社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（12）
2004／12／20

混合主成分分析を用いた音声認識

シュスター・マイク†　堀　貴明†　中村　篤†

† 日本電信電話（株）ＮＴＴコミュニケーション科学基礎研究所　〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{schuster,hori,ats}@cslab.kecl.ntt.co.jp

あらまし　音声認識のための出力分布表現に対する混合主成分分析（MPPCA: Mixture of Probabilistic Principal Component Analysis）の適用について述べる．MPPCAによれば，混合分布において，少数の有効パラメータのみを持つ制約付き共分散行列で全共分散行列を近似することができる．またこの有効パラメータの数によってモデルの複雑度を適正に制御することができる．まず，MPPCAの基本的な構成を紹介し，つづいて簡単な拡張によりモデルの複雑度が容易に決定可能となることを示す．さらに，音声認識システムにMPPCAを実装した際発生する数値計算上の問題とその対処についても言及する．実験として共有状態数5,000，総ガウス分布数80,000のQuinphone HMMにおいてMPPCAを適用したモデルを日本語自然発話音声を用い評価した結果，対角共分散モデルでの誤り率22.2%が19.7%に削減され，全共分散モデルと同等の性能が得られたことを示す．あわせて全共分散モデルを上回る性能を得るための課題について考察する．

キーワード　音声認識, 共分散モデリング, 混合主成分分析

Mixtures of Probabilistic Principal Component Analyzers in Speech Recognition

Mike SCHUSTER†, Takaaki HORI†, and Atsushi NAKAMURA†

† Nippon Telegraph and Telephone Corporation, NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0237 Japan
E-mail: †{schuster,hori,ats}@cslab.kecl.ntt.co.jp

Abstract　This paper describes the application of Mixtures of Probabilistic Principal Component Analyzers (MPPCA) for modeling the observation distributions in a speech recognition system. The MPPCA model is a mixture of Gaussians with a constrained covariance approximating a full covariance with less effective parameters whose complexity can be controlled by the user. The paper summarizes the necessary basics of the MPPCA model, describes a simple extension of the basic model to set the user-defined complexity of the constrained covariance in a more automatic way and describes how to deal with numerical problems occuring for typical speech recognition systems. The MPPCA model is tested against a diagonal covariance and a full covariance model for our so far best acoustic model with 5000 quinphone clustered states and 80000 Gaussians total on a large, spontaneous Japanese speech task. Results show that we can improve error rates on the standard test set from 22.2% to 19.7% by moving to full covariances. For several MPPCA models tested we reach the same error rates with less effective parameters but fail to improve over using full covariances, for which possible reasons are discussed.

Key words　Speech recognition, covariance modeling, Probabilistic Principal Component Analysis

1. Introduction

Most speech recognition systems use Hidden Markov Models (HMMs) with Gaussian mixtures to model the observation distributions. In most cases these are diagonal Gaussians which have a number of advantages over other covariance structures that made them popular for use in speech recognition systems. Advantages include a) robust estimation of covariance parameters since their number grows only linearily with the dimensionality d of the feature space, b) simple implementation, c) memory efficient accumulation of sufficient statistics and storage of models, d) several schemes

for fast likelihood calculation known and e) the often heard property that a mixture with diagonal covariances can approximate any density given its parameters are correctly chosen. This last point is from a practical standpoint a little misleading since in most applications there is not enough data and/or parameters in the mixture as well as convergence problems in case of many parameters to be able to approximate the desired density with arbitrary accuracy.

The main disadvantage of using diagonal covariances is that feature dimensions are assumed to be uncorrelated within a Gaussian. Although typical features used for speech recognition (e.g. MFCCs with first and second delta features appended) are globally nearly uncorrelated they are less uncorrelated in specific subspaces which are modeled by individual Gaussians, therefore making more elaborate covariance structures an interesting alternative for building better models. The most extreme choice is to use full covariances which generally improve results but because the number of parameters in this case grows quadratically with the feature dimensionality d these are for high d easily prone to overfitting when trained with typical maximum likelihood procedures. Also, estimating and using full covariances in practice generally causes some numerical problems which are in more detail discussed below.

There are several covariance structures which try to aim at a level of complexity between *radial* (diagonal covariance structure with one shared parameter for all diagonal elements) and full covariances, most notably *semi-tied covariances* [1], *factored sparse inverse covariances* [2], the *extended maximum likelihood linear transform (EMLLT)* [3], the *subspace of precisions and means (SPAM)* [4]~[6] and finally the *mixture of inverse covariances (MIC)* model [7]~[9].

One recently published general covariance model which hasn't been explored much for use in speech recognition is the *mixture of probabilistic principal component analyzers (MPPCA)* [10], [11] which is characterized by an approximation of the full covariance matrix to reduce the number of free parameters and can be regulated by a user-defined intrinsic dimensionality $q < d$. Extensions of this model include a Bayesian approach where the intrinsic dimensionality is completely inferred from the data, leaving no significant settable parameters for the covariances [12], [13]. An application of a Bayesian PCA model for phoneme classification is described in [14].

The paper is structured as follows: In section 2. we review basics of the MPPCA model and discuss our approach to dealing with some numerical and implementation problems occuring for the large number of Gaussians as used in our speech recognition system. We decribe a simple (non-Bayesian) extension of the model to automatically set the intrinsic dimensionality q per Gaussian. Section 3. describes experiments with the MPPCA model using a large, spontaneous Japanese database (CSJ corpus). Finally, section 4. summarizes and discusses some ideas for improved MPPCA modeling approaches in classification tasks.

2. The MPPCA Model

This section summarizes the for this paper relevant parts of the Mixture Of Probabilistic Principal Component Analysis model according to [11] and explains our method of dealing with occuring numerical problems. It also shows a simple, but non-Bayesian way of setting the intrinsic dimensionality q per Gaussian automatically.

2.1 Mixtures of Probabilistic Principal Component Analyzers

Using the notation from [11] a regular mixture of Gaussians for d-dimensional data vectors $\mathbf{t}$ is defined as

$$p(\mathbf{t}) = \sum_{i=1}^{M} \pi_i p(\mathbf{t}|i) \tag{1}$$

with M being the number of mixtures and π_i being the ith mixing coefficient, and defining

$$p(\mathbf{t}|i) = (2\pi)^{-d/2}|\mathbf{C}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu}_i)^T\mathbf{C}_i^{-1}(\mathbf{t}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

as the ith Gaussian with $\mathbf{C}_i$ being its $d \times d$ symmetric covariance matrix and $\boldsymbol{\mu}_i$ being the mean of the data belonging to Gaussian i. The parameters $\mathbf{C}_i$, $\boldsymbol{\mu}_i$ and π_i are estimated by the EM algorithm from the N training data vectors to maximize its log-likelihood $\mathcal{L} = \sum_{n=1}^{N} ln\{p(\mathbf{t}_n)\}$.

In the MPPCA model the complexity of the covariance matrix can be controlled by the settable intrinsic dimensionality $1 \leqq q_i < d$ per Gaussian i, therefore allowing a finer control of the total complexity between a radial, diagonal or full covariance. Reference [11] shows that the constrained covariance matrix with $\mathbf{I}_d$ being a d-dimensional unit matrix can be modeled as

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}_d \tag{3}$$

with the maximum likelihood solution of the $d \times q$ matrix $\mathbf{W}$ being

$$\mathbf{W}_{ML} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I}_q)^{1/2} \tag{4}$$

where the columns of the $d \times q$ matrix $\mathbf{U}_q$ are the eigenvectors (sorted by their eigenvalues) of the data sample covariance matrix

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu}_n)(\mathbf{t}_n - \boldsymbol{\mu}_n) \tag{5}$$

and corresponding eigenvalues λ_i in the diagonal matrix $\boldsymbol{\Lambda}_q$.

The maximum likelihood solution for σ^2 is given by

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i \quad (6)$$

with $q < d$ being the parameter to set per Gaussian. The *noise* σ^2_{ML} can be interpreted as the average loss per discarded dimension.

In practice not $\mathbf{C}$ directly but its inverse is needed, which is for $q \ll d$ efficiently calculated as

$$\mathbf{C}^{-1} = \frac{1}{\sigma^2}(\mathbf{I}_d - \mathbf{W}(\sigma^2\mathbf{I}_q + \mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T) \quad (7)$$

avoiding the explicit inversion of the $d \times d$ matrix $\mathbf{C}$, an $\mathcal{O}(d^3)$ process.

Estimating the parameters for the MPPCA model means therefore after collecting full covariance statistics to pick q per Gaussian, do a (sorted) eigenvalue decomposition of the sample covariance matrix $\mathbf{S}$, use Eq.(6) to find σ^2_{ML}, calculate $\mathbf{W}_{ML}$ using Eq.(4) and finally generate the constrained inverse covariance matrix $\mathbf{C}^{-1}$ using Eq.(7) or invert $\mathbf{C}$ directly after using Eq.(3).

It should be noted here that although use of the MPPCA model reduces the number of effective parameters compared to the full covariance matrix and therefore can lead to better generalizing models, log-likelihood calculation speed and storage requirements are not reduced, unless $\mathbf{C}^{-1}$ is built out of $\mathbf{W}$ and σ^2_{ML} on demand each time it is needed, which will not be efficient for most applications.

2.2 Dealing with numerical problems

In practical applications it can happen that the sample covariance $\mathbf{S}$ is not positive definite, which it needs to be a) to be invertable by Cholesky decomposition and b) to be not ill-conditioned in terms of all non-negative eigenvalues and "useful" log-likelihoods if used in a Gaussian. Especially in applications with many thousand Gaussians like large speech recognition systems it is necessary to have a safe method to deal with non-positive definite sample covariance matrices. In our system we take care of this before updating $\mathbf{C}$ by a) flooring the diagonal elements of $\mathbf{S}$ with a small positive value and b) successively dividing the off-diagonal elements of $\mathbf{S}$ by 2 (therefore decreasing existing correlations) as long as $\mathbf{S}$ is not positive definite, which can be checked by Cholesky decomposition. This method guarantees that all $\mathbf{C}$ are always positive definite with the minor approximation that some Gaussians will have artificially decreased correlations.

2.3 Automatic determination of q

The standard maximum-likelihood MPPCA requires to set q per Gaussian manually. To find the (possibly different) best q for each Gaussian such that the model approximates best the true density means to search through all possible combinations of q for all Gaussians, which is even for moderately sized M and d impossible. An automatic way of determining q can be done using a Bayesian framework [12], [13], but implementation of these methods is slightly involved and hasn't been tried here.

A simple to implement and intuitively straight-forward method of automatically selecting q per Gaussian within the maximum-likelihood framework is by trying to keep a fixed relative portion of the sample covariance. Determine q such that the total loss in variance

$$\sigma^2_{loss} = \sum_{q+1}^{d}\lambda_i \quad (8)$$

per Gaussian is a fixed relative portion of the total variance

$$\sigma^2_{total} = \sum_{1}^{d}\lambda_i \quad (9)$$

such that q is subject to optimization of the inequality

$$\frac{\sum_{q+1}^{d}\lambda_i}{\sum_{1}^{d}\lambda_i} \leqq r \quad (10)$$

with a *single* parameter $0 < r \leqq 1$ for *all* M Gaussians of the mixture. Parameter q is easily determined per iteration from above inequality by increasing q as long as it is $\leqq r$. This can be done for all M Gaussians leading to possibly different intrinsic dimensionalities q_i per Gaussian all governed by the single parameter r.

3. Experiments & Results

Experiments were run to compare performance of the MPPCA model to a regular diagonal and full covariance model on a large speech recognition task. We used the A-set (male & female) lecture speech subset ($186k$ utterances, 230h) of the CSJ Japanese spontaneous speech database for training and test set 1 (10 lectures, 26515 words, perplexity 78.7 for trigram, out-of-vocabulary rate 2.4%) with a $30k$ dictionary and trigram for testing [15], which is currently one of the standard speech recognition benchmarks in Japan. Acoustic features were standard 39-dimensional MFCCs.

We used our so far best models on this task (5000 quinphone decision tree clustered states, 16 Gaussians with diagonal covariances per state) to run a single-pass retraining (aligning Forward-Backward with our diagonal model to calculate Gaussian occupation probabilities, but accumulating statistics for the new covariance mode) to estimate for all 80000 Gaussians and mixture weights new parameters depending on the wished covariance mode. Since for full covariances and for all MPPCA models the same complete full covariance sufficient statistics are needed, we only had to do a single run through the data, which is possibly slightly suboptimal but saves the time of iteratively realigning the

complete training data with full covariance models. Besides full covariances we built MPPCA models for different values of r leading to various average, minimum and maximum q as shown in table **Tab.1**. Around 30 sample covariance matrices were not positive definite which was fixed using the method explained in section 2. 2.

r	average q	min/max q
0.95	14.57	6/25
0.98	20.16	8/29
0.99	24.05	10/33
0.995	27.33	11/35
0.998	31.42	12/37

Tab. 1 Different values of relative noise threshold r and resulting average, minimum and maximum q for constrained covariance matrix in MPPCA model.

For decoding we used our single pass, full cross-word SOLON FST decoder with fully compiled FST networks decoding a complete lecture at a time, with a high beam of 250 and a maximum of 10000 hypotheses allowed at any time to avoid search errors. **Tab.2** shows final error rates for diagonal, full and all tested MPPCA covariances.

covariance	error (%)
diagonal	22.2
MPPCA ($r = 0.95$)	21.1
MPPCA ($r = 0.98$)	20.1
MPPCA ($r = 0.99$)	19.9
MPPCA ($r = 0.995$)	19.7
MPPCA ($r = 0.998$)	19.7
full	19.7

Tab. 2 Error rates for several different covariance structures using testset 1 of the CSJ database.

First of all the results show that for better recognition rates it certainly makes sense to use full covariances inspite of the vast increase of parameters to estimate – the improvement over using models with diagonal covariances is surprisingly large even though the diagonal covariance model is with 80000 Gaussians already rather big.

Although producing better results for some speakers, unfortunately none of the MPPCA models is overall better than the full covariance model. We expected a better model for the densities and therefore better generalization based on the experiments described in [11] and on our own informal small-scale density estimation experiments that were used to test the basic implementation.

After running the experiments we realized that at least one problem are the relative differences in the variances of the features. Since the eigenvalue decomposition of the sample covariance matrix depends on the individual variances of the dimensions they will need to be scaled, either all to unity or even better weighted by their relative importance for classification, to not loose important information when cutting off the small eigenvalues. For example, the delta-energy will have a lot smaller variance than the energy itself and therefore runs chance to be partially deleted by the PPCA procedure, although we know that the delta-energy is much more important for classification.

4. Conclusions

We tested an interesting constrained full covariance model (Mixture of Probabilistic Principal Component Analyzers) for speech recognition on a large Japanese speech recognition task. While all MPPCA models produced better results than using a diagonal covariance model, they failed to improve over simply using full covariances, although some of the MPPCA models achieved equal error rates with less effective parameters. We suspect that the main reason for not improving over using full covariances is that the feature variances were not appropriately scaled according to their relative importance for classification, which needs to be tested in the future.

Using full covariances or MPPCA models with a noise threshold r close to one for all 80000 Gaussians in our 5000 state quinphone system improved our so far best error rate on the CSJ testset 1 from 22.2% to 19.7%.

文　献

[1] M. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 7:272–281, 1999.

[2] J. A. Bilmes. Factored sparse inverse covariance matrices. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 1009–1012, Istanbul, Turkey, 2000.

[3] P. Olsen and R. Gopinath. Modeling inverse covariance matrices by basis expansion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 945–948, Orlando, 2002.

[4] S. Axelrod, R. Gopinath, and P. Olsen. Modeling with a subspace constraint on inverse covariance matrices. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2177–2180, Denver, 2002.

[5] S. Axelrod, R. Gopinath, P. Olsen, and K. Visweswariah. Dimensional reduction, covariance modeling, and computational complexity in asr systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 912–915, Hong Kong, 2003.

[6] K. Visweswariah, P. Olsen, and S. Axelrod. Maximum likelihood training of subspaces for inverse covariance modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 896–899, Hong Kong, 2003.

[7] V. Vanhoucke and A. Sankar. Mixtures of inverse covariances. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 900–903, Hong Kong, 2003.

[8] V. Vanhoucke and A. Sankar. Variable length mixture of inverse covariances. In *Proceedings of the European Con-*

ference on Speech Communication and Technology, pages 2605–2608, Geneva, Switzerland, 2003.

[9] Vincent Vanhoucke. *Mixture of inverse covariances*. PhD thesis, Department of Electrical Engineering, Stanford University, 2003.

[10] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, 1999.

[11] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[12] C. M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks*, volume I, pages 509–514, Edinburgh, UK, 1999.

[13] C. M. Bishop. Bayesian PCA. In S. A. S. M. S. Kearns and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 382–388. MIT Press, Cambridge, MA, 1999.

[14] O. W. Kwon, T. W. Lee, and K. Chan. Application of variational Bayesian PCA for speech feature extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 825–828, Orlando, 2002.

[15] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition using the corpus of spontaneous japanese. In *Proceedings of the Spontaneous Speech Processing & Recognition Workshop*, pages 135–138, Tokyo, Japan, 2003.