

Large Vocabulary Speech Recognition for Read and Broadcast Czech^{*}

W. Byrne¹, J. Hajic², P. Ircing³, F. Jelinek¹, S. Khudanpur¹, J. McDonough¹,
N. Peterek², J. Pstuka³

¹ Johns Hopkins University, Baltimore MD, USA

² Charles University, Prague, Czech Republic

³ University of West Bohemia in Pilsen, Czech Republic

Abstract. We describe read speech and broadcast news corpora collected as part of a multi-year international collaboration for the development of large vocabulary speech recognition systems in the Czech language. Initial investigations into language modeling for Czech automatic speech recognition are described and preliminary recognition results on the read speech corpus are presented.

1 Introduction

In this paper we describe data collection efforts to develop an infrastructure for the development of Czech language automatic speech recognition (ASR) systems. Our goal is to gather a rich variety of spoken language material so that Czech language speech processing systems can be developed for research and eventually for other purposes. Our initial collection is a Czech read speech corpus intended primarily for initial and baseline experiments. In addition, Czech language Voice of America (VOA) broadcasts are also being recorded and transcribed. These databases are being collected in the early stages of a three year research program. As such, only preliminary investigations intended to assess the effectiveness of known techniques will be reported here. Our ultimate research plans are more ambitious, however. Unlike English and other languages for which high performance ASR is available, Czech is a highly inflected, relatively free word order language. It has been observed elsewhere that this leads to significant rates of out-of-vocabulary words (OOVs) even with very large language models [1]. It is our goal to address this problem directly by incorporating pronunciation modeling, morphological modeling, and parsing early in the recognition process [2–5]. The corpora and preliminary experiments described here are initial steps towards these goals.

^{*} This work was supported by the projects KONTAKT No. ME293 and No. VS96151 of the Ministry of Education in Czech Republic and by NSF Grants No. IIS-9810517 and No. IIS-9820687.

2 Czech Language Speech Corpora

2.1 CUCFN Read Speech Corpus

Our initial effort in data collection under this project is the Charles University Financial News Corpus (CUCFN). The first stage of this corpus consists of recordings of read economic news taken from the Ceskomoravsky Profit Journal. Speech read by fluent Czech speakers was recorded in quiet conditions at 22KHz with 16 bit resolution. Recording were made simultaneously with both a HMD410 Sennheiser head-mounted, close-talking microphone and a desk-mounted K6 Sennheiser microphone. Speech from 29 male speakers and 23 female speakers has been collected and verified. Most subjects were native speakers of common Czech, except for some speakers with marked regional accents from North Moravia and South Moravia. There was also one native Russian speaker and one native Macedonian speaker. Each speaker read a randomly selected section of 100 news sentences as well as 40 common 'enrollment' sentences for use in adaptation and normalization experiments. The first stage of corpus contains a total of 7280 sentences yielding slightly more than 17 hours of speech. On average each utterance contains 17.0 words and has a duration of 9.1 seconds. This read speech corpus is meant to serve as a well-defined environment for use in developing new Czech speech recognition systems.

A second stage of data collection is currently ongoing. General news stories from the Lidove Noviny newspaper [6] read by 50 additional speakers are being recorded.

2.2 CZBN Broadcast News Corpus

We are also collecting a Czech language Broadcast News corpus. Satellite transmissions of Voice of America broadcasts are being recorded by the Linguistic Data Consortium (LDC) and transcribed at the University of West Bohemia according to protocols developed by LDC for use in Broadcast News LVR evaluation. The recordings span the period February 8 through May 4, 1999. The corpus consists of 46 recordings of 30 minute news broadcasts yielding a total of 23 hours of broadcast material. Portions of the shows containing music, speech in music, or other non-speech material are marked, but speech found during those periods is not transcribed. Based on an initial assessment of the transcribed material, this yields approximately 19:30 minutes of transcribed material from each 30 minute broadcast. The average sentence has a duration of 8.6 seconds and contains 18.3 words.

Collection and transcription has begun only recently, hence no ASR systems yet exist. This corpus will be a focus of one of the projects to be undertaken at the 1999 NSF Language Engineering Workshop held at Johns Hopkins University.

3 Czech Language Pronunciation Dictionary

We use the phonetic mappings presented by Psutka [4, 5] to derive pronunciations for large vocabulary speech recognition experiments. Our phone set consists

of 38 phonemes. Several phones in the original phone set [4] were discarded because the words in which they occur were infrequently observed in the acoustic training data (stage 1 of the CUCFN data). As described [4], pronunciations

Table 1. Modifications to Czech phonetic alphabet.

Description or example	Number of occurrences	Replacement
voiced counterpart of “c”	13	“c”
voiced counterpart of “ch”	10	“ch”
pneumatika	9	“e u”
tramvaj	3	“m”
long “o”	49	“o”

are derived from lexical forms directly from spelling with supplemental rules of transformation. The rules have the form $A \rightarrow B/C_ (D|E)$, which indicates that letter A is transformed to letter B when preceded by C and followed by either D or E . This yields an intermediate representation which is then mapped directly to the phone set. These rules capture coarticulation, context effects, and changes in voicing due to context.

A difficulty in the use of these rules is their treatment of words of non-Czech origin. For example the rule $t \rightarrow t' / _ (i|i')$ transforms *ticho* to *t'icho*, and the rule $n \rightarrow \check{n} / _ (i|i')$ transforms *nic* to *ňic*. However, both these rules when applied to the word *administrativa* yield *admiňistrat'iva*, which should be pronounced as *administrativa*. As these foreign words are observed they are added to a list of exceptions. However pronunciation of non-native words remains a problem.

4 Czech Language Modeling

Statistical language modeling for Czech poses several interesting challenges not encountered in some of the more widely studied languages such as English or French. The usual N-gram approach to language modeling runs into two main hurdles when used for Czech. The highly inflected nature of Czech causes a proliferation of word types if words are viewed as white-space-separated groups of characters without attaching any further meaning to them. At the same time Czech also has a nearly-free word order, further compounding the gathering of good statistics of word collocation.

The highly inflected nature of Czech is clearly evident in the CUCFN corpus described in Section 2.1 and used for experiments reported here. The 16.5 million word language model (LM) training portion of the corpus contains 415K distinct words. The 64K most frequent words, a typical vocabulary size for English automatic speech recognition systems, cover only 95% of the training tokens and the coverage by the 100K most frequent words increases to only 96.4%. Of the 415K distinct words 241K appear only once or twice in the corpus. Thus a fairly high

incidence of out of vocabulary (OOV) words can be expected on any realistic recognition task if each word form is treated as a distinct word in the language model.

A portion of about 19K words was held out from the CUCFN collection and used as a test set in experiments reported here. It contains about 7K distinct words, and even for a 415K word vocabulary extracted from the LM training corpus, we find that 1.6% of the tokens in the test set are not seen in training! For more realistic recognition vocabulary sizes, this OOV rate is even higher: 30% for a 5K vocabulary, 22% for 10K, 18% for 15K, ..., 6.7% for 63K. This should be compared with typical OOV rates of 1-2% for a 64K-word vocabulary encountered on the English *Wall Street Journal* (WSJ) corpus, and 3-4% for a 64K-word vocabulary encountered on the Spanish *Broadcast News* corpus. While a detailed analysis of the incidence of OOVs due to new inflected forms as opposed to, say, new proper nouns *etc.* remains to be performed, Czech does appear to have a higher rate of inflected forms than languages like Spanish.

It is thus clear that a brute-force expansion of the vocabulary is not the ideal approach for Czech language modeling. We are investigating several alternative solutions to this problem, including class based language models using automatically derived morphological word classes, in our ongoing work. In this paper, however, we only report results with word N-gram language models using either a small vocabulary of 5K words for pilot experiments or a large vocabulary of 63K words for more realistic evaluations.

The small (5K word vocabulary) speech recognition system was built primarily as a development step in order to validate the proper engineering of component modules such as the acoustic models, the dictionary and the language model. To circumvent the 30% OOV rate of the 5K-word vocabulary for the pilot experiments, which could make it difficult to pinpoint faulty modules in case of high recognition error rates, it was decided to close the LM vocabulary on the test set. All (4500) words in the test set reference transcriptions which did not appear in the list of 5K most frequent words in the training corpus were added to the 5K word vocabulary, resulting in a 9.5K-word vocabulary.

A bigram and a trigram language model were trained for the pilot experiments with the 9.5K word vocabulary and the 16.5 million word corpus. A bigram and a trigram language model were also estimated for the large vocabulary experiments with the 63K word vocabulary from the same corpus.

Table 2 shows the perplexity of the test set under various language models. Note that the perplexity of the English WSJ corpus, which is comparable in terms of genre (newspaper text) and content (financial news), is in the range of 180-220 for a bigram model and 110-160 for a trigram model, depending on vocabulary size, amount of training text, *etc.*

The relatively high perplexity of the models for the 9.5K vocabulary perhaps merits some explanation: when we close the 5K vocabulary by including 4500 additional words which occur in the test set, we end up estimating a model which “knows” that these words are infrequent even though they have made an appearance in a word list otherwise comprising the most frequent words.

N-gram Order	9.5K Vocab (0% OOVs)	63K Vocab (6.7% OOVs)
1gram	5895	2579
2gram	1829	737.5
3gram	1664	657.4

Table 2. Test set perplexity for various vocabulary sizes and N-gram order

As a result, they often get an even smaller probability than what a *uniform distribution* would assign to a set of 9.5K words.

5 Acoustic Modeling

Initial large vocabulary speech recognition experiments have been conducted on the CUCFN corpus using HTK, the hidden Markov model toolkit [7]. The recognition system is based on a continuous density HMM trained on approximately 10 hours of read speech from 22 speakers taken from a subset of the first stage of the CUCFN corpus. The speech features parameterization employed in training and test are mel-frequency cepstra, including both delta and delta-delta sub-features; cepstral mean subtraction is applied to all features on a per utterance basis. Triphone state clustering was carried out using broad acoustic phonetic classes similar to those used for English. The final model comprises approximately 2,800 state clusters each with 12 mixture Gaussian observation distributions.

5.1 CUCFN Development Test Set Results

A portion of the first stage of the CUCFN data corpus was set aside as a development test set. The test set contained speech by 7 male and 7 female speakers. The test set consisted of 1399 sentences containing 19376 words in total.

The 9.5K word bigram language model of Section 4 was used for initial recognition experiments on this test set. Results are reported in Table 3. As discussed earlier, this was done only as initial validation of the model and is reported here for interest only.

A second experiment was performed with a fair language model. The 63K word bigram model of Section 4 was used in an initial recognition pass. Lattices produced using the AT&T large vocabulary decoder [8] were then used with the 63K word trigram for further acoustic rescoring. As can be seen in comparison with the initial language model cheating experiment, the increased language model size and the OOVs significantly reduce accuracy. Overall we found that performance was relatively consistent across speakers with the exception of the worst subject, a female native speaker of Russian.

Table 3. Word error rates on the CUCFN development test set.

Type	Language Model	OOV Rate (%)	Word Accuracy (%)
	Vocabulary Size (words)		
2gram	9,500	0.0	78.66
2gram	63,000	6.7	69.88
3gram	63,000	6.7	71.33

6 Conclusion

We have described a project for the collection of spoken Czech language databases and have presented results of preliminary language modeling and speech recognition experiments. These experiments suggest that well-known acoustic modeling techniques should work well for large vocabulary Czech language speech recognition. However the relatively free word order of Czech and number of inflected word forms pose difficulties for language modeling.

Acknowledgments

Satellite news broadcast recordings were done under contract by the Linguistic Data Consortium, Philadelphia, PA, USA. We gratefully acknowledge use of the large vocabulary decoder provided by M. Riley and F. Pereira of ATT and the use of the SRI Language Modeling Toolkit provided by A. Stolcke of SRI. Language modeling data has been provided by the Lidove Noviny Publishers, Prague, Czech Republic.

References

1. P. Geutner, M. Finke, P. Scheytt: Adaptive Vocabularies for Transcribing Multilingual Broadcast News. ICASSP. 1998.
2. J. Hajic: Morphology Unification Grammar. PhD Thesis. Charles University. Prague. 1984.
3. J. Hajic: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning. Karolinum Press. Charles University. Prague. 1998. p. 106-132.
4. J. Psutka: Communication with Computer by Speech. Academia, Prague 1995, 287pp. (in Czech)
5. J. Nouza, J. Psutka, J. Uhlir: Phonetic Alphabet for the Recognition of Czech. Radioengineering, vol.6, no.4, pp.16-20, 1997.
6. <http://www.lidovenoviny.cz>
7. S. Young et al.: The HTK Book. Entropic Inc. 1999
8. M. Mohri, M. Riley, D. Hindle, A. Ljolje, F. Pereira: Full Expansion of Context-Dependent Networks in Large Vocabulary Speech Recognition. ICASSP 1998.