Comparison of various feature decorrelation techniques in automatic speech recognition

J.V. Psutka, Luděk Müller

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

ABSTRACT

The design of an optimum front-end module for an automatic speech recognition system is still a great effort of many research teams all over the world. Prepared paper wants to contribute partly to these discussions. It is especially aimed at feature decorrelation techniques based on Maximum Linear Likelihood Transform (MLLT) applied at a different level of matrix clustering. Also the comparison of the MLLT with other decorrelation techniques will be discussed.

Key words: speech parameterization, decorrelation techniques, Maximum Linear Likelihood Transform

1. INTRODUCTION

The effort to select suitable features containing relevant information for speech recognition brought researchers after thorough analysis of process of human hearing to parameterization techniques based on mel frequency cepstral (MFCC) or perceptual linear prediction (PLP) cepstral coefficients. An experience with speech recognition showed that it is beneficial to use also delta and delta-delta coefficients which decrease the word error rate (WER) but simultaneously increase the dimension of feature space (usually 3 times). Even though the original set of features of the MFCC or the PLP parameterizations is more or less correlated then after addition of delta and delta-delta features the information redundancy of elements in feature vectors increases. Moreover, a large amount of (correlated) features make the training and recognition process more difficult.

Let us remind that speech in LVCSR systems is predominantly modeled by hidden Markov models (HMMs). As fundamental attributes of this concept can be considered output probabilities tied to each state of model and modeled by multidimensional Gaussian distributions (simply by Gaussians) or more exactly by mixtures of Gaussians. An application of mixtures of Gaussians for output probability modeling results from an effort both to catch the possible non-Gaussian nature of density functions which are associated with a particular state and to model mutual correlation of elements in feature vectors (especially in case of Gaussian distributions which are determined by only diagonal covariance matrices). Because all output probabilities should be computed for each incoming feature vector it is useful notably for real-time applications to reduce often huge amount of computations which increase with a size of the dimension of feature space and also with the number of Gaussians.

To reduce computation burdens associated with evaluating output probabilities we can apply some of following techniques:

- To execute decorrelation of feature vectors and to use rather diagonal then full covariance matrices for modeling of output probabilities. For these purposes usually some orthogonal transforms based on DCT (Discrete Cosine Transform) or NPS (Normalization of Pattern Space) are applied [1].
- To reduce a dimension of pattern space using projection of feature vectors from the original space to the space with lower dimension. A typical approach is based on PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis) or HLDA (Heteroscedastic LDA).
- To pass from the triphone- to the monophone-based structure of models where an influence of suppressed dependencies among features is alleviated mainly by enhancing number of Gaussians in individual states of monophone models. LVCSR systems with triphone-based structure work typically with 30 to 100 thousand of Gaussians, whereas systems working with monophone-based structure uses from 5 to 10 thousand Gaussians [2], [3].

There are of course many other clever approaches, which speed up computations or choose only relevant states with associated Gaussians for evaluations. Generally, it is possible to say, that both the pass from the triphone- to the monophone-based concept and also various transformation techniques on one hand decrease the number of computations but on the other hand they cause increasing the word error rate (WER). Moreover, in case of transformations applied in a level of feature vectors, there is usually unfeasible to find the only transformation which could decorrelate all elements of feature vectors of all states.

Recently the new approaches have been designed, which alleviate above mentioned decrease of recognition accuracy (Acc) simultaneously with preserving high computation efficiency. These techniques is based on Maximum Linear Likelihood Transform (MLLT), which supposes that one or more transformation matrices will be tied with states (more exactly tied with covariance matrices belonging to individual states) of hidden Markov models. Transformations are designed according to a group of Gaussians which should be decorrelated. If the transformation is built separately for each individual mixture (Gaussian), then we could obtain the system which is identical with the system using full covariance matrices (of course we wouldn't get any computation savings in this case). Therefore it is reasonable to find a suitable tradeoff between selected groups of Gaussians that should be decorrelated by individual transformation matrices and the recognition accuracy.

2. EXPERIMENTAL CONDITIONS

All experiments were performed with the high-quality speech corpus. This corpus is a read-speech database consisting of the speech of 100 speakers. Each speaker read a same portion of 40 sentences. The database of text prompts from which the sentences were selected was obtained in an electronic form from the web pages of Czech newspaper publishers. A special consideration was given to the selection of the mentioned set of 40 sentences, since they provide a representative distribution of

the more frequent triphone sequences (reflecting their relative occurrences in natural speech). The corpus was recorded in the office where only the speaker was present. Recordings were performed using the notebook IBM TP 760 ED owing to a very silent operation of this computer (it hasn't any fan). Each sentence was recorded simultaneously by two microphones. The close-talking microphone (Sennheisser HMD410-6) yielded utterances of a high-quality, the desk microphone (Sennheisser ME65) recorded utterances including common office noise. The prompting/recording sessions yielded totally about 80 hours of speech, all of which was digitized into pairs of single-channel files at 44.1 kHz with 16-bit resolution.

Test set for high-quality speech was selected from utterances of speakers who were not included in training set. The test set consists of 100 sentences randomly selected from utterances of 4 different speakers (4 speakers x 25 sentences) The vocabulary in all our test tasks contained 475 different words. Since several words had multiple different phonetic transcriptions the final vocabulary consisted of 528 items. There were no OOV words. In all recognition experiments a language model based on zerograms was applied. It means that each word in the vocabulary is equally probable as a next word in the recognized utterance. For that reason the perplexity of the task was 528.

3. FEATURE SPACE TRANSFORMATION

During feature extraction and pattern space decorrelation experiments we tested such techniques as discrete cosine transform (DCT) and linear discriminant analysis (LDA). The goal of mentioned techniques is to find a transformation, which transforms given pattern space to the space with decorrelated features and/or to the space of lower dimension.

Discrete Cosine Transform (DCT)
Discrete cosine transform is used in order to decorrelate features in the pattern space. This is the standard method applied to the log-energies of output filters (LogEF) during the MFCC parameterization. DFT is defined as

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

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 $\boldsymbol{y}$.

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

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{i=1}^{c} P_i \boldsymbol{S}_i , \quad (2)$$

where P_i is the a priory 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

$$\boldsymbol{S}_i = \mathrm{E}\{(\boldsymbol{v} - \mu_i)(\boldsymbol{v} - \mu_i)^{\mathrm{T}}\} , \quad (3)$$

where μ_i is the mean vector of the class i . Between-class scatter matrix is defined as

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{i=1}^{c} P_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}} , \quad (4)$$

where μ is the global mean vector

$$\boldsymbol{\mu} = \sum_{i=1}^{c} P_i \boldsymbol{\mu}_i . \quad (5)$$

It is well known 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

$$\boldsymbol{y} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{v} \quad (6)$$

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 after the transformation usually the same (equal to n).

Recognition experiments
In recognition experiments we used log-energies of 27 output filters of the MFCC parameterization as features for a basic description of speech patterns. For comparison of individual transformation techniques we used HMMs based on monophone structure. The goal was to investigate an influence of a number of mixtures (Gaussians) assigned to individual states of HMMs to the recognition accuracy. In Table 1 and 2 you can find results of many experiments in case of diagonal and/or full covariance matrices belonging to individual Gaussian mixtures. The second column shows results for feature vectors built as 12 DCT coefficients + 12 delta + 12 delta-delta. The third column brings results after additional processing of above mentioned 36 dimensoin feature vector using LDA transform to the space with dimension of 26. And finally the fourth column gives results of LDA transform of feature vectors composed of log-energies of 27 output filters of the MFCC parameterization + corresponding 27 delta + 27 delta-delta to the dimension of 36.

Number of mixtures	Diagonal Covariance Matrices		
	DCT(36)	**DCT**(36)→ **LDA**(26)	**LDA**(81→36)
1	73.31	49.21	24.18
4	86.44	66.57	49.78
8	89.53	71.45	62.84
12	90.39	74.68	64.79
16	91.18	76.97	68.79
20	92.61	78.05	68.94

Table 1. Recognition accuracy for increasing number of mixtures modeled by diagonal covariance matrices.

Let us notice that for **DCT**(36)→ **LDA**(26) in Table 2 it was able to compute only HMMs with 8 Gaussians. It was probably owing to ill-conditioned matrices during estimation of parameters of covariance matrices (probably due to a large space and only few examples for some phonemes).

Number of mixtures	Full Covariance Matrices		
	DCT(36)	**DCT**(36)→ **LDA**(26)	**LDA**(81→36)
1	80.27	81.92	84.92
4	90.17	89.67	90.24
8	92.47	91.97	92.97
12	92.55	92.15	-
16	92,62	92.27	-
20	93,11	92.47	-

Table 2. Recognition accuracy for increasing number of mixtures modeled by full covariance matrices.

4. MAXIMUM LINEAR LIKELIHOOD TRANSFORMATION

Current speech recognition systems use HMMs with continuous parameters which are represented for each state by a Gaussian Mixture Model (GMM). A standard GMM with parameters given by $\boldsymbol{\Theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^m$ is of the form

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{m} \pi_j N(\boldsymbol{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{7}$$

where m is the number of components, π_j is the j-th component weight satisfying requirements: $\pi_j \in R,\ \pi_j \geq 0, \Sigma_{j=1}^m \pi_j = 1$, $\boldsymbol{\mu}_j$ is the component mean $\boldsymbol{\mu}_j \in R^d$ and where $\boldsymbol{\Sigma}_j$ is a square covariance matrix of rank d. As was mentioned in Section 1, almost all currently used systems work with covariance matrices of diagonal form. This approach has in comparison with the full covariance model an evident advantage especially owing to lower computation and storage burdens and also due to robust parameter estimation. But this can be done only with assumption of elements of the feature vector which are independent. MLLT introduces a new form of a covariance matrix, which allows sharing a few full covariance matrices over many distributions. Instead of having a distinct covariance matrix for every component in the recognizer, each covariance matrix consists of two elements: a non-singular linear transformation matrix $\boldsymbol{A}$ shared over a set of components and the diagonal elements in the matrix $\boldsymbol{\Lambda}_j$. Inverse covariance (precision) matrix $\boldsymbol{\Sigma}_j^{-1}$ is of the form

$$\boldsymbol{\Sigma}_j^{-1} \approx \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_j \boldsymbol{A} = \sum_{k=1}^{d} \lambda_k^j \boldsymbol{a}_k \boldsymbol{a}_k^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{\Lambda}_j$ is a diagonal matrix with entries $\boldsymbol{\Lambda}_j = diag(\lambda_k^j)$ and where a_k^{T} is the k^{th} row of the transformation matrix $\boldsymbol{A}$.

In contrast to other methods which follow similar goals, this technique fits within the standard maximum-likelihood criterion used for training HMM's. Parameters of the MLLT model $\Theta = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j, \boldsymbol{A}\}_{j=1}^m$ are estimated using a generalized expectation-maximization (EM) algorithm. A function that should be optimized with respect to π_j, $\boldsymbol{\mu}_j$, $\boldsymbol{\Lambda}_j$, $\boldsymbol{A}$ is as follows

$$\Phi(\Theta, \hat{\Theta}) = \sum_{j=1}^{m} \pi_j \log \pi_j + \pi_j / 2 \ \log \det\left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_j \boldsymbol{A}\right) - \\ - \pi_j / 2 \ trace\left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Lambda}_j \boldsymbol{A}\, \boldsymbol{W}_j\right), \tag{9}$$

where

$$\boldsymbol{W}_j = \sum_{i=1}^{N} \gamma_{ij} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}} \Big/ \sum_{i=1}^{N} \gamma_{ij}, \tag{10}$$

where $\{\boldsymbol{x}_i\}_{i=1}^N$ is a given training set, γ_{ij} is the aposteriori probability of j-th component of GMM and given vector $\boldsymbol{x}_i$. Optimizing (9) is unfortunately nontrivial especially because of $\boldsymbol{\Lambda}_j$ and $\boldsymbol{A}$ (optimizing π_j, $\boldsymbol{\mu}_j$ is done similarly to the classical HMM's). The idea is to cycle through alternating estimations of $\boldsymbol{A}$ and $\boldsymbol{\Lambda}_j$ while keeping one of them fixed. The estimation of $\boldsymbol{\Lambda}_j$ for fixed $\boldsymbol{A}$ is done by

$$\boldsymbol{\Lambda}_j = diag\left(\boldsymbol{A}\ \boldsymbol{W}_j\ \boldsymbol{A}^{\mathrm{T}}\right). \tag{11}$$

The estimation of $\boldsymbol{A}$ for fixed $\boldsymbol{\Lambda}_j$ has an elegant iterative algorithm

$$\hat{\boldsymbol{a}}_k = \boldsymbol{c}_k\, \boldsymbol{G}_i^{-1} \sqrt{\left(\frac{\sum_{j=1}^{m} \sum_{i=1}^{N} \gamma_{ij}}{\boldsymbol{c}_i\, \boldsymbol{G}_i^{-1} \boldsymbol{c}_i^{\mathrm{T}}}\right)}, \tag{12}$$

where $\boldsymbol{a}_k^{\mathrm{T}}$ is the k-th row of the transformation matrix $\boldsymbol{A}$, $\boldsymbol{c}_k$ is the k-th row vector of cofactors of the current estimate of $\boldsymbol{A}$ and

$$\boldsymbol{G}_i = \sum_{j=1}^{m} 1/\lambda_{i,j}^2\ \boldsymbol{W}_j \sum_{i=1}^{N} \gamma_{ij} \quad . \tag{13}$$

The mathematical theory of this algorithm [5] can be supported by following steps:

1) Initialization of transformation matrix $\boldsymbol{A}$ with an identity matrix.
2) Estimation of means and component weights is made using standard HMM formulas.
3) Estimation of the set of component specific $\boldsymbol{\Lambda}_j$ (11) using the current estimate of the transformation matrix $\boldsymbol{A}$.
4) Estimation of the transformation matrix $\boldsymbol{A}$ using current set of $\{\boldsymbol{\Lambda}_j\}_{j=1}^m$. Go to the step 3) until convergence is satisfied.
5) Continuation by the step 3) until convergence is satisfied.

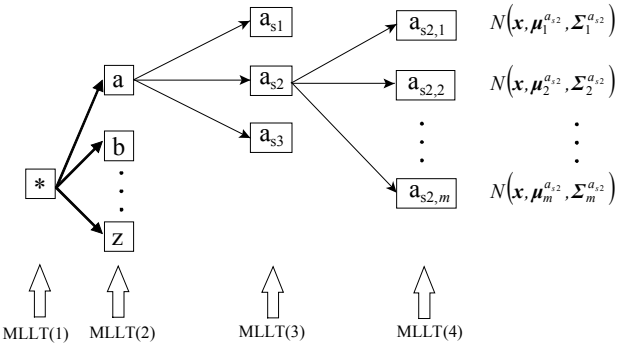

Figure 1. Different level of covariance matrices clustering.

As was described above the MLLT finds a global transformation matrix $\boldsymbol{A}$ for a „set" of Gaussians. This set can be defined for example from the topological or the phonetic point of view. Presented paper proposes some experiments with topological point of view. There are basically four different levels of clustering, which are indicated in Figure 1:

(1) There is only one transformation matrix for all components of all states of all monophones.

(2) There is one transformation matrix for all components of all states of individual monophone (it means that each monophopne has it's "own" transformation matrix).
(3) In this case one transformation matrix is connected to all components of individual state.
(4) Each component has it's own transformation matrix. This approach is very similar to the full covariance GMM model, because each symmetric positive definite (SPD) matrix has clear decomposition to the diagonal and transformation matrix (eigen values and eigen vectors).

The goal of following experiments was to explore how the different level of clustering influences the recognition results (Acc). The experiments were performed simultaneously for two different feature sets: first one was **DCT**(36) and second one worked with **LDA**(26) after transforms **DCT**(36)→**LDA**(26). All experiments were made with 3-state monophone based models where each state was represented by GMM. Owing to the extremely time-consuming computation burdens, all tests in this case were carried out using only 8 components for each GMM. Recognition results are shown in Table 3.

	# of ***A***	**DCT**(36)	**DCT**(36)→ **LDA**(26)
Diagonal	-	89.53	71.45
MLLT(1)	1	89.68	81.13
MLLT(2)	39	90.82	88,24
MLLT(3)	117	92.04	89.42
MLLT(4)	936	92.25	91.78
Full	-	92.47	91.97

Table 3. Accuracy of recognition for a different level of covariance matrices clustering in MLLT.

In the first row the recognition results of a "baseline" system with diagonal covariance matrix (no additional transformation was applied) is given. This model was used as a starting point for the MLLT optimization process. The second row shows the first case of MLLT clustering (one transformation matrix for all mixtures of all states and all monophones). Since we had 39 different monophones there were used 39 different transformation matrices, one for each monophone. Third variant uses 3*39=117 transformation matrices (one transformation matrix was assigned to each state of each monophone). Finally, one transformation matrix was used for each component of each state and each monophone. In this case there were 3*39*8=936 transformation matrices. Table 3 also mentions results obtained using full covariance model.

5. CONCLUSION

As was introduced in Section 1, decorrelation of features can be solved by a transformation of pattern space on a level of feature vectors. Experiments described in Table 1 and 2 indicate very good quality of the DCT in case of diagonal covariance matrices (here the DCT overcame the LDA). From results obtained for full covariance matrices it is evident better property of a transform based on the LDA. From results given in Table 1 and 2 it is also evident that correlation dependencies and possible non-Gaussian nature of density functions can be modeled by an increased number of Gaussian mixtures associated with a particular state of a HMM. The feature decorrelation technique based on the Maximum Linear Likelihood Transform brings evident improvements in comparison with concept of pure diagonal covariance matrices (see results in Table 3). It is important in what level of covariance matrices clustering the MLLT technique will be employed. The more transformation matrices (***A***) is used the better recognition accuracy is obtained but the more computations accompany this process. Therefore it is reasonable to find a suitable tradeoff between selected groups of Gaussians that should be decorrelated by individual transformation matrices and the recognition accuracy. Next steps of our research will be aimed at finding this compromise and also at extending this approach to the triphone-based concept

6. ACKNOWLEDGEMETS

This work was funded by the Academy of Science of the Czech Republic, project 1QS101470516.

7. REFERENCES

[1] Psutka, J.V., Müller, L.: Optimization of some parameters in the speech-processing module developed for the speaker independent ASR system. –In: Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics SCI'2003, Orlando, U.S.A., 2003, pp. 414-418.

[2] Psutka, J.V., Müller, L.: Building Robust PLP-based Acoustic Module for ASR Application. –In: Proceedings of the 10th International Conference Speech and Computer SPECOM'2005, Patras, Greece, 2005, pp.761-764.

[3] Pražák, A., et al.:: Automatic Online Subtitling of the Czech Parliament Meetings. TSD 2006. *Lecture Notes in Artificial Intelligence,* Springer-Verlag, Berlin, Heidelberg, 2006.

[4] Gales, M.J.F.: Semi-Tied Covariance Matrices for Hidden Markov Models. *IEEE Trans. on Speech and Audio Proc.*, vol.7, no.3, 1999, pp.272-281.

[5] Olsen, P.A., Gopinath, R.A.: Extended MLLT for Gaussian Mixture Models or Modeling Inverse Covariances by Basis Expansion. *IEEE Trans. on Speech and Audio Processing*, January 2004.