Czech Speech Recognizer for Multiple Environments *

Ondřej Glembek, Martin Karafiát, Lukáš Burget, Jan Černocký

Faculty of Information Technology, Brno University of Technology
E-mail: glembek@fit.vutbr.cz

Abstract

This paper presents our work on building a large vocabulary continuous speech recognition (LVCSR) system for Czech, capable of operation in multiple environments. SpeeCon and Temic speech databases were used to define a data-set for training acoustic models, attention was paid to unification of these two resources. The test set was also defined using these corpora with careful choice of segments not overlapping with the training data. The system was completed by a language model trained on Czech National corpus. The recognition was performed using DUCoder – an LVCSR stack decoder. Experimental results on the LVCSR task give a reference score of the system for future improvements.

1 Introduction

Automatic speech recognition (ASR) becomes an important part of many applications in IT as computation availability grows rapidly. Typical ASR tasks include speech recognition (i.e., automatic speech-to-text transcription), speaker identification, language identification, etc. No matter what the task is, good ASR performance is achieved by precise training of acoustic and language models using big amount of training data (i.e., transcribed acoustic utterances), which correspond to the desired task. For the Czech language, only telephone speech and broadcast news are publicly available. However, our task is to build a universal large vocabulary continuous speech recognition (LVCSR) system which would recognize wide-band speech in real environments, especially lectures.

This paper presents our work on building such system using multi-environment sources of data. We dispose two large speech databases—Czech SpeeCon, and Temic—for acoustic modelling and Czech National Corpus for language modelling.

The paper is organized as follows: section 2 describes the two speech databases, their structure, differences, unification, and separation into training and test sets. Section 3 describes how acoustic and language models were trained and their relation to the evaluation of the recognition. In Section 4, we present experimental results. A summary and future directions in research are presented in section 5.

*This work was partially supported by Grant Agency of Czech Republic under project No. 102/05/0278. Lukáš Burget was supported by post-doctoral grant of Grant Agency of Czech Republic No. 102/06/P383.

2 Data

Czech SpeeCon[1] is a speech database collected in the frame of EC-sponsored project "Speech Driven Interfaces for Consumer Applications". The database consists of 550 sessions, each comprising one adult speaker. The sessions were recorded in four different environments: office, entertainment, public place, car. Speakers taking part in recordings were selected with respect to achieve specified coverage from the point of view gender, age, and speaker dialects.

The recording device disposed four channels which were used for recording the same utterance with different characteristics (i.e., different types of microphone). The following setup was used: Sennheiser ME 104 and Nokia Lavalier HDC-6D for close talk, Sennheiser ME 64, Haun MBNM-550 E-L, AKG Q400 Mk3 T, and Peiker ME15/V520-1 for medium distance, and Haun MBNM-550 E-L for far distance.

The contents of the corpus is divided into four main sections: free spontaneous items (an open number of spontaneous topics out of a set of 30 topics), elicited spontaneous items, read speech (phonetically rich sentences and words, numbers, digits, times, dates, etc.), and core words. Out of this set, we chose free spontaneous items, and a subset of read speech comprising phonetically rich sentences and words.

The database was annotated orthographically including correcting the phonetic form of utterances [2]. To ensure maximum quality, all transcriptions were automatically checked for syntax, spelling, etc. These checks were based on comparison with already checked lexica. Selected annotations were hand checked, especially for usage of annotation marks.

Temic is a Czech speech data collection comprising 710 speakers collected for the TEMIC Speech Dialog Systems GmbH in Ulm[2] at Czech Technical University in Prague in co-operation with Brno University of Technology and University of West Bohemia in Plzen. Speaker coverage and content if items are similar to SpeeCon.

The audio data were recorded all in car under different conditions and in different situations (e.g., engine on, engine off, door slam, wipers on, etc.). The situations were transcribed by special marks in the annotation data. There were basically two different setups: Sennheiser ME102 as a close talk with AKG Q400 MK3 T as a far talk; and AKG Q400 MK3 T and Peiker ME27 both as far talk.

Annotation system differs in the two databases. It was necessary to unify the systems without loss of significant information. In both SpeeCon and Temic, leading underscore was used with subword fragmens, or synthetic words without sense used in uttered e-mail or web addresses, i.e. Internet items. Mispronunciations use single asterisk prepended to the word. Totally non-understandable word/speech is marked by double asterisk . Non-speech acoustic events were annotated by special marks which were either excluded or mapped to silence. The disturbing-noise marks were left in the label files for experiments. Furthermore, Temic prepends tilde symbol to foreign words. The marks were left in the label files as they are marked the same way in the dictionary. There is a difference in annotating spelled items in the two databases. While SpeeCon uses different variants of capital letter notation, Temic prepends a dollar to the lower-case letter names. We unified this by making all spelled items capital and we append underscore mark whenever the spelled letter is its name.

The phonetic alphabet uses 43 different phonetic elements which are covered in phonetically rich material. We excluded the phoneme "schwa" and mapped it to the silence

[1] http://www.speechdat.org/speecon/index.html

[2] http://www.temic-sds.com/english

model as there were little data to train it on. Encoding of the phonetic forms uses SAMPA[3].

Training and Test Sets Before splitting the unified collection into a training and test sets, we pruned away non-balanced and short utterances like city names, numbers, etc. Then, incorrect utterances (utterances with misspelled items, uncertain internet words, etc.) were excluded. The resulting set of sessions was into two disjoint sets. We decided to have at least three hours of test data with no single-word utterances. In order to eliminate similar phoneme contexts, we excluded those utterances from training set, whose prompts match at least one utterance in the test set. We ended up with 59 hours of training data and 3 hours of test data.

3 Recognition System

Acoustic modeling Speech features are 13 PLP coefficients augmented with their first and second or derivatives (totaly 39 coefficients) with cepstral mean and variance normalization applied per conversation side. Acoustic models are based on left-to-right 3 state word-internal triphone HMMs with states tied according to phonetic decision tree clustering. For all experiments, baseline system was entirely trained using HTK tools [1].

We trained the system on the described training set under maximum likelihood framework and we used close talk microphone setup. We started by training 2-Gaussian-per-state monophone models followed by repetitive increasing of number of Gaussian mixtures by two and retraining models (four iterations of Baum-Welch) until 16-Gaussian models were trained. Using these, we performed forced alignment, expanded monophones to context-dependent word-internal triphones, and performed clustering, which resulted in 3920 tied states. Then we used the same retraining iteration as for monophones and we produced 8-, 16-, and 32-Gaussian word-internal HMMs[4].

Language model We used the same LM as in [6]. Czech National Corpus[5] was used as the main source for textual data both for LM creation and dictionary generation. The corpus was provided by colleagues form Faculty of Informatics of Masaryk University in Brno. Its size is 4.6 GB[6] and its vocabulary comprises almost three milion words. We eliminated the vocabulary by specifying a word frequency threshold, which was experimentally chosen 30. This step reduced the vocabulary size to approximately one seventh of its original size. We used SRILM[7] toolkit with default options to create a trigram language model with Good-Turing discounting and Katz back-off smoothing. The resulting LM was merged with a smaller LM trained on lecture data [6]. With our testset, the OOV was 239 and the perplexity was 543.56. Such high number was caused by the size of the LM. Our test set vocabulary is only a 7000 words which is a mere fraction of the LM vocabulary.

Recognition We used the Duisburg University HDucode [4] as LVCSR decoder. It's a multi-stack LVCSR decoder which performs the search for the most probable word

[3]http://www.phon.ucl.ac.uk/home/sampa/home.htm

[4]number of Gaussian mixtures for silence models is double

[5]http://ucnk.ff.cuni.cz/

[6]Note the physical size of the corpus file. Working with this amount of data was extremely time and space consuming therefore the corpus was split into blocks of 600000 sentences and processed in parts.

[7]http://www.speech.sri.com

sequence in recognition systems that make use of the HMMs and backoff n-gram LMs. The disadvantage of the decoder is itsinability of use of cross-word HMMs.

4 Results

The results were evaluated using word accuracy. Two important constants needed to be tuned: the word insertion penalty (WIP) and grammar scale factor (GSF). Watching the best accuracy, we set the WIP to -10 and GSF to 12. The following results were achieved: 71.79% for 8-Gaussian, 74.44% for 16-Gaussian, and 73.64% for 32-Gaussian models. Lower accuracy with 32-Gaussian models was caused by insufficient amount of training data.

5 Conclusions

In this paper we presented a baseline Czech LVCSR recognition system which works with 74.44% word accuracy using 16-Gaussian word-internal models. Close talk microphone setup was artificially used in order to tune the system, however future work will be oriented to variable setup. In our experiments we used standard maximum likelihood training of acoustic models, which serve as a reference recognition system. No adaptation or other advanced training techniques were used, however this is the future direction of our research. We plan to use cross-word models and perform adaptation on user and channel as well as use other training techniques: vocal tract length normalization (VTLN), minimum phoneme error (MPE), heteroscedastic linear discriminant analysis (HLDA), posterior based features [7], etc.

References

[1] Young, S., Evermann, G., Hain, T., Kershaw, D., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P.: *The HTK Book*. Entropics Cambridge Research Lab., Cambridge, UK, 1996

[2] Pollák, P., Černocký, J.: *Orthographic and Phonetic Annotation of Very Large Czech Corpora with Quality Assessment*, Czech Technical University in Prague, Faculty of Electrical Engineering, Brno University of Technology, Faculty of Information Technology

[3] Moore, J., Kronenthal, M., Ashby, S.: *Guidelines for AMI Speech Transcriptions*, 10 February 2005, Version 1.2

[4] Willet, D., Neukirchen, C., Rigoll, G.: *Ducoder — The Duisburg University LVCSR Stack Decoder*. Department of Computer Science, Faculty of Electrical Engineering, Gerhard-Mercator-University, Duisburg, Germany

[5] Park, A., Hazen, T.J., Glass, J.R.:*Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary selection and Language Modeling*, MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA, 2005

[6] Glembek, O.: *Automatic Lecture Indexing Using Voice Recognition*, Diploma thesis, Brno University of Technology, Faculty of Information Technology, 2005

[7] Karafiát, M., Grézl, F., Černocký, J. (2004): *TRAP based features for LVCSR of meting data*, In INTERSPEECH-2004, 437-440.