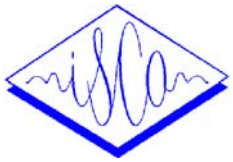

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# Using Various Language Model Smoothing Techniques for the Transcription of a Weather Forecast Broadcasted by the Czech Radio

*Luděk Müller and Josef Psutka*

University of West Bohemia, Department of Cybernetics, Univerzitní 8, 306 14 Plzeň, Czech Republic
muller@kky.zcu.cz, psutka@kky.zcu.cz

## ABSTRACT

This paper presents an experimental speech recognition system used to transcribe a weather forecast broadcasted by the Czech radio[1]. The system is based on the HMM with mixture Gaussian continuous densities and is designed as a speaker independent. To overcome very sparse training data various language models supported by smoothing of model parameters based on the leaving-one-out technique, discounting and backing-off approach were tested. The results of recognition experiments are discussed in the paper.

## 1. INTRODUCTION

Presented paper concerns the recognition of continuous Czech speech. The solved task is to recognize and transcribe a weather forecast transmitted by the Czech radio. It seems to be a task with a relatively small vocabulary and a low task perplexity but with very sparse text data for the language modeling. The goal of our work was to develop the first experimental recognition system for continuous Czech speech, to evaluate it for several kinds of language models and to use it as a reference for various experiments in our future research.

In this article experiments with two kinds of language models and with several various language model parameters are described. The *n*-gram discounting language model for joint probabilities and the *n*-gram absolute discounting language model with backing-off for conditional probabilities were tested. As a consequence of very sparse text training corpus a smoothing technique was used to model unseen events. The references to a case of the system working with no language model as well as the results for an approximated but faster implementation of a language model are given.

## 2. DATA ACQUISITION

The speech data were obtained as follows. The radio signal was sampled and saved on a PC equipped with a radio card. Then the sampled data were (off-line) brought to a computer providing both the training and the recognition. Weather forecast radio speech data were collected for the acoustic modeling and weather forecast radio text data for the language model construction.

The total number of available utterances was 2251. Data were divided into 2151 training and 100 test utterances. Each recorded utterance was listen, carefully checked, and manually annotated. So-called detailed orthographic transcription with aspect to a Czech language was used during this phase. Subsequently the annotation of each utterance was automatically transcribed to its phonetic baseform. The phonetic baseform is a sequence of Czech phones that are defined according to the Czech Phonetic Alphabet. Because no Czech word pronunciation dictionary is available a set of Czech phonological production rules was defined and used to obtain a very precise phonetic baseform for each phrase used for both the training and the recognition. After the phonetic transcription phase each fonetic baseform was automatically transformed into the so-called post-transcription form suitable for an automatic acoustic modeling procedure. The phonetic vocabulary containing phonetic baseforms for each word in vocabularies was generated during this post-transcription step with respect to the fact that more than one post-transcription form for one word can exist. To illustrate this process the following example is given:

Utterance:[2]

```
tlaková tendence slabý pokles
```

Detailed orthographic transcription:

```
[LIP_SMACK]  [LOUD_BREATH]  tlaková
tendence slabý pokles
```

Phonetic transcription:

```
sil  [LIP_SMACK][LOUD_BREATH] t l a k
o v aa  sp  t e n d e n c e  sp s l a
b ii  sp  p o k l e s  sil
```

Phonetic post-transcription:

```
sil  t l a k o v aa  sp  t e n d e n
c e  sp  s l a b ii  sp  p o k l e s
sil
```

[1] This work was supported by the Grant Agency of Czech Republic grant no. 102/96/K087.

[2] The meaning is "pressure tendency weak decrease".

## 3. THE ACOUSTIC PROCESSOR

The aim of the acoustic processor is to convert the continuous speech signal into a sequence of feature vectors. The parameterization process used in our system is based on the Mel-Frequency Cepstral Coefficients (MFCCs). The audio signal of the Czech weather forecast FM radio speech is digitized at 16 kHz sample rate and 16 bit resolution. The pre-emphasized acoustic waveform is then segmented into 25 millisecond frames every 10 ms. Hamming window is applied to each frame and MFCCs are computed to generate final acoustic vectors. We used a filterbank with 26 triangular filters, 13 MFCCs including the 0th coefficient. The delta and delta delta MFCCs are computed and appended to the static MFCCs for each speech frame.

## 4. ACOUSTIC MODELING

The basic speech unit of our system is a phone. A phone is sound of speech corresponding to one phoneme or to a short pause or a long silence. The list of all Czech phones was established and is given in [2].

Each individual phone is represented by a hidden Markov model (HMM). Each HMM excluding model of the short pause is left-to-right and has five states connected by arcs. The three inner states are emitting states each of them with two output arcs (transitions) - self-loop and forward transition. The entry and exit states are non-emitting. To model a long silence the standard topology of HMM was modified by adding extra skip and backward transitions in the model. A short pause has a three-state left-to-right model with only one emitting state between the non-emitting entry and exit states and an extra direct transition from entry to exit state. The output probability distribution function assigned to each state is expressed by mixture of multivariate Gaussians where each Gaussian has a diagonal covariance matrix. The number of mixtures for each model was obtained experimentally and there were used 8 mixtures for each model.

We have assumed so far that only one HMM is required per phone, and since 45 phones are needed for Czech, it may be thought that there should be only 45 HMMs in our system at all. However the consideration of contextual effect leads to context modeling. Our system uses triphones instead phones to model phones in different context. Because the number of Czech triphones is large the phonetic decision trees were used to tie states of Czech triphones what gives us more robust estimates for the parameters of the tied-states. That is because all the data associated with each individual state can be pooled and after tying several states share the same output probability distribution. There are about 2500 tied-states in the system and thus 2500 different distribution functions in comparison with 79 507 all possible different Czech triphones and 79 507*3 different triphone states.

For finding Maximum-Likelihood estimates of HMM parameters we used the Baum-Welch algorithm.

## 5. LANGUAGE MODELING

In our task the statistical language model based on joint probabilities with a discount factor [3] was first applied. Secondly, the language model based on conditional probabilities with discounting and backing-off [1], [3] was tested. The solved task was unfortunately supported by very sparse data. We had only 2252 sentences at our disposal rewritten from real transmitted weather forecasts containing 540 distinct words. This "corpus" was split into the training part with 1264 sentences containing 16917 words and the test part with 988 sentences (12188 words). Using a training text uni-, bi- and trigram statistics were counted for both types of the models and the perplexity of the training and the test corpora was verified. Note we supposed that all 540 words of the vocabulary are known but because the whole corpus was split into two parts randomly so the 101 words were unseen in the training corpus and their unigram statistics had to be also estimated.

**Language model with discount factor and joint probability**

The language model with discount factor and joint probability uses relative frequencies of events (words) related to the all $N$ words in the training text. The relative frequencies are then expressed as $N(h,w)/N$, where $N(h,w)$ is a count of observations for joint event $(h,w)$. The problem is how to cover events that were not seen in the training text but could be observed with nonzero probability in the test text. To solve it we performed a redistribution of a part probability mass among unseen events by means of count dependent discount factors $\lambda_{N(h,w)}$. Thus we obtained the language model with discount factor and joint probabilities $\bar{p}(h,w)$

$$\bar{p}(h,w)=\begin{cases} R/N & \text{for } N(h,w)=R \\ (1-\lambda_{N(h,w)})\dfrac{N(h,w)}{N} & \text{for } 0<N(h,w)<R \\ \dfrac{1}{n_0(.,.)}\displaystyle\sum_{h'w':0<N(h',w')<R}\lambda_{N(h',w')}\frac{N(h',w')}{N} & \text{for } N(h,w)=0\,, \end{cases} \quad (1)$$

where $R$ being the maximum count among events of the training text and $n_0(\,.\,,\,.\,)$ is the number of distinct joint events $(h,w)$ that occurred exactly 0 times.

As can be seen the discount factor is only unknown parameter in the distribution (1). To find optimal values of discount factors the maximum likelihood criterion extended by the principle of leaving-one-out was used. The likelihood function was constructed using estimated probabilities (1) and with the leaving-one-out criterion

$$L(\lambda_{N(h,w)})=\log\prod_{n=1}^{N}p_v(h_n,w_n)=\sum_{hw}N(h,w)\log p_v(h,w)\,, \quad (2)$$

where $p_v(h,w)$ being the likelihood function (1) modified by a successive leaving of particular events. Using the partial derivatives with respect to $\lambda_{N(h,w)}$ and setting them equal to zero we get the relation for the discount factors

$$\lambda_r=1-\frac{(r+1)\,n_{r+1}(.,.)}{r\,n_r(.,.)}\left[1-\frac{R\,n_R(.,.)}{N}\right] \quad (3)$$

where $n_r(.,.)$ is the total number of joint events $(h,w)$ that occurred exactly $r$ times and $r = 1, \ldots, R-1$. By the substitution $\lambda_r = \lambda_{N(h,w)}$ derived in (3) for the discount factors in (1) we obtained the expected forms of estimations $\bar{p}(h,w)$.

Perplexities of the training and the test corpora computed for this language model are summarized in the Tab.1. To compute estimations $\bar{p}(h,w)$ the distribution of the counts $n_r(.,.)$ should be known for $r = 1, \ldots, R$. Regarding a poor training text many of these counts remain equal to zero and have to be estimated by the linear interpolation from the non-zero counts.

| **Language model with discount factor and joint probability** | | | |
|---|---|---|---|
| | | Training corpus | Test corpus |
| **Unigram** | Entropy ($\hat{H}$) | 7.154 | 7.318 |
| | Perplexity ($PP$) | 142.453 | 159.620 |
| **Bigrams** | Entropy ($\hat{H}$) | 2.433 | 4.395 |
| | Perplexity ($PP$) | 5.440 | 21.033 |
| **Trigrams** | Entropy ($\hat{H}$) | 1.093 | 5.697 |
| | Perplexity ($PP$) | 2.157 | 51.890 |

Tab.1 Statistics for the training and the test corpora of the language model with discount factor and joint probability.

**Language model with absolute discount factor, backing-off and conditional probability**

The language model with absolute discount factor and backing-off uses the conditional probabilities and is defined as

$$\bar{p}(w|h) = \begin{cases} \dfrac{N(h,w) - b_h}{N(h,.)} & \text{for } N(h,w) > 0 \\ b_h \dfrac{W - n_0(h,.)}{N(h,.)} \dfrac{\beta(w|\bar{h})}{\sum_{w':N(h,w')=0} \beta(w'|\bar{h})} & \text{for } N(h,w) = 0, \end{cases} \quad (4)$$

where $b_h$ denotes the discount factor dependent on the history $h$, $W$ is the number of distinct words in the vocabulary, $N(h,.)$ is the number of joint events $(h,.)$ for a fixed history $h$, $n_r(h,.)$ is the number of distinct words $w$ that were seen following history $h$ exactly $r$ times and finally $\beta(w|\bar{h})$ is the backing-off distribution for generalized history $\bar{h}$. Note that in our case we define the generalized history $\bar{h}$ as follows:

- if $(h,w)$ is a specific trigram $(u,v,w)$ then the generalized history is defined as bigram $(v,w)$
- if$(h,w)$ is a specific bigram $(v,w)$ then the generalized history is defined as unigram $(w)$.

So in the generalized history $\bar{h}$ trigrams are smoothed by bigrams, bigrams are smoothed by unigrams which again may be smoothed by zerograms.

Using the similar smoothing techniques as in previous language model (that is a maximum likelihood criterion completed by leaving-one-out approach) we can obtain the following relations for $b_h$ and $\beta(w|\bar{h})$:

$$b_h = \frac{n_1(h,.)}{n_1(h,.) + 2n_2(h,.) + \sum_{r=3}^{R} \frac{r\, n_r(h,.)\,(1 - b_h)}{r - 1 - b_h}} \quad (5)$$

which is computed as iteration formula for all history $h$ and

$$\beta(w|\bar{h}) = \frac{N(\bar{h},w)}{\sum_{w'} N(\bar{h},w')}, \quad (6)$$

where $N(\bar{h},w)$ is the number of joint events $(\bar{h},w)$. Perplexities of the training and the test corpora computed for this language model completed by all described smoothing techniques are summarized in the Tab.2.

| **Language model with discount factor, backing-off and conditional probability** | | | |
|---|---|---|---|
| | | Training corpus | Test corpus |
| **Unigram** | Entropy ($\hat{H}$) | 6.995 | 7.241 |
| | Perplexity ($PP$) | 127.525 | 151.260 |
| **Bigrams** | Entropy ($\hat{H}$) | 2.419 | 4.192 |
| | Perplexity ($PP$) | 5.345 | 18.279 |
| **Trigrams** | Entropy ($\hat{H}$) | 1.494 | 4.332 |
| | Perplexity ($PP$) | 2.817 | 20.140 |

Tab.2 Statistics for the training and the test corpora of the language model with discount factor, backing-off and conditional probability.

## 6. RECOGNITION RESULTS

In experiments with the speech recognition the 100 utterances (sentences) were randomly chosen from the test corpus. Only bigrams were used in both types of language models. The recognition phase based on the Viterbi decoding algorithm was performed with the support of the HTK toolkit. For an evaluation recognition results we used two standard measures – Correctness (Co) and Accuracy ($Ac$) defined as

$$Co = (N - D - S) / N \times 100\% = H / N \times 100\%, \quad (7)$$

$$Ac = (N - D - S - I) / N \times 100\% = (H - I) / N \times 100\%, \quad (8)$$

where $N$ is the total number of evens (words, sentences) in the reference transcription, $S$ is the number of substitution errors, $D$ is the number of deletion errors, $I$ is the number of insertion errors and $H = (N - D - S)$.

The results of several recognition experiments are summarized in the Tables 3-6. To understand items in the tables we explain established abbreviations. Tests were accomplished with the training sets marked as *T_mmm_XX / M*, where *mmm* means the number of months between acquisition of training and test data, *XX* indicates the number of bits per sample of speech and *M* is the number of mixtures in the Gaussian continuous densities. The case in which no language model was used for comparison is denoted as *Z* (Zerogram). The BJFm means the bigram LM with discount factor and joint probability, the BCFm indicates the bigram LM

with discount factor, backing-off and conditional probability and the BCBn means the language model with absolute discount factor, backing-off and conditional probabilities provided by a simplified recognition network. Mentioned simplification is made through the fact that each word transition probability in the bigram model can be expressed from the back-off weight of the previous word and from the unigram probability of the following word. By itself this fact does not simplify the network. The network simplification is then reached via merging the back-off nodes and sharing a single node. However this is an approximation because there can be generally two transitions for a word pair - back-off transition and explicit bigram probability.

To balance the information between the acoustic and language model the fixed transition penalty $p$ and the grammar scale factor $s$ were incorporated in our system. The log likelihood of word $w_j$ following word $w_i$ with associated observation sequence $\boldsymbol{O}$ for a bigram language model is given by

$$\log(p(\boldsymbol{O}, w_k \mid w_l)) = \log(p(\boldsymbol{O} \mid w_k)) + s\log(p(w_k \mid w_l)) + p\,. \quad (7)$$

| T_000_16 / 1 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **LM** | | Z | BJFm | BCBn | BCFm | | |
| *p* | | 0 | 0 | 5 | 3 | 5 | 5 |
| *s* | | 5 | 5 | 10 | 25 | 10 | 50 |
| **Sentence** | *Co* | 9.00 | 63.00 | 66.00 | 77.00 | 77.00 | 71.00 |
| | *N* | 100 | 100 | 100 | 100 | 100 | 100 |
| | *H* | 9 | 63 | 66 | 77 | 77 | 71 |
| | *S* | 91 | 37 | 34 | 23 | 23 | 29 |
| **Word** | *Co* | 94.44 | 98.12 | 98.75 | 97.85 | 98.03 | 95.97 |
| | *Ac* | 61.92 | 95.88 | 96.06 | 97.04 | 96.77 | 95.16 |
| | *N* | 1116 | 1116 | 1116 | 1116 | 1116 | 1116 |
| | *H* | 1054 | 1095 | 1102 | 1092 | 1094 | 1071 |
| | *D* | 5 | 13 | 3 | 9 | 7 | 17 |
| | *I* | 363 | 25 | 30 | 9 | 14 | 9 |
| | *S* | 57 | 8 | 11 | 15 | 15 | 28 |

Tab.3. Recognition results for test data T_000_16 / 1

| T_000_16 / 8 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **LM** | | Z | BJFm | BCBn | BCFm | | |
| *p* | | 0 | 0 | 5 | 3 | 5 | 5 |
| *s* | | 5 | 5 | 10 | 25 | 10 | 50 |
| **Sentence** | *Co* | 31.00 | 85.00 | 88.00 | 82.00 | 87.00 | 75.00 |
| | *N* | 100 | 100 | 100 | 100 | 100 | 100 |
| | *H* | 31 | 85 | 88 | 82 | 87 | 75 |
| | *S* | 69 | 15 | 12 | 18 | 13 | 25 |
| **Word** | *Co* | 97.04 | 98.21 | 98.84 | 98.12 | 98.84 | 96.24 |
| | *Ac* | 82.53 | 97.85 | 98.48 | 97.58 | 98.48 | 95.79 |
| | *N* | 1116 | 1116 | 1116 | 1116 | 1116 | 1116 |
| | *H* | 1083 | 1096 | 1103 | 1095 | 1103 | 1074 |
| | *D* | 4 | 8 | 3 | 7 | 3 | 22 |
| | *I* | 162 | 4 | 4 | 6 | 4 | 5 |
| | *S* | 29 | 127 | 10 | 14 | 10 | 20 |

Tab.4. Recognition results for test data T_000_16 / 8

| T_000_13 / 8 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **LM** | | Z | BJFm | BCBn | BCFm | | |
| *p* | | 0 | 0 | 5 | 3 | 5 | 5 |
| *s* | | 5 | 5 | 10 | 25 | 10 | 50 |
| **Sentence** | *Co* | 30.00 | 85.00 | 88.00 | 82.00 | 86.00 | 75.00 |
| | *N* | 100 | 100 | 100 | 100 | 100 | 100 |
| | *H* | 30 | 85 | 88 | 82 | 86 | 75 |
| | *S* | 70 | 15 | 12 | 18 | 14 | 25 |
| **Word** | *Co* | 97.04 | 98.21 | 98.84 | 98.12 | 98.75 | 96.24 |
| | *Ac* | 82.35 | 97.85 | 98.48 | 97.58 | 98.39 | 95.79 |
| | *N* | 1116 | 1116 | 1116 | 1116 | 1116 | 1116 |
| | *H* | 1083 | 1096 | 1103 | 1095 | 1102 | 1074 |
| | *D* | 4 | 8 | 3 | 7 | 3 | 22 |
| | *I* | 164 | 4 | 4 | 6 | 4 | 5 |
| | *S* | 29 | 12 | 10 | 14 | 11 | 20 |

Tab.5. Recognition results for test data T_000_13 / 8

| T_024_13 / 8 | | | | | | | |
|---|---|---|---|---|---|---|---|
| **LM** | | Z | BJFm | BCBn | BCFm | | |
| *p* | | 0 | 0 | 5 | 3 | 5 | 5 |
| *s* | | 5 | 5 | 10 | 25 | 10 | 50 |
| **Sentence** | *Co* | 10.00 | 59.00 | 54.00 | 38.00 | 54 | 28 |
| | *N* | 100 | 100 | 100 | 100 | 100 | 100 |
| | *H* | 10 | 59 | 54 | 38 | 54 | 28 |
| | *S* | 90 | 41 | 46 | 62 | 46 | 72 |
| **Word** | *Co* | 91.89 | 95.23 | 95.23 | 93.24 | 95.23 | 88.38 |
| | *Ac* | 73.24 | 94.86 | 94.23 | 91.80 | 94.23 | 87.30 |
| | *N* | 1110 | 1110 | 1110 | 1110 | 1110 | 1110 |
| | *H* | 1020 | 1057 | 1057 | 1035 | 1057 | 981 |
| | *D* | 9 | 15 | 14 | 38 | 14 | 85 |
| | *I* | 207 | 4 | 11 | 16 | 11 | 12 |
| | *S* | 81 | 38 | 39 | 37 | 39 | 44 |

Tab.6. Recognition results for test data T_024_13 / 8

## 7. CONCLUSION

Note that the Czech language belongs to the set of Slavic languages and the results of our task can be beneficial for speech recognition of the others Slavic languages including the Russian as well. Furthermore we intend to make the experimental system more robust, accurate and faster in the future and subsequently we propose to construct continuous Czech speech recognition systems for other problem tasks as well.

**References:**


[1] Jelinek, F.: Statistical Methods for Speech Recognition, MIT Press, Cambridge, 1997.

[2] Nouza, J., Psutka, J., Uhlíř, J.: Phonetic Alphabet for Speech Recognition of Czech. *Radio-engineering*, Vol. 6, pp. 16-20, December 1997

[3] Young, S., Bloothooft, G.: Corpus-Based Methods in Language and Speech Processing. Kluwer Academic Publishers, Dordrecht, 1997.