Hebrew, Yiddish). Some of the foreign words appeared in isolation, but there were also stretches of continuous segments e.g. in Slovak.

The test set contained 19,465 different words and 156,315 tokens. The coverage of the test set vocabulary by subsets of the training set transcriptions is presented in Table 5.

[1] Slovakia was part of the former Czechoslovakia from 1918 to 1992; Carpatho-Ukraine from 1918 to 1938. People living there were frequently using the Slovak and Ukrainian languages, respectively.

# of testimonies	*# of types*	*# of tokens*	**OOV rate** *types (tokens)*
50	12.9k	80.9k	19.7k (12.6 %)
100	20.3k	165.0k	15.1k (9.6 %)
150	26.1k	239.3k	13.0k (8.3 %)
200	31.5k	326.1k	11.5k (7.4 %)
250	36.0k	405.2k	10.5k (6.7 %)
300	40.9k	507.4k	9.4k (6.0 %)
336	43.7k	565.5k	9.0k (5.8 %)

Table 5. Test set coverage with increasing training set size

The OOV rate of 5.8% with a 43,700-word lexicon is high compared to other languages; in many English language tasks, an OOV rate of less than 2% can be obtained with a vocabulary of 50,000 words. The simplest way to decrease the OOV rate would be to increase the lexicon through words found by transcribing more data. However, a rough extrapolation of the data in Table 5. indicates that we would need several thousands of transcribed testimonies to reduce the OOV rate to below 1%; that much speech is not available.

3. LANGUAGE MODEL CORPUS CREATION

Given that it is impractical to create enough language model training data by transcribing speech, we investigated the use of other text collections to complement the transcriptions. To our knowledge, there are no large collections of transcriptions of spontaneous, spoken Czech. We therefore investigated the use of general text collections.

We first attempted to use the Lidove Noviny (LN) corpus, which is a large collection of general news text. We found however that the most frequent 60,700 words in the corpus yielded only a 9.6% OOV rate on the test set testimonies. We conclude from these results that language models built from general news text collections are not suitable for our task.

We then considered another larger and more diverse text collection. The Czech National Corpus (CNC) contains approximately 400 million words (tokens) taken mostly from newspaper articles, but also from other sources, such as novels. Of course, it was not possible to determine beforehand whether or not this collection contains enough in-domain language model data to construct a robust language model for our domain; the collection is too large to investigate it manually in this way. However, given the diversity of the CNC collections, we suspected that it would contain passages, perhaps from novels or other dialog-type texts that would be useful in our domain. We therefore investigated the possibility of using automatic methods to select sentences from the CNC that are similar in language usage, lexicon, and style to the sentences in the training set transcriptions.

3.1 Automatic Selection of Language Model Text

A statistical test was established to determine whether sentences from the CNC were similar to the transcribed testimonies. We employ techniques developed for the construction of topic and domain specific language models [5]. While our goal is the same originally motivated the development of these techniques, namely to filter a large collection of text to find relevant data based on examples, the problem faced here is complex in that there are simultaneously stylistic and domain specifics to be considered. For example, finding scholarly text related to the Holocaust might not help in language modeling for this task because of the style mismatch between formal and spontaneous forms. Ideally, we would like to tune the search for either style or content. We do not address this problem here, but note that the ability to filter out in-domain material of inappropriate style would be valuable.

Two unigram language models were created: P(.|CNC) was estimated from the CNC collection, and P(.|Tr) was estimated from the acoustic training set transcriptions. A likelihood ratio test was applied to each sentence in the CNC, using a threshold **t**: a sentence **s** from the CNC was added to the filtered set (named CNCS) if P(**s**|CNC)< **t** P(**s**|Tr) . This is a simple way of assessing whether sentences from the CNC are closer to the testimony transcriptions than to the bulk of the CNC corpus itself. The test threshold effectively allowed us to determine the size of selected sub-corpus CNCS. Gradually decreasing the threshold yields smaller and smaller sub-corpora that, ideally, are more and more similar to the testimony transcriptions. A threshold of 0.8 created a CNCS containing about 3% of the CNC (approx. 16M tokens).

The test set coverage obtained by of all these collections is summarized in Table 6. The filtered version of the Czech National Corpus (CNCS) has greatly improved the OOV rate relative to the similarly sized collection of LN news text.

Corpus	**# of types**	**# of tokens**	**OOV (token)**
Tr	43.7k	565k	9.0k/5.8%
LN	60.7k	33.2M	15.0k/9.6%
CNCS	61.0k	15.8M	9.7k/6.2%
Tr + CNCS	81.9k	565k + 15.8M	6.0k/3.8%

Table 6. Test Set Coverage by Language Model Corpora

The CNCS corpus does not have as good coverage as the testimony transcriptions (Tr) themselves, but merging them together (denoted Tr+CNCS) yields an OOV rate lower than that obtained with either of the individual corpora (due to their substantially different vocabularies); we will show below that this fact can be successfully exploited.

4. SPEECH RECOGNITION

The acoustic training set consisted of approximately 84 hours of speech. The data was parameterized as 15 dimensional PLP cepstral features. Features were computed at a rate of 100 frames per second. Cepstral mean subtraction was applied per utterance. Cross-word acoustic models were trained using the HTK Toolkit [5]. The resulting models had approximately 6000 states and 96k Gaussians.

Language models were trained using the SRILM Toolkit [6]. All language models were word bigrams estimated using Katz's discounting. When building language models for speech recognition experiments we attempted to exploit all available text resources in order to achieve as good recognition results as possible.

Two different language models were estimated using the acoustic training set transcriptions. The model **LM-Tr** is trained on the acoustic training set transcriptions using only the training set vocabulary. Another model **LM-Tr-C** has a vocabulary that was augmented to cover the test set

transcriptions; this is of course a "cheating" language model. It was created to assess the influence of OOV words on the recognition Word Error Rate (WER). The model **LM-LN** was trained using the Lidove Noviny corpus, and finally, the model **LM-CNCS** was trained using the CNCS (the filtered CNC corpus).

4.1 Speech Recognition Performance

The test set consisted of 500 sentences selected at random from the testimonies held aside as a test collection; 50 sentences were selected at random from each of the 10 test speakers. Results with each of the language models are summarized in Table 7. The effect of OOVs is readily apparent; the difference in performance between **LM-Tr** and **LM-Tr-C** is almost exactly the OOV rate with respect to the acoustic training set transcriptions. The worst performance is obtained under the Lidove Noviny language model (**LM-LN**). We note that the degradation relative to **LM-Tr** is not explained by the **LM-LN** OOV rate alone; the LN corpus itself is clearly mismatched to this task. The language model trained on the filtered CNC (**LM-CNCS**) achieves performance that is worse than the one obtained with the **LM-Tr** model but significantly better than with the **LM-LN.**

LM	# of types	OOV	WER
LM-Tr-C	43.9k	0.0 %	40.23 %
LM-Tr	43.7k	5.8 %	45.91 %
LM-LN	60.7k	8.8 %	59.75 %
LM-CNCS	61.0k	6.2 %	52.99 %

Table 7. ASR performance and language model domain

Although the **LM-CNCS** language model did not perform as well as the models trained on the acoustic training set, we took advantage of their different vocabularies and created a merged model **LM-Tr-CNCS**: we merged the **CNCS** and **Tr** vocabularies, retrained bigram language models in each domain (obtaining **LM-Tr'** and **LM-CNCS'**), and used the SRILM Toolkit to interpolate them linearly using the following formula:

$$P_{LM\_Tr\_CNC\_S}(.) = \lambda P_{LM\_Tr'}(.) + (1-\lambda) P_{LM\_CNC\_S'}(.)$$

Word error rates are reported in Table 8 for some values of the parameter λ. For λ = 0.0, which corresponds to training on the CNCS corpus alone, we find an improvement relative to the 52.99% WER result reported above due to the expanded vocabulary. Similar gains are observed at the value 1.0 which corresponds to training the language model on the acoustic training set alone. Clearly some benefit is obtained in each case through a merging of the vocabularies between domains. Further gains can be found at intermediate values of the interpolation constant with a peak at about λ = 0.75.

These results validate the statistical filtering used to select the CNCS collection. We find improvement both from an enlarged vocabulary and from increased predictive power obtained by merging the filtered language model with an entirely in-domain language model.

λ	WER
0.00	51.10 %
0.25	46.68 %
0.50	44.31 %
0.75	**43.92 %**
1.00	44.99 %

Table 8. Results of experiments for the merged language model **LM-Tr-CNCS**, merged vocabulary (81.9k words)

5. CONCLUSIONS

We have presented automatic speech recognition results from our initial investigations into the challenging problem of transcribing oral histories. In this domain the initial modeling difficulty encountered is the limited availability of acoustic and language model training data. We have focused on the latter problem here, with the goal of automatically searching very large general corpora for in-domain language model training data. We employ a statistical search for similar data guided by a statistical model trained on whatever in-domain transcriptions are available. We find that performance gains result from both increased vocabulary coverage as well as from improved predictive power. An overall reduction in Word Error Rate from 45.91% to 43.92% is achieved. Work is ongoing to develop language models that explicitly model language changes when users switch from formal to spontaneous Czech. Such detailed models should complement topic specific language models tuned to the domain of the VHF collections.

6. ACKNOWLEDGEMENTS

This work has been funded by the 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 MSM234200004 and LN00A063.

7. REFERENCES

[1] B. Ramabhadran, M. Picheny and J. Huang, "Automatic Transcription of Speaking Styles in Spoken Archives - English ASR for the MALACH project". ICASSP'03.

[2] J. Psutka, P. Ircing, J.V. Psutka, V. Radová, W.J. Byrne, J. Haji_, S. Gustman, and B. Ramabhadran, "Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments", TSD'2002, LNAI 2448, Springer-Verlag Berlin Heidelberg 2002, pp. 253-260.

[3] W. Byrne, F. Jelinek, P. Ircing, P. Krbec, J. Haji_, J. Psutka, S. Khudanpur, "On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language – Czech", In: 7th EUROSPEECH'2001, Denmark, Aalborg, 2001, pp. 487-490.

[4] P. Ircing, J. Psutka, "Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary", TSD'2001, LNAI 2166, Springer-Verlag Berlin Heidelberg 2001, pp. 273-277.

[5] R. Iyer, M. Ostendorf and H. Gish, "Using Out-of-Domain Data to Improve In-Domain Language Models", IEEE Signal Processing Letters, 4(8) 221-223, 1997

[6] The HTK Toolkit. User's Manual. Entropics plc, 1995.

[7] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit", In: ICSLP'2002, Denver 2002, pp. 901-904.