Towards Automatic Transcription of Spontaneous Czech Speech in the MALACH Project*

Josef Psutka[1], Pavel Ircing[1], Josef V. Psutka[1], Vlasta Radová[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, 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

Abstract. Our paper discusses the progress achieved during one-year effort with building the Czech LVCSR system for an automatic transcription of spontaneously pronounced testimonies in the MALACH project [1]. The difficulty of this task stems from the highly inflectional nature of the Czech language and is further multiplied by the presence of many colloquial words in spontaneous Czech speech and also by the need to handle emotional speech filled with disfluencies, heavy accents, age-related coarticulation and language switching. In this paper we concetrate mainly on the acoustic issues - the proper choice of the front-end parameterization, handling the non-speech events in acoustic modeling and especially the unsupervised usage of the MLLR adaptation technique. A method for selecting suitable language model data is also briefly mentioned.

1 Introduction

The goal of the MALACH project [1] is to use automatic speech recognition (ASR) and information retrieval (IR) techniques to provide improved access to the large multilingual spoken archives created by the VHF. The initial stages of building the ASR component were reported in a paper published in the TSD proceedings last year [2].

* 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

The paper described the recording conditions under which the data were obtained along with means of data digitization and further processing. A detailed description of the speech annotation procedure was also given. The analysis of the text corpus obtained by the annotation of the speech files revealed that the MALACH data differ significantly from other available Czech text corpora. This dissimilarity is caused by the fact that existing text resources are from inappropriate domains (broadcast news, newspaper articles) and also by the substantial difference between formal and colloquial Czech on both the lexical and the syntactic level.

This fact, together with the well-known highly inflectional nature of the Czech language and the presence of proper names and foreign words, made the initial ASR performance reported in [2] suffer from a lack of appropriate language model training data.

In our current paper we report the results of a one-year effort put into improving the ASR compoment of the MALACH project.

2 Properties of Collected Corpus

First of all, we have finished the process of annotation of the Czech testimonies. The final training set consists of 336 speakers (145 males and 191 females) yielding a total of approximately 84 hours of speech data. The training set contains 43,702 different words and 565,517 tokens (running words).

The percentages of problematic word classes (colloquial words, personal and geographical names, foreign words) remain essentially the same as those found in the initial portion of the testimonies [2]. Detailed numbers concerning the whole training set are given in [3]. We would only like to stress out that colloquial words constitute 8.9% of words in the vocabulary and 6.8% of tokens in the corpus. According to our knowledge, such high percentages of colloquial words do not appear in any other language besides Czech.

The test set consists of completely transcribed testimonies from 5 male and 5 female speakers (approximately 23 hours of transcribed speech) which were not included in the training data. It contains 19,465 different words and 156,315 tokens. The test set OOV rate with a 43.7k training set vocabulary is 5.8%. It is a high number compared to other languages; in many English language tasks, an OOV rate of less than 2% can be obtained with a 50k vocabulary.

3 Building Robust System for Automatic Transcription

3.1 Front End

Several tests were performed in order to determine the best parameterization of the acoustic data. We have experimented with MFCC and PLP parameterization and also looking for an optimal number of coefficients (see [4] for methodology).

The best results were achieved using 15 PLP cepstral coefficients with both delta and delta-delta sub-features. Therefore one feature vector contains 45 coefficients. Feature vectors were computed at a rate of 100 frames per second. Cepstral mean subtraction was applied to all features on a per utterance basis.

3.2 Acoustic Models

The basic speech unit used in our acoustic module is a triphone. Each triphone is represented by a three-state HMM with a continuous output probability distribution assigned to each state. This probability distribution is represented by 16 Gaussians mixtures with diagonal covariance matrices.

Initially, we trained a separate model also for each of the non-speech events (see the list in [2]). Then we came across a problem how to incorporate those models into the decoder.

One possibility is to add each non-speech event model at the end of the phonetic baseform in each pronunciation lexicon entry. However, such approach would multiply the lexicon size by 11 (the number of distinct non-speech events in our system) and is therefore highly unpractical.

The second approach involves treating non-speech events as independent words. Then we face a problem with language modeling, since non-speech events appear in the utterances regardless of a lexical context and therefore their probabilities cannot be captured by the standard language modeling techniques.

Thus a third approach was devised. We took the sets of Gaussian mixtures from all the non-speech event models including the standard model for a long pause (silence - `sil` - see [6]). Then we weighted those sets according to the state occupation statistics of the corresponding models and compounded the weighted sets together in order to create a robust "silence" model with 176 Gaussian mixtures. The resulting model was incorporated into the pronunciation lexicon so that each phonetic baseform in the lexicon is allowed to have either the short pause model (`sp`) or the new robust `sil` model at the end.

The described technique proved to be very efficient in handling the non-speech events that often appear in the survivor's testimonies.

3.3 Language Models

The high OOV rate and the incompatibility between the testimonies and available text resources raises a need for obtaining more of appropriate language model training data. The seemingly simplest way to achieve this is to transcribe more MALACH data. However, the analysis presented in [3] indicates that we would need several thousands of transcribed testimonies to reduce the OOV rate under 1%. Such effort would be in conflict with the project aim (automatic transcription of the testimonies) and moreover so many testimonies do not even exist.

We have devised a statistical technique for selecting suitable training data from a large unstructured text collection - the Czech National Corpus (CNC).

Experiments described in [3] showed that merging of the vocabularies from the testimony transcriptions and from the newly selected data substantially decreased the OOV rate and the interpolation of the corresponding models improved the ASR performance (see Table 1).

3.4 Speaker Adaptation

While both acoustic and language modeling of the VHF data is extremely difficult, the project has a considerable advantage over other large vocabulary ASR tasks in the following two aspects:

1. There is a lot of speech available from each speaker
2. There is no need for a real-time recognition.

The first feature presents an opportunity for speaker dependent, long-term acoustic model adaptation, whereas the second one allows a usage of multi-pass recognition strategies.

Since the transcription of the testimonies is supposed to be automatic, we would need an unsupervised speaker adaptation technique. Currently we do not have such a tool at our disposal; however, employing a two-pass recognition in combination with the supervised MLLR adaptation we can accomplish the desired effect - an unsupervised adaptation.

The procedure is as follows. After the first recognition pass, the output of the decoder is used as the reference transcription for computing an MLLR transformation. Then the acoustic models are adapted using this transformation and a second recognition pass is performed. Therefore all data available from a given speaker are used for adaptation without any human effort put into the manual annotation of the adaptation data.

Even though the first-run recognition accuracy is still far from perfect (about 55%), a substantial performance improvement was achieved using the technique described above (see Table 1).

4 ASR results

The ASR test data consist of 500 sentences selected at random from the testimonies held aside as a test collection (see Section 2); 50 sentences were selected at random from each speaker.

Feature extraction was performed with our own software [4]. Acoustic modeling and computation of MLLR transformations were carried out using HTK Toolkit [6] and language models were trained using the SRILM Toolkit [7]. All language models used in the experiments are word bigrams estimated using Katz's discounting. Finally, the AT&T decoder [8] was used for the actual recognition.

Results of the experiments described in the paper are summarized in Table 1. The first line (#1) shows the best performance reported in our paper published in the TSD 2001 proceedings. It is included in the table just to illustrate the progress of our research. Note that the system #1 is trained using approximately half of the final training data.

Table 1. ASR result summary

Experiment	Vocabulary Size	OOV rate [%]	Accuracy [%]
#1	23k	8.19	42.08
#2	43.7k	5.8	54.09
#3	81.9k	3.8	56.62
#4	81.9k	3.8	60.60
#5	81.9k	3.8	62.14

After transcribing the entire training set and tuning front-end and acoustic model parameters (see Section 3.1 and Section 3.2, respectively), we have observed a great improvement of the system performance (Experiment #2).

Both systems #1 and #2 used only the testimony transcriptions as the language model training data. Interpolating the transcription language model from Experiment #2 with the model estimated using the data statistically selected from CNC (see Section 3.3 and [3]) further improved the accuracy (Experiment #3).

Table 2. Speaker adaptation results for individual speakers

Speaker	Accuracy [%]		
	#3	#4	#5
1	52.10	59.12	60.68
2	47.10	50.84	52.51
3	59.81	60.75	61.87
4	60.99	68.96	70.88
5	59.34	59.34	61.08
6	63.33	68.09	68.49
7	52.62	54.63	56.79
8	62.64	64.84	65.57
9	55.48	57.89	59.65
10	39.23	53.38	58.20
Total accuracy [%]	56.62	60.60	62.14

In experiment #4 we used the speaker adaptation technique described in Section 3.4. As can be seen from Table 1, we have achieved a significant gain in terms of the recognition accuracy.

Experiment #5 was carried out in order to find out how much we can boost the accuracy using the correct (that is, manually annotated) transcriptions as the reference data for the MLLR adaptation. Table 1 shows that the additional improvement over Experiment #4 is not so big.

Table 2 presents the results of the speaker adaptation experiments for individual speakers. It can be seen that the accuracy improvement gained by the adaptation varies greatly depending on the speaker, ranging from 0.00% for Speaker 5 to 14.15% for Speaker 10. The table also shows that the benefit of the supervised adaptation is rather consistently around 1.5% with the exception of Speaker 10.

5 Conclusions and future work

Our paper presented the importance of the careful design of every component used in the system for automatic speech recognition, especially when dealing with spontaneous speech with a lot of disfluencies and non-speech events.

We have shown that the system can greatly benefit from the MLLR adaptation even though the correct reference transcriptions are not available and the output of the decoder has to be used.

We have also briefly outlined the technique used for finding appropriate language model training data. A detailed description of the method can be found in [3].

In our further research we would like to concentrate on the language modeling, especially on proper handling of colloquial words.

References

1. http://www.clsp.jhu.edu/research/malach
2. 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.
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, L. Müller, J. V. Psutka: Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. Proceedings of Eurospeech 2001, Aalborg, 2001.
5. J. Psutka, V. Radová, L. Müller, J. Matoušek, P. Ircing, D. Graff: Large Broadcast News and Read Speech Corpora of Spoken Czech. Proceedings of Eurospeech 2001, Aalborg, 2001.
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.