Advances in acoustic modeling for the recognition of Czech

Jiří Kopecký, Ondřej Glembek, Martin Karafiát

Speech@FIT, Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
{kopecky,glembek,karafiat}@fit.vutbr.cz

Abstract. This paper presents recent advances in Automatic Speech Recognition for the Czech Language. Improvements were achieved both in acoustic and language modeling. We mainly aim on the acoustic part of the issue. The results are presented in two contexts, the lecture recognition and SpeeCon+Temic test set. The paper shows the impact of using advanced modeling techniques such as HLDA, VTLN and CMLLR. On the lecture test set, we show that training acoustic models using word networks together with the pronunciation dictionary gives about 4-5% absolute performance improvement as opposed to using direct phonetic transcriptions. An effect of incorporating the "schwa" phoneme in the training phase shows a slight improvement.

Key words: Automatic Speech Recognition, LVCSR system, acoustic modeling, HLDA, VTLN, CMLLR, lectures recognition

1 Introduction

In the framework of e-learning, more and more lectures and seminars are recorded, streamed to the Internet and stored to archives. To add value to the recordings, users are allowed to search in the records of the lectures and browse them efficiently. Large vocabulary continuous speech recognition (LVCSR) is used to produce recognition lattices to cope with standard word and phrase indexing and search.

Although a lot of work has been done in the Czech domain in the past years ([4, 9], advanced techniques of acoustic modeling, such as HLDA, VTLN, CMLLR, discriminative training, etc., have been studied more thoroughly for English tasks. Our work aims at incorporating these techniques for the Czech spontaneous speech, especially lectures recognition.

In section 2, description of advanced techniques used for acoustic modeling is presented. Description of all data is given in section 3. Section 4 contains some information about the used recognizer through our experiments. Section 5 shows achieved results on different test sets. The paper concludes with a summary and states future work in section 6.

2 Acoustic modeling techniques

The investigated techniques apply standard speech recognition based on context-dependent Hidden Markov models (CD-HMM) [13]. The following techniques were used in our experiments. Their setup was based on previous experiments run on the English tasks [11]

2.1 HLDA

Heteroscedastic linear discriminant analysis (HLDA), which was first proposed by N. Kumar [5, 6], can be viewed as a generalization of Linear Discriminant Analysis (LDA). LDA is a data driven technique looking for linear transformation allowing for dimensionality reduction of features. Like LDA, HLDA assumes that classes obey multivariate Gaussian distribution, however, the assumption of the same covariance matrix shared by all classes is relaxed. HLDA assumes that n-dimensional original feature space can be split into two statistically independent subspaces: While in p useful dimensions (containing discriminatory information), classes are well separated, in $(n-p)$ nuisance dimensions, the distributions of classes are overlapped. In our case, the classes are the Gaussian mixture components.

2.2 CMLLR

Maximum likelihood linear regression (MLLR) is an adaptation technique based on estimating linear transformations for groups of model parameters by maximizing the likelihood of the adaptation data [2]. Unlike MLLR, which allows different transforms for the means and variances, *constrained MLLR* (CMLLR) aims at estimating a single transformation for both the means and variances. This constraint allows to apply CMLLR online by transforming the features [3].

2.3 VTLN

Vocal tract length normalization (VTLN) is a speaker based normalization based on warping of frequency axis by speaker dependent warping factor [7]. The normalization of vocal tract among the speakers has a positive effect on reduction of inter-speaker variability. The warping factor is typically found empirically by a searching procedure which compares likelihoods at different warping factors. The features are repeatedly coded using all warping factors in searching range, typically 0.8-1.2, and the one with best likelihood is chosen.

3 Data

3.1 Training Data

Czech SpeeCon[1] is a speech database collected in the frame of EC-sponsored project "Speech Driven Interfaces for Consumer Applications". The database

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

consists of 550 sessions, each comprising one adult speaker. The sessions were recorded in four different environments: office, home, public place, car. Speakers taking part in recordings were selected with respect to achieve specified coverage regarding gender, age, and speaker dialects.

The content of the corpus is divided into four 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. To ensure maximum quality, all transcriptions were automatically checked for syntax, spelling, etc. These checks were based on comparison with already checked lexicon. 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 of the items are similar to SpeeCon. The audio data were all recorded in car under different conditions and in different situations (e.g., engine on, engine off, door slam, wipers on, etc.). The annotation systems used in these databases were unified without loss of significant information.

Utterances matching the following criteria were pruned out: non-balanced and short utterances (e.g. city names, numbers), broken utterances (containing misspelled items, uncertain internet words, etc.). We ended up with 59 hours of data, 56 hours of which were left for training.

3.2 Test Data

We created two different test sets through the work on Czech recognition system:

- SpeTem test set – contains about 3 hours of speech and is derived from the same corpus as the training data.
- Lecture test set – the target domain of our work is decoding of lectures. Hence we have chosen two lectures recorded and transcribed on our faculty as the second test set: The first lecture from the "Information Systems Project Management" (IRP) course in total time 1.6 hours of speech and the second lecture from "Multimedia" (MUL) course containing about 1 hour

3.3 Language Model Data

We used a general bigram language model (LM) for all decoding of our experiments and acoustic model comparison. Furthermore, SpeTem test set was

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

expanded by a trigram LM. Both LM's were trained on the Czech National Corpus [1]. The subset chosen for training contains nearly 500M word forms. This is an extremely heterogeneous corpus that consists of texts pertaining to different topics and thus can serve as the basis of a general language model. The corpus contains 2.8M different word forms. At the stage of vocabulary construction, we included in the vocabulary only those words that appear in the corpus at least 30 times. That resulted in the smaller vocabulary of 350K words. Even such a vocabulary is presently considered as extremely large for LVCSR tasks. However, we did not want to reduce it any further because inflectional nature of the Czech language calls for larger vocabularies (as compared to English) in recognition of continuous speech [12]. Good-Turing discounting with Katz backoff was used for language model smoothing, singleton N-grams were discarded.

3.4 Data Processing

The phonetic alphabet uses 43 different phonetic elements which are covered in phonetically rich material. It covers 29 consonants, 11 vowels, 3 diphones. The monophone set further includes one special model for silence and also all speaker or background noises and finally the last model representing short pauses between words. Encoding of the phonetic forms uses modified SAMPA[3].

Originally, the handling of "schwa" was rudimentary and for example in spelled items, there were only plosive models (such as 't', 'd', etc.) followed directly by silence; we excluded the phoneme "schwa" and mapped it to the silence model. Because of our training databases contain precious phonetic transcriptions, we could use them directly for the training. Acoustic models trained on this base will be called as *version .v0.* For all experiments, the baseline system was trained using HTK tools [13],

The following work led us to complete our phoneme set with the "schwa" phoneme. Presently, we are also using our own training toolkit STK [4] developed at Speech@FIT group, which allows training from phoneme networks. We created them from word transcriptions and pronunciation dictionary included in training databases and used them in the training process instead of straight phonetic string. This approach allows more freedom by choosing the correct pronunciation variant of each word. This multi-pronunciation occurs mainly in foreign and non-literary words in our training set. The influence of this newer acoustic models (marked as *version .v1*) is investigated in section 5.

It was not clear what exactly brings the improvement achieved by acoustic models in version .v1 - the new phoneme schwa or training from networks? Therefor we decided to train another acoustic models (*version .v2*) where the schwa was mapped on silence model again but the training process was done from phoneme networks. This work is still in the beginning, therefore table 2 is not complete yet. However it has been shown, that training using network, brings most of the improvements.

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

[4] Lukáš Burget, Petr Schwarz, Ondřej Glembek, Martin Karafiát, Honza Černocký: STK toolkit, http://speech.fit.vutbr.cz/cs/software/hmm-toolkit-stk-speech-fit

Acoustic models	2gram decoding	3gram expansion
xwrd.sn2	22.33	20.92
xwrd.sn2.hlda	21.46	19.35
xwrd.sn2.cmllr	20.38	18.17
xwrd.sn2.vtln0	20.47	18.48
xwrd.sn2.vtln1	20.24	18.20
xwrd.sn2.vtln2	20.19	18.34
xwrd.sn2.vtln3	20.31	18.43
xwrd.sn2.vtln4	20.31	18.26
xwrd.sn2.vtln5	20.41	18.41
xwrd.sn2.vtln1.hlda	19.87	17.45
xwrd.sn2.vtln1.cmllr	**19.03**	**16.93**
xwrd.sn2.vtln1.cmllr.hlda	19.28	17.30

Table 1. *Comparison of different advanced techniques in acoustic modeling on SpeTem test set.*

4 Recognizer

Speech features are 13 PLP coefficients augmented with their first and second derivatives (39 coefficients in total) with cepstral mean and variance normalization applied per conversation side. Acoustic models are based on left-to-right 3 state cross-word triphone HMMs with states tied according to phonetic decision tree clustering. Number of tied states was tuned around 4000 in the phase of state clustering. After one phase of retraining, clustering was performed once more.

5 Experimental results

All results are presented in terms of word error rate (WER).

As can be seen in table 1, each technique gives some improvement – about 1% (from HLDA), almost 2% (thanks to CMLLR and VTLN) absolutely. Vtln0 represent the acoustic vtln-models trained on the output from non-vtln models; vtln1-5 was trained on previous iteration of the vtln models. We also tried to combine these techniques. For this purpose we used vtln1 models and CMLLR, HLDA transformations. The results are presented in the second part of table 1. Surprisingly, the combination of both transformations with VTLN performs worse than its combination separately. However, we are still able to improve our results by 4% by using these advanced modeling techniques.

Table 2 show some results on lecture test set in WER [%] by using only 2gram decoding network. Presently, we are working on this issue, so the table is not complete so far. Three versions of acoustic models are compared:

- .v0 acoustic models without schwa trained straight from phoneme strings
- .v1 acoustic models with schwa trained from phoneme network
- .v2 acoustic models without schwa trained from phoneme network

Acoustic models	IRP			MUL		
	.v0	.v1	.v2	.v0	.v1	.v2
xwrd.sn2	52.78	48.12	48.59	61.79	56.33	57.57
xwrd.sn2.hlda	51.48					
xwrd.sn2.cmllr	51.38					
xwrd.sn2.vtln0	45.69	**42.47**		**54.12**	54.19	
xwrd.sn2.vtln1	45.44					
xwrd.sn2.vtln2	44.93					

Table 2. *Results achieved on the Lecture test set.*

What we can see is the significant improvement by using VTLN adaptation. The change of training method from phoneme string to phoneme network helps too. The influence of new "schwa" model is not so noticeable but even now we can see a little improvement of acoustic models even though "schwa" is omitted in the decoding part.

6 Conclusions

We have used some advanced acoustic modeling techniques, which were successfully tested in the English LVCSR. Not only their effect was visible on the SpeTem test, but mainly on the target lecture test set, where the baseline system gave poor results.

The effect of network training is eminent in lecture decoding, therefore we need to complete our experiments. Another improvement is expected from discriminative training of our acoustic models [8] and usage of posterior features [10]. We also expect improvement by integrating the "schwa" phoneme into decoding networks.

Acknowledgments

This work was partly supported by Ministry of Trade and Commerce of Czech Republic under project FT-TA3/006 and by Ministry of Interior of Czech Republic under project VD20072010B16. The hardware used in this work was partially provided by CESNET under project No. 201/2006.

References

1. Český národní korpus (Czech National Corpus). Technical report, Ústav Českého národního korpusu FF UK, Praha, Česká republika, 2005.
2. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

3. M. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical Report CUED/FINFENG/TR291, Cambridge University, 1997. Also available as `http://citeseer.ist.psu.edu/article/gales98maximum.html`.
4. Nouza J., Ždánsky J., Červa P., and Kolorenč J. Continual On-line Monitoring of Czech Spoken Broadcast Programs. In *International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP 2006)*, pages 1650–1653, Pittsburgh, USA, 2006.
5. N. Kumar. *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, John Hopkins University, Baltimore, 1997.
6. N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communcation*, 26:283–297, 1998.
7. L. Lee and R. Rose. Speaker Normalization Using Efficient Frequency Warping Procedures. In *Proc. ICASSP 1996*, pages 339–341, Atlanta, GA, USA, May 1996.
8. D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University Engineering Dept, 2003.
9. J. Psutka. *Komunikace s počítačem mluvenou řečí*. Academia, Praha, 1995.
10. František Grézl, Martin Karafiát, Stanislav Kontár, and Jan Černocký. Probabilistic and bottle-neck features for LVCSR of meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pages 757–760, Hononulu, USA, 2007.
11. Hain Thomas, Wan Vincent, Burget Lukáš, Karafiát Martin, Dines John, Vepa Jithendra, Garau Giulia, and Lincoln Mike. The AMI System for the Transcription of Speech in Meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, pages 357–360, Hononulu, USA, 2007.
12. E.W.D. Whittaker. *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*. PhD thesis, Cambridge University, 2000.
13. S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book*. Entropics Cambridge Research Lab., Cambridge, UK, 2002.