Optimization of some parameters in the speech-processing module developed for the speaker independent ASR system

J. V. PSUTKA
and
Ludek MÜLLER

Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic

{psutka_j, muller@kky.zcu.cz}

ABSTRACT

This paper deals with looking for an optimum parameterization in automatic speech recognition systems working with the speech transferred over a telephone channel. The performed experiments were supported by a large collection of training data provided from telephone calls of at least one thousand speakers. MFCC and PLP cepstral parameterizations were tested with the aim to find the optimal number of filters and coefficients. Temporal patterns describing several adjacent frames of a given frame were verified in connection with techniques ensuring feature extraction and decorelation of pattern space.

Key words: speech parameterization, optimization of front-end, linear discriminate analysis, pattern space normalization

1. INTRODUCTION

The robust parameterization of speech used in automatic speech recognition (ASR) systems is still a great objective of many research teams. In tasks solved in many laboratories the software toolkit HTK [1] is often used and we can see that for processing of speech of various quality unsuitable default settings of the HTK's front-end are frequently applied. Such settings do not sometimes agree with the theory of human hearing, which is incorporated in the two most frequent auditory-based parameterization techniques – MFCC and PLP. Simultaneously we can observe under-dimensioned and/or over-dimensioned front-ends, which respect neither the application tasks nor the real working conditions as time and memory demand. The present paper addresses this problem and aims to contribute to the discussion concerning the selection of an optimum number of filters distributed in the frequency axis as well as the number of enumerated coefficients. In contrast with [2], [3], all experiments using MFCC and PLP parameterizations were performed with continuous speech of telephone quality employing voices of a large group of one thousand speakers for building HMMs.

The second aim of this paper is to compare several techniques of feature extraction and pattern space decorelation. In the performed experiments we substituted a conventional feature vector describing speech in one frame by longer temporal vector composed of vectors of features of several adjacent frames. For feature extraction and pattern space decorelation we tested such techniques as the linear discriminant analysis (LDA), principal component analysis applied on the between-class scatter matrix, normalization of pattern space, and cosine transform. Owing to the extremely time-consuming computation burden, all tests in this case were carried out using training corpus of only one hundred speakers. The quality of settings of all front-ends was evaluated by the accuracy (*Acc*) defined as

$$Acc = (N - D - S - I) / N \times 100\% ,$$

where N is the total number of words in the reference transcription, S is the number of substitution errors, D is the number of deletion and I the number of insertion errors.

All experiments deal with telephone-based speaker independent continuous speech recognition. Speech data was taken from the Czech telephone corpus. This corpus consists of read speech transmitted over a telephone channel. More than one thousand speakers were asked to read various sets of 40 sentences. The digitization of an input analog telephone signal was provided by a telephone interface board DIALOGIC D/21D at 8 kHz sample rate and converted to the mu-law 8-bit resolution. Speech data was annotated using special annotation software Transcriber 1.4.1, which is freely available from the web site http://www.ldc.upenn.edu/. The corpus was then randomly divided so that recordings of one thousand speakers created the training part and the remaining part (one hundred of different speakers) formed the test part of the corpus. From this training part 100 sentences were randomly selected to create test data for all our experiments. The lexicon of the task contained 475 different words. In all experiments a language model based on a zero-gram was applied.

2. AUDITORY-BASED FRONT-END

The MFCC and PLP-based front-ends attempt to model the auditory processing up to activation of the inner hair cells by the basilar membrane vibrations. This simulation is usually performed through a selective model which is implemented by a filter bank whose center frequencies are spaced along the frequency axis which satisfies the critical-band scale and whose particular filter widths correspond to the theory of critical bandwidths [4]. The most common critical-band scales are the mel-scale and the bark-scale in which the filters are distributed along the frequency axis approximately linear up to about 1000 Hz and logarithmic above 1000 Hz. The number of critical bands depends on the whole frequency bandwidth and/or the sampling frequency F_s. For the telephone frequency band with F_s=8 kHz approximately from 15 to 17 critical bands are recommended.

2.1 Experiments with MFCC parameterization

The computational algorithm of the MFCC parameterization is realized by the bank of symmetric overlapping triangular filters spaced linearly in a mel-frequency axis, according to auditory perceptual considerations. The spacing as well as the bandwidth of the particular filters is determined by a constant mel-frequency interval. In our case the spacing was approximately 145 mels and the width of the triangle was 290 mels. So, for telephone frequency band (0 - 2146 mels) with the sampling frequency F_s=8 kHz we obtained (using critical-band concept) about 15 filters distributed up to the Nyquist frequency. The MFCC parameterization was accomplished by the computation of mel-cepstral coefficients $c(1)$, ... , $c(M)$. In practice these coefficients are usually obtained by applying an inverse DCT (Discrete Cosine Transform) to the log-energy of the filter bank outputs in order to decorelate the parameter (pattern) space. The set of cepstral coefficients is usually complemented by the coefficient $c(0)$, which approximates the average log-energy of the signal (this coefficient is often replaced by the energy computed directly from the signal). In this set of experiments our front-end is, in addition, complemented by the computation of time derivatives (delta plus delta-delta) of the corresponding static features. This means that our front-end provides feature vectors with the dimension of $3M$.

In this case the experiments were realized for an increasing number of filters from 4 to 21 (in the step 2) and for an increasing number of coefficients from 4×3=12 to 18×3=54. The results of all experiments are given in Table 1. Since the greatest accuracy (*Acc*) oscillates between 84 and 86%, we indicated in Table 1, for a clearer survey better survey, by gray color the items with recognition accuracy higher than 84%.

# of coef. **# of filters**	**12**	**15**	**18**	**21**	**24**	**30**	**36**	**42**	**48**	**54**
7	78.6	81.4	83.7	**84.1**	-	-	-	-	-	-
9	77.1	81.6	83.8	83.7	**84.9**	-	-	-	-	-
11	78.1	81.1	83.3	83.5	83.8	**85.0**	-	-	-	-
13	75.9	80.0	82.3	83.2	83.8	**85.6**	**85.9**	-	-	-
15	76.3	80.7	81.6	82.6	**84.1**	**85.0**	**84.6**	**85.4**	-	-
17	76.6	79.8	82.6	82.7	83.8	**84.6**	**84.7**	**85.5**	**84.6**	-
19	76.1	79.2	81.3	82.3	**84.2**	**84.1**	**85.0**	**84.6**	**84.4**	**84.5**
21	75.6	78.3	81.0	83.0	83.7	**84.5**	**85.1**	**85.1**	**85.3**	**84.8**

Table 1. Recognition results for various numbers of filters and parameters in MFCC parameterization.

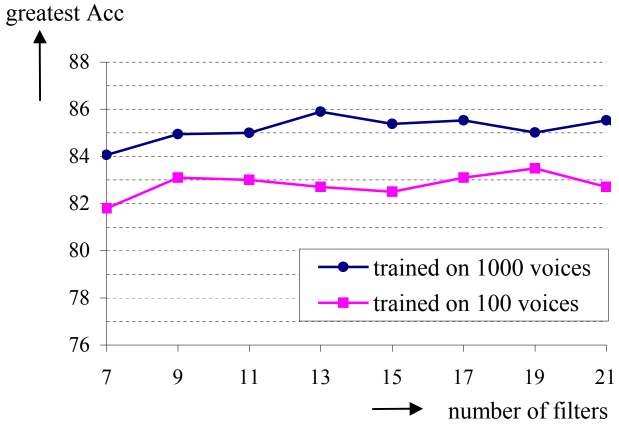

Figure 1. Dependence of the greatest Acc on the number of filters used in the MFCC parameterization.

The dependence of the greatest accuracy for a given number of filters is depicted in Figure 1. To compare results of these experiments (based on training corpus of 1000 different speakers) with those obtained using voices of only 100 speakers (these experiments were reported in [3]) we complemented Figure 1 by a curve which expresses corresponding dependence (see Figure 1). The dependence of the greatest results of accuracy for a given number of coefficients is depicted simultaneously with the results got for 100 speakers [3] in Figure 2.

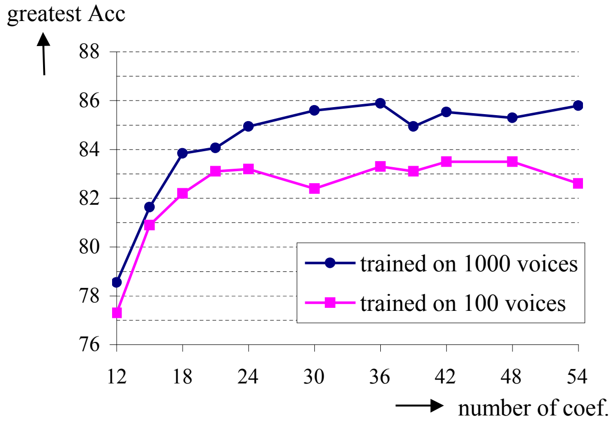

Figure 2. Dependence of the greatest Acc on the number of coefficients used in the MFCC parameterization.

2.2 Experiments with PLP parameterization

The front-end used in this case was based on the PLP parameterization described in [4]. For the transformation of a power speech spectrum to a corresponding auditory spectrum the PLP combines three components from psychophysics of hearing: the critical-band spectral selectivity, the equal-loudness curve and the intensity-loudness power law. To carry out this process we have to perform following steps:

- Computation of short-term speech spectrum
- Nonlinear frequency transformation and critical-band spectral resolution. The modeling of these phenomena is performed in the PLP either by the nonlinear transformation of frequencies from the Herz into the Bark scale and by the construction of masking curves that simulate critical-band of hearing and are modeled by the band-pass filters. The centers of filters are spaced in the Bark domain linearly with the step approximately 1 Bark. As the speech signal covers the range from 0 to 4 kHz the corresponding range in the Bark scale was 0 ÷ 15.57 Bark and we used M=17 filters spaced linearly with the step of 0.973 Bark. The 0^{th} filter had a center in the value of 0 Bark, the last $(M\text{-}1)^{st}$ filter was centered in the value of 15.57 Bark.
- Critical-bands adjustment to the curves of equal-loudness
- Weighted spectral summation of power spectrum samples
- Enforcing the intensity-loudness power-law
- All-pole spectrum approximation
- Transformation of the PLP-coefficients to the PLP-cepstral representation. The PLP-cepstral coefficients $c(1)$, ... , $c(Q)$ are computed by the standard approach from the Q PLP predictive coefficients. For the final acoustic modeling we extended the original PLP-cepstral representation with derived delta and delta-delta features. In fact, the dimension of the pattern space in which the acoustic models of triphones were built, increased from Q to 3 Q.

A number of experiments was also performed for this parameterization in which the bank of filters was increased from 4 to 21 (in the step 2) and the number of coefficients was enumerated from 4×3=12 to 18×3=54. The results achieved in recognition experiments are given in Table 2.

# of coef. **# of filters**	**12**	**15**	**18**	**21**	**24**	**30**	**36**	**42**	**48**	**54**
7	78.8	79.7	78.2	79.1	-	-	-	-	-	-
9	76.2	81.5	83.8	**85.1**	81.3	-	-	-	-	-
11	75.8	80.7	**84.6**	**84.9**	**84.2**	**84.6**	-	-	-	-
13	77.0	80.1	**85.5**	**84.1**	**84.8**	**85.9**	**84.2**	-	-	-
15	78.5	80.9	**84.8**	**84.4**	**85.6**	**85.8**	**85.9**	**85.7**	-	-
17	75.8	80.5	83.7	**85.1**	**84.4**	**85.0**	**85.0**	**85.0**	**84.2**	-
19	76.9	80.2	**84.6**	**84.7**	**84.3**	**84.6**	**86.6**	**85.1**	**86.0**	**84.4**
21	76.1	79.8	83.9	**85.0**	83.9	**85.5**	**86.5**	**86.0**	**85.7**	**85.1**

Table 2. Recognition results for various numbers of filters and parameters in PLP cepstral parameterization

Dependence of the greatest Acc on the number of filters and on the number of coefficients are given in Figure 3 and 4. Similarly to the MFCC parameterization (Figure 1 and 2), two curves were obtained for the ASR system trained on 100 and 1000 speakers respectively.

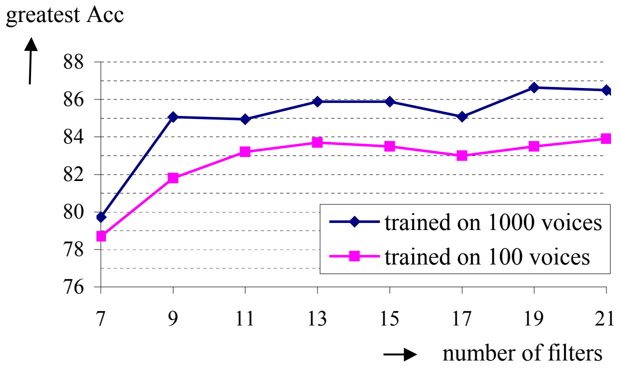

Figure 3. Dependence of the greates Acc on the number of filters used in the PLP parameterization.

2.3 Comparison of both parameterization techniques

In all experiments the HMMs based on triphones were used. The number of Gaussians and states for individual types of parameterizations moves from 30k Gaussians and 3k6 states for lower number of filters and parameters up to 50k Gaussians and 6k2 states for higher number of filters and parameters (similarly for both MFCC and PLP), see Table 1 and 2. Looking at Tables 1 and 2 we can see that there are areas of coefficients and filters with high and relatively stable recognition accuracy. For the MFCC and PLP parameterizations the greatest *Acc* (*Acc* higher than 84%) was achieved for 13 ÷ 19 filters and 10×3 ÷ 16×3 coefficients (MFCC) and for 11 ÷ 19 filters and 7×3 ÷ 16×3 coefficients (PLP) respectively. This is a slightly higher number of filters, especially for the MFCC parameterization, than was reported in [3] for tasks with training set of 100 people. It is evident that these "optimal" settings are approaching the recommended number of filters and satisfy the theory of critical bandwidths to a greater extent. However, the area of suitable settings for the PLP is much larger and the recognition results achieved are higher by 0.5 to 1% on average in comparison with the MFCC. Also the number of coefficients for the PLP does not have to be so high, which could bring computation savings (useful for building real time ASR systems).

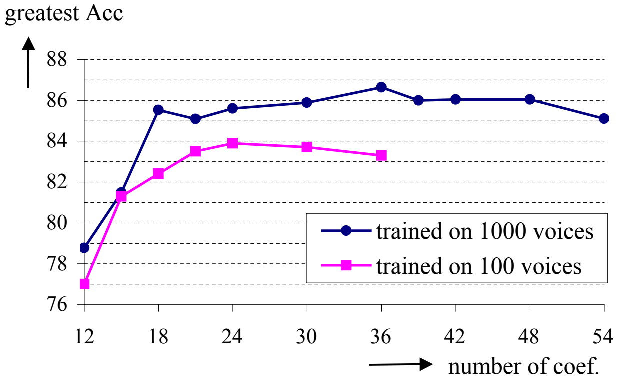

Figure 4. Dependence of the greatest Acc on the number of coefficients used in the PLP parameterization.

The results depicted in Figures 1÷4 bring expected yet interesting information. Front-ends built on HMMs trained with voices of one thousand people run better by 2% on average (given by *Acc*) than those trained on the group of one hundred people. Analyzing the results of the recognition experiments, we can mention that this improvement was achieved by better covering several test voices, which were incorrectly matched by the "old" models trained on the corpus built from speech of 100 people.

We would also like to address another interesting problem which deals with modeling speech by monophone- and/or triphone-based HMMs. Our recent experiments with monophone-based HMMs using the set of 100 training voices showed, that there are no differences in recognition accuracy between triphone-based HMMs with 8 mixtures and monophone-based HMMs using at least 50 or 60 mixtures. However, our new results, obtained with new triphone-based HMMs with 8 mixtures trained on the corpus of 1000 speakers (see Tables 1 and 2), exceeded the monophone structure trained on the same corpus by 2%. Figure 5 shows the dependency of the recognition accuracy on the number of mixtures in monophone-based HMMs working with the MFCC parameterization (15×3=45 parameters). Apart from this slight degradation of the *Acc* we have to notice that these relatively outstanding results were achieved with the set of monophone-

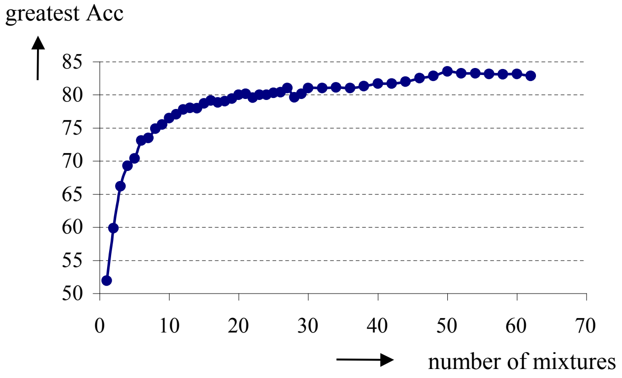

Figure 5. Dependence of the Acc on the number of mixtures for monophone-based HMMs with MFCC parameterization.

based HMMs, which were surprisingly described only by 7k Gaussians and 135 states! Further improvements of the *Acc* could be obtained by the cepstral mean normalization, amplitude normalization, linear discriminant analysis etc. The first two mentioned techniques were tested in [3]. Our next section is devoted to the several feature extraction methods in the context of so called temporal patterns.

3. TEMPORAL PATTERNS AND FEATURE EXTRACTION

In this section three series of experiments with temporal patterns [5] are described. A technique of temporal patterns substitutes each conventional feature vector by longer temporal pattern consisting of feature vectors of several adjacent frames. Our experiments were performed using parameters of

a) given frame (the conventional case) $\boldsymbol{v}(k) = \boldsymbol{x}(k)$,

b) two adjacent frames $\boldsymbol{v}(k) = [\boldsymbol{x}(k\text{-}1), \boldsymbol{x}(k), \boldsymbol{x}(k{+}1)]^{\mathrm{T}}$,

c) four adjacent frames (two from each side) $\boldsymbol{v}(k) = [\boldsymbol{x}(k\text{-}2), \boldsymbol{x}(k\text{-}1), \boldsymbol{x}(k), \boldsymbol{x}(k{+}1), \boldsymbol{x}(k{+}2)]^{\mathrm{T}}$.

As features for basic description of patterns, the log-energies of 15 output filters of the MFCC parameterization were enumerated (neither delta nor delta-delta features were used), see paragraph 2.1. These sets of features were subjected to further processing which aims both at extraction of smaller subsets of informative features and decorelation of pattern space. Owing to extremely time-consuming computation burdens all tests in this case were done using training corpus of only one hundred speakers. In addition, to ensure the same conditions for all tests we used in all experiments monophone-based HMMs with 8 mixtures (only 1k Gaussians and 126 states). The set of 100 test sentences stayed the same as in the last experiments. During feature extraction and pattern space decorelation experiments we tested such techniques as a linear discriminant analysis (LDA), principal component analysis applied to the between-class scatter matrix (PCAc), normalization of pattern space (NPS), and discrete cosine transform (DCT). Now we briefly explain individual techniques.

The goal of all mentioned techniques is to find a transformation matrix $\boldsymbol{W}^{\mathrm{T}}$, which transforms the given pattern space to the space of lower dimension and/or to the space with decorelated features. In order to carry out discriminant analysis it is necessary to determine individual phoneme classes and use phonetically labeled speech corpus. For these purposes the speech corpus of all 100 training speakers was phonetically labeled using 42 Czech phone units.

3.1 Linear discriminant analysis (LDA)

For the *c*-class problem the linear discriminant analysis involves $c-1$ discriminant functions. Thus, the projection is from the original *n*-dimensional feature space to a $m{=}(c-1)$-dimensional space. What we seek now is a transformation matrix $\boldsymbol{W}^{\mathrm{T}}$, which in some sense maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix. In our case the within-class scatter matrix $\boldsymbol{S}_{\mathrm{W}}$ is defined as

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{i=1}^{c} P_i \boldsymbol{S}_i ,$$

where P_i is the a priori probability of the class i and $\boldsymbol{S}_i$ is the covariance matrix computed from the feature vectors belonging to the phoneme class i . $\boldsymbol{S}_i$ can be expressed as

$$\boldsymbol{S}_i = \mathrm{E}\ \{(\boldsymbol{v} - \boldsymbol{\mu}_i)(\boldsymbol{v} - \boldsymbol{\mu}_i)^{\mathrm{T}}\} ,$$

where $\boldsymbol{\mu}_i$ is the mean vector of the class i . Between-class scatter matrix is defined as

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{i=1}^{c} P_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}} ,$$

where $\boldsymbol{\mu}$ is the global mean vector

$$\boldsymbol{\mu} = \sum_{i=1}^{c} P_i \boldsymbol{\mu}_i .$$

It is well known [6] that the rows of an “optimal” transformation matrix $\boldsymbol{W}^{\mathrm{T}}$ are the generalized eigenvectors that correspond to the largest eigenvalues of the matrix ($\boldsymbol{S}_{\mathrm{W}}^{-1}\,\boldsymbol{S}_{\mathrm{B}}$). The input vector $\boldsymbol{v}$ of dimension n from the original pattern space can be then transformed to the “optimum” space of dimension $m{=}c-1$ (there are only $c-1$ nonzero eigenvalues) in accordance with the equation

$$\boldsymbol{y} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{v}$$

Let us notice that if the dimension n of the original feature space is lower than $m{=}c-1$, then a dimension of a new pattern space stays usually the same after the transformation (equal to n).

3.2 Discrete Cosine Transform (DCT)

Discrete cosine transform is used in order to decorelate features in the pattern space. This is the standard method applied to the log-energies of output filters (LogEF) during the MFCC parameterization, see paragraph 2.1. DFT is defined as

$$y_j = \sum_{i=1}^{n} v_i \cos\left[\frac{\pi j}{n}(i - 0.5)\right], \qquad \text{for } j = 0, 1, \ldots, n$$

where v_i is *i*-th coordinate of the input vector $\boldsymbol{v}$ and y_j is *j*-th coordinate of the corresponding output vector $\boldsymbol{y}$. This transformation can be easily expressed in the matrix notation.

3.3 Normalization of pattern space (NPS)

Normalization of pattern space is usually applied in order to decorelate features in the space. The transformation matrix $\boldsymbol{G}^{\mathrm{T}}$ ensuring this transformation should satisfy the relation

$$\boldsymbol{G}^{\mathrm{T}}\,\boldsymbol{S}_{\mathrm{W}}\,\boldsymbol{G} = \mathbf{1} ,$$

where $\mathbf{1}$ is the identity matrix and $\boldsymbol{S}_{\mathrm{W}}$ the within-class scatter matrix. The solution of this equation is

$$\boldsymbol{G}^{\mathrm{T}} = \boldsymbol{\Lambda}^{-1/2}\,\boldsymbol{C}^{\mathrm{T}} ,$$

where $\boldsymbol{\Lambda}$ is the diagonal n by n matrix of eigenvalues and $\boldsymbol{C}^{\mathrm{T}}$ is n by n matrix with rows created by the orthonormal eigenvectors that correspond to the eigenvalues of the within-class scatter matrix $\boldsymbol{S}_{\mathrm{W}}$. This transformation does not change the dimensionality of a pattern space.

3.4 Principal Component Analysis (PCAc)

In this case the transform matrix was formed from m eigenvectors corresponding to the m largest eigenvalues of the between-class scatter matrix $\boldsymbol{S}_{\mathbf{B}}$. There are maximum $m{=}c-1$ nonzero eigenvalues. This means that the transformation can be done to the space with the maximum dimension $c-1$.

In Table 3 you can find the results of related experiments. To explain conditions of individual experiments we proposed the following notation: (*LogEF_n*) depicts log-energies of *n* output filters, (*diag*) means that final covariance matrices were diagonalized – coefficients out of the main diagonal were set to zero, (*full*) indicates that during experiments full covariance matrices were used, (*LDA_n*) means the linear discriminant analysis with output vectors of dimension *n*, (*NPS*) specifies the normalization of the pattern space, (*PCAc*) means the principal component analysis applied to the between-class scatter matrix.

	Accuracy		
	***n*=15**	***n*=45**	***n*=75**
LogEF_n → diag	39.7	40.3	42.1
LogEF_n → DCT → diag	45.3	53.4	59.2
LogEF_n → full	47.3	73.7	74.3
	n → m		
	15→15	**45→41**	**75→41**
LogEF_n → LDA_m → diag	53.1	67.5	66.7
LogEF_n → LDA_m → DCT → diag	44.6	49.3	43.4
LogEF_n → NPS+PCAc_m → diag	46.2	63.9	66.3
LogEF_n → LDA_m → NPS+PCAc → diag	45.6	64.7	66.7
LogEF_n → LDA_m → full	61.9	70.9	75.5

Table 3. Results of experiments with several feature extraction and decorelation techniques applied to temporal patterns.

For a given number of 8 mixtures and a monophone-based HMMs structure, we can compare individual techniques of the feature selection and pattern space decorelation:

- decorelate techniques based on the DCT give distinctly worse results than those obtained in other techniques; compare (*LogEF_n→DCT→diag*) versus (*LogEF_n→LDA_m→diag*) or (*LogEF_n→NPS+PCAc_m→diag*) or (*LogEF_n→LDA_m→NPS+PCAc →diag*)
- LDA slightly improves recognition accuracy, see (*LogEF_n →LDA_m→diag*) versus (*LogEF_n→ DCT→diag*) or (*LogEF_n →NPS+PCAc_m→diag*)
- full covariance matrices distinctly exceed diagonal matrices (for the same number of mixtures)
- temporal patterns used in our experiments did not bring better recognition results in comparison with delta+delta-delta representation (see results depicted in Figure 5).

4. CONCLUSION

The results achieved in this paper confirmed, conclusions described in [3], but on a substantially larger portion of speech data. Both parameterizations (MFCC and PLP) are comparable but the PLP one provides slightly better and robust (stable for a larger number of coefficients and filters) results. Feature extraction techniques and pattern space decorelate methods tested in our experiments will also have to be estimated in the future on the triphone-based HMMs or monophone-based HMMs with a larger number of mixtures.

5. ACKNOWLEDGEMETS

This work was funded by the Ministry of Education of the Czech Republic, project No. MSM234200004 and the Grant Agency of the Czech Republic, project No. 102/02/0124.

6. REFERENCES

[1] Young, S. et al.: The HTK Book. User's Manual. -In: http://htk.eng.cam.ac.uk/, 1997.

[2] Müller, L., Psutka, J.V.: Selection of an Optimum Speech Parameterization for Continuous Speech Recognition System Using a Telephone Channel. –In: Proc. of SCI'2001, Orlando, U.S.A., 2001, pp. 542-545.

[3] Psutka, J., Müller, L., Psutka, J.V.: Comparison of MFCC and PLP Parameterization in the Speaker Independent Continuous Speech Recognition Task. –In: Proc. of EUROSPEECH'2001, Denmark, Aalborg, 2001, pp. 1813-1816.

[4] Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. –In J. Acoust. Soc. Am. 87, (1990), pp.1738-1752.

[5] Hermansky, H., Sarma, S.: TRAPS – Classifiers of Temporal Patterns. –In: Proc. of ICSLP'98, Sydney, 1998.

[6] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, Inc., New York, 2001.

[7] Veth, J., Boves, L.: Channel normalization techniques for automatic speech recognition over the telephone. –In: Speech Communication, 25 (1998), pp.149-164.

[8] Hermansky, H., Sharma, S., Ellis, D., Jain, P., Kajarekar, S.: Feature extraction using non-linear transformation for robust speech recognition on the Aurora database. -In: Proc. ICASSP 2000, Turkey, 2000.