THE RECOGNITION OF CZECH LECTURES BY LVCSR SYSTEM

Jiří Kopecký
Doctoral Degree Programme (4), FIT BUT
E-mail: kopecky@fit.vutbr.cz

Supervised by: Jan Černocký
E-mail: cernocky@fit.vutbr.cz

ABSTRACT

This paper presents a summarization of our recent work in automatic speech recognition of Czech lectures. We will familiarize with a data preparation and a system training itself as well as its testing (evaluating of the system performance). For purpose of this paper, the complex speech processing issue is quite simplified. We show that training acoustic models using phoneme networks gives about 4-5% absolutely performance improvement as opposed to using direct phonetic transcriptions. An effect of incorporating the "schwa" phoneme in the training phase shows a slight improvement.

1 INTRODUCTION

As the name of this paper suggests, LVCSR systems can be used for automatically textual recording of the speech (like parliament or court acts, meetings, lectures, etc.), subtitle creation or almost all other speech processing tasks (like keyword spotting, language identification, . . .). Our work aims at the Czech spontaneous speech recognition – lectures especially.
Section (2) describes basics of a **L**arge **V**ocabullary **C**ontinues **S**peech **R**ecognition (LVCSR) system, their principles of functionality and possible usage. In section (3) will be shown all steps needed be done before the system training, followed by the acoustic models training itself. The next section (4) contains all about an issue of system testing. Section 5 shows achieved results and a summary of different systems. The paper concludes with states of future work in section 6.

2 LVCSR SYSTEM

What does it mean "LVCSR" system? Basically, you can imagine some kind of software (decoder) with any audio record on its input (=continues speech recognition; not only single words of phrases), which produces transcription (textual representation what was said) of the speech on its output. And large vocabulary means, that these systems can recognize tens or hundreds thousands of different words. But there are still some words, which stay unrecognized – so called Out Of Vocabulary (OOV) words (mainly proper names, less often words, . . .).

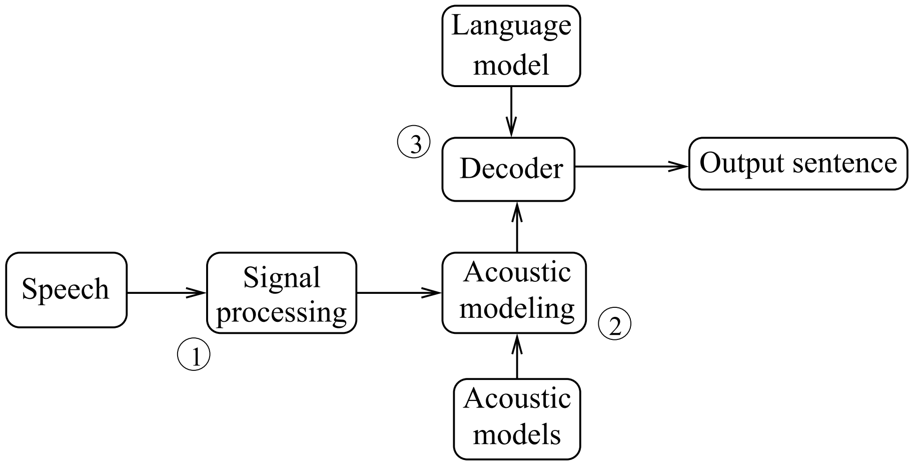

Figure 1: Data flow through the common LVCSR system

2.1 PARTS OF THE SYSTEM

You can see the data flow through the system on the figure 1. Initially the feature extraction is needed (1). Afterwards modified input audio data are recognized by the acoustic Hidden Markov Models (HMM) and the phonetic strings or nets are produced (2). A language model provides correct word connections and we finally obtain readable and hopefully reasonable output (3).
We have introduced a lot of new phrases in the last paragraph. Let us explain you all of them in two following sections more sophisticated.

3 SYSTEM TRAINING

3.1 TRAINING DATABASES

For our purpose, we have got 2 czech audio databases with relevant transcriptions – SpeeCon and Temic. We used only whole sentences for initial version of the czech system training – it contains about **56 hours** of speech. Another **3 hours** are used for a cross-validation (a tuning of the system characteristics and settings). More details about these databases are found in SpeeCon and Temic.

3.2 TRANSCRIPTION PREPARATION

First we have to prepare all audio records and corresponding transcriptions = textually written what was said. The databases generally contain word transcriptions and pronunciation dictionary (like in our case). In the dictionary, there are all words through the database and corresponding sequences of monophones representing all possible pronunciations of each word.
The training tool requires specific form – there has to be labels with additional information of start and end time for each sentence used for training. We create phoneme string or net from the

words according to the pronunciation dictionary and training toolkit. And now, we can move on the next step – feature extraction from input acoustic data.

3.3 AUDIO DATA PREPARATION – FEATURE EXTRACTION

Simple audio record contains a lot of useless information for a speech processing. Hence we modify the audio data to take out only suitable data for classifier recognizer – so called **feature extraction**. Shortly, speech signal is divided into the typically 25ms segments with 10ms shift, where speech is regarded as stationary. A multiplication of speech waveform with hamming window can be used for this purpose. The power Fourier spectrum is computed for every speech segment. And perceptual linear predictive analysis [2] leads to creation of final PLP coefficients (organized into so called feature vectors).

3.4 ACOUSTIC TRAINING – HMM MODELING

Hidden Markov Model (HMM) can be understood as a finite state machine. It makes transition from a one state to another with some probability and every discrete time step is generated a feature vector (based on state distribution). HMM could represent each word from the dictionary, but we would not be able to train such a huge amount of models. Hence the words are split into smaller units (according to pronunciation dictionary) and these are trained. We are using 3-state context dependent HMM with two additional non-emitting (non-distribution) state used for connection with another models (used for creating whole words etc.).
The state model distribution is modeled using mixture of multivariate Gaussians (GMM) with diagonal covariance matrices. We can estimate the HMM parameters from known and correct training transcription. The most popular training scheme is maximum likelihood (ML) parameter estimation by using the Baum-Welsh algorithm [5]. The goal is to find such setting of values, that maximizes the likelihood of training data.
We work with set of 45 models of monophones – 29 consonants, 11 vowels, 3 diphones. One special model for silence and all speaker or background noise and the last model representing short pause between words. After few training iteration, these monophones are expanded into cross-word (xwrd) models and retrained.

3.5 SCHWA ISSUE AND TRAINING FROM NETS

We ware using pretty simplified approach till now. Originally, 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 by HTK tools [5] on this base will be called as *xwrd.v0*.
The following work led us to complete our phoneme set with the “schwa” phoneme. Presently, we are also using our own training toolkit STK 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 *xwrd.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 (*xwrd.v2*) where the schwa was mapped on silence model again but the training process was done from phoneme networks. In table 1 is shown the answer, what brings most of the improvements.
For now, we have got tree different acoustic models. And we would like to find out, what set is better. How to test LVCSR systems will be show in the followed section, as well as all needful parts for decoding.

4 SYSTEM PERFORMANCE

4.1 TEST DATA

Lecture decoding is the main aim of our work for now. Hence we have chosen two lectures recorded and transcribed on our faculty: first from the "Information Systems Project Management" (IRP) course in total time 1.6 hours of speech and the second from "Multimedia" (MUL) course containing 1 hour of speech.

4.2 LANGUAGE MODEL – LM

Language model significantly participates on final performance. Basically, it forms word sentences by choosing the most probably phoneme sequence (from the phoneme network obtained after acoustic decoding). All our experiments mentioned in this paper are decoded by general bigram LM trained on the Czech National Corpus extended at Masaryk University in Brno [1]. This general LM is able to use for any kind of task, but the specialized LM brings significant improvement. More details about this issue you can find in paper [4].

4.3 WORD ERROR RATE (WER) – EVALUATION MATRIC

Every system has to be uniformly evaluated. Decoders make three types of mistakes: deletion error (the word is missing in recognized output at all), insertion error (the word is recognized but should not be) and substitution error (recognized word is different from expected word). When we have got correct transcriptions of the test data, we can compute ratio between wrong recognized words and all words to obtain the final **W**ord **E**rror **R**ate (WER - as the sum of all errors divided by all words in the test set). The smaller value (in percents) is the better.
There is one more influence on the final WER. As we mentioned earlier, large vocabulary can not include all possible words of concrete language. The words occurring in the test set and not in the dictionary (OOV words) inflicts errors not only in the word itself but affects the surrounding words too. Hence we try have as less OOV rate as possible to obtain the best possible WER.

5 RESULTS AND CONCLUSIONS

Table 1 shows results on two lectures and tree acoustic models (as was mentioned before in 3.5). On the first view, we can see a significant improvement (4.5-5.5% absolutely) of the xwrd.v1

Lecture	Acoustic model set		
	xwrd.v0	xwrd.v1	xwrd.v2
IRP	52.78	**48.12**	48.59
MUL	61.79	**56.33**	57.57

Table 1: *Overall results overview. All numbers show WER in percents for different LVCSR systems.*

system against xwrd.v0 system. It is also a little better then xwrd.v2 system (especially in the MUL lecture, where more abbreviations occurs). It means, that training using network, brings most of the improvements.

The overall high values of WER are not so pleasant. We are dealing with natural speech but we are using general LM. Only the change of the language model to some more specialized one improves the results about 10% absolutely (by the xwrd.v0 system).

6 FUTURE WORK

There exists a lot of advanced acoustic modeling techniques (like HLDA, VTLN, CMMLR, . . .) with significant influence on the improvement of system performance. Some of them we have already implemented for the xwrd.v0 system (for example the VTLN adaptation [3] improved results about 7% absolutely). These techniques had to be also done for the newer acoustic models (xwrd.v1 and xwrd.v2).

The acoustic models could also be adapted for specific task – by another lectures in our case. All acoustic models will also be trained on whole databases (not only sentences) to obtain more precise acoustic models. Integration of “schwa” into decoding process should be the first issue to be work out.

REFERENCES

[1] Český národní korpus (Czech National Corpus). Technical report, Ústav Českého národního korpusu FF UK, Praha, Česká republika, 2005.

[2] H. Hermansky. Perceptual linear predictive (PLP) analysis for the speech. *J. Acous. Soc. Am.*, pages 1738–1752, 1990.

[3] L. Lee and R. Rose. Speaker Normalization Using Efficient Frequency Warping Procedures. In *Proc. ICASSP 1996*, pages 339–341, Atlanta, GA, USA, May 1996.

[4] T. Mikolov. Language Models for Automatic Speech Recognition of Czech Lectures. In *Proc. EEICT*, 2008.

[5] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book*. Entropics Cambridge Research Lab., Cambridge, UK, 2002.