Covariance Matrix Enhancement Approach to Train Robust Gaussian Mixture Models of Speech Data

Jan Vaněk, Lukáš Machlica, Josef V. Psutka, Josef Psutka

University of West Bohemia in Pilsen, Univerzitní 22, 306 14 Pilsen
Faculty of Applied Sciences, Department of Cybernetics
{vanekyj,machlica,psutka_j,psutka}@kky.zcu.cz}

Abstract. An estimation of parameters of a multivariate Gaussian Mixture Model is usually based on a criterion (e.g. Maximum Likelihood) that is focused mostly on training data. Therefore, testing data, which were not seen during the training procedure, may cause problems. Moreover, numerical instabilities can occur (e.g. for low-occupied Gaussians especially when working with full-covariance matrices in high-dimensional spaces). Another question concerns the number of Gaussians to be trained for a specific data set. The approach proposed in this paper can handle all these issues. It is based on an assumption that the training and testing data were generated from the same source distribution. The key part of the approach is to use a criterion based on the source distribution rather than using the training data itself. It is shown how to modify an estimation procedure in order to fit the source distribution better (despite the fact that it is unknown), and subsequently new estimation algorithm for diagonal- as well as full-covariance matrices is derived and tested.

Keywords: Gaussian Mixture Models, Full Covariance, Full Covariance Matrix, Regularization, Automatic Speech Recognition

1 Introduction

Gaussian mixture models (GMMs) are very popular models of multivariate probabilistic distributions in various domains, including speech and speaker recognition domains. For given training data set, one is confronted with three mutually dependent problems:

- How complex the model should be? How many Gaussians? Diagonal- or full- covariance matrix?
- How to estimate model parameters to fit also to unseen data?
- Is the model numerically stable? Variances may go approach zero and full-covariance matrices may be ill-conditioned.

The problem with numerical stability can be handled in relatively easy way. In the case of low variances, the problematic model component can be simply discarded or a minimum variance threshold can be specified. The threshold is usually a fixed fraction of the global variance of the entire data set [1]. From the other hand, the threshold introduces an additional prior, in the form of a magic number, into the estimation algorithm

and it may be dubious. In the case of full-covariance matrices, smoothing or shrinkage methods can be used [2], [3]. They are based on the lowering of the off-diagonal values of the covariance matrix.

Choosing a proper model complexity is the main challenge [4–8]. It is the trade-off between an accurate training data fit and a generalization ability to unseen data. The trade-off can be handled easily if one knows both information. But, in our case, we have only the first half: the training data. The generalizations ability to unseen data can be only estimated.

The first class of solutions involves penalization of the training data fit by the model complexity. The most popular criteria are Bayesian information criterion (BIC) [9] and Akaike's information criterion (AIC) [10]. BIC and AIC penalize the mean of the log-likelihood of training data by the number of model parameters. BIC penalization depends moreover on a number of training data samples. These criteria may work nicely in most cases. However, they depend on the log-likelihood of the training data. It means that they depend also on a way how the minimum variance is handled. In the case of outliers, the variance will go to zero and the log-likelihood will go to infinity. An extreme example is the mixture model of Dirac functions placed on locations of training data samples. This model is completely incorrect, but it gives the best BIC and AIC. Additional weak point in BIC is the number of samples. The samples are assumed mutually independent, but it is not true in most of real cases (e.g. speech data). This may be solved by a tunable gain of the penalization part in the BIC, but it brings an additional magic constant into the training set-up.

In the second class of solutions an unseen-data performance is estimated via cross-validation technique. In the simplest case, the available data are split into two parts - training and development. The training part is used to train the model parameters and the development part is used to evaluate the model performance. In a more complex case - the true cross-validation - data are split into more parts and an one-leave-out approach with all combinations is used. The cross-validation works well on real data, but it has also disadvantages. The first one is much higher computational requirements, which grow with the number of data splits. The second disadvantage resides in fact that the result varies with the number of splits and data distribution between these parts.

The approach proposed in this paper is based on an assumption that the training data and the testing data are generated from the same source distribution. When this is not true, one should use an appropriate normalization and/or adaptation technique to compensate for the difference as much as possible. The key part of the approach is to use a criterion based on the source distribution rather than the training data itself. Naturally, the source distribution is unknown. But, we are able to modify the estimation procedure in order to fit the distribution better despite the fact that it is unknown. Based on a criterion, we have derived how the covariance matrices need to be enhanced, and we have proposed a new estimation algorithm for diagonal as well as full covariance matrices. Also, a very useful feature of the algorithm is the ability to leave out the redundant Gaussians. Therefore, the final GMM has an optimal number of components. Moreover, such enhanced full covariance matrices are well-conditioned. Thus, this feature prevents numerical stability issues. The proposed approach may be understood also as the extreme case of cross-validation, where each data sample forms a new part.

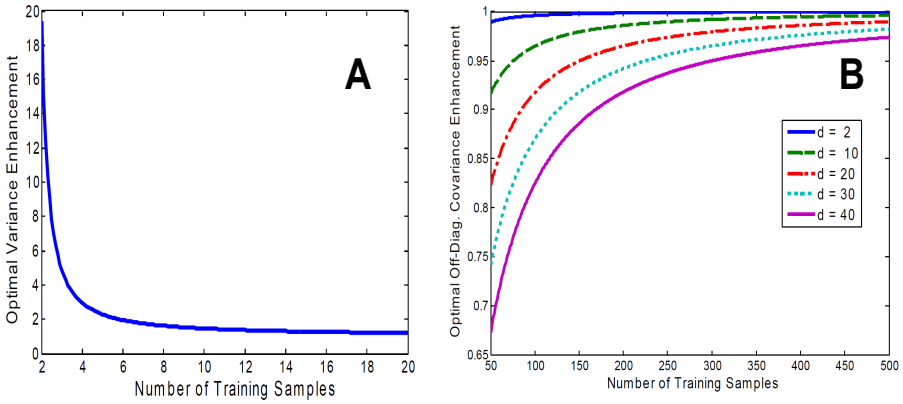

Fig. 1. *Fitted functions of the ratio between the optimal estimate of variance of the source-distribution and the optimal estimate of variance (Figure A) and covariance (Figure B) of training data for various numbers of training samples.*

2 Robust GMM Training

2.1 Estimation of Single Gaussian Model

Assume we have i.i.d. random data set $X = \{x_1, x_2, \ldots, x_n\}$ sampled from univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ (the real-speech data case will be discussed later in the paper). The Maximum Likelihood (ML) estimates of sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ are given by the well-known formulas:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{2}$$

From Central Limit Theorem, it can be derived that the estimate of the sample mean $\hat{\mu}$ has normal distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ and the estimate of the sample variance $\hat{\sigma}^2$ has Chi-square distribution with variance equal to $\frac{2\sigma^2}{n-1}$. These estimates give the best value of the ML criterion for the training data set X. On the other hand, these estimates are not optimal, and they do not achieve the best value of the ML criterion for unseen data generated from the source distribution $\mathcal{N}(\mu, \sigma^2)$.

We performed a very large amount of Monte Carlo simulations for various lengths of data sets, and we have found out that the ML estimate of the optimal variance of the source distribution should have a higher value. The difference grows especially for data sets containing few samples. We fitted a function of the ratio between the optimal estimate of variance of the source-distribution and the optimal estimate of variance of

training data. The function is given in dependence on the number of training samples and it is shown in Figure 1 A. The ratio function can be used to enhance the variance estimate $\hat{\sigma^2}$ given by the equation (2). The enhanced variance estimate $\tilde{\sigma}^2$ is calculated in following way:

$$\tilde{\sigma}^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n - 1.25}\right)^{3.5} \tag{3}$$

In the multivariate case, assuming a diagonal covariance matrix, the same variance enhancement can be used for individual dimensions. The variance enhancement also handles numerical stability issues. The enhanced estimate cannot be close to zero even for outliers, because the enhanced variance grows fast when only a few samples are available.

In the case of multivariate data and full covariance model, the covariance matrix should be enhanced in the same way - by multiplication with the coefficient from equation (3). Moreover, the off-diagonal part of the matrix need to be corrected. We performed some additional Monte Carlo simulations to fit the optimal function for the off-diagonal part. The optimal correction coefficient was found in the interval $\langle 0, 1 \rangle$. It means that for small training sets some suppression of the off-diagonal elements is needed. This is similar to smoothing and shrinkage methods [2], [3]. The optimal value of the correction coefficient depends on the number of training samples n and on the number of dimensions d. The fitted functions are shown in Figure 1 B. The enhanced estimate of the off-diagonal element $\hat{s}_{ij}$ of the covariance matrix, which was enhanced by equation (3) already, is

$$\tilde{s}_{ij} = \hat{s}_{ij} \left[1 - \left(\frac{d}{d - 1 + n}\right)^{1.4}\right]. \tag{4}$$

Covariance matrix enhanced this way is also well-conditioned. The Monte Carlo simulation and the fit was done for dimensions in range from 2 to 50. Therefore, the fit may be inaccurate in cases with significantly higher dimensions.

2.2 Real Speech Data

Real speech data (e.g. MFCC or PLP vectors augmented by delta and acceleration coefficients) are not i.i.d.. Subsequent feature-vectors are mutually dependent. This is the consequence of the speech processing itself. There is an overlap of the FFT window and the delta and acceleration coefficients are computed from neighbouring feature-vectors. Also, the speech production is continuous, i.e. time dependent. It means that n used in formulas (3) and (4) cannot be directly the number of feature-vectors.

An number of independent feature-vectors $\tilde{n}$ needs to be estimated too from the given data set of n real feature-vectors. For such an estimation we used a normalized mean of absolute differences of consecutive feature-vectors:

$$\tilde{\delta} = \frac{1}{d} \sum_{j=1}^{d} \frac{1}{n\hat{\sigma}_j} \sum_{i=2}^{n} |x_{ij} - x_{i-1,j}|. \tag{5}$$

In the case of dependent data, the difference $\tilde{\delta}$ is smaller than for the independent data. We simulated various filter lengths and many filter shapes (e.g. Hamming, Blackman, Triangle, Rectangular) to analyze a relation between the data dependency influenced by the filter and the difference $\tilde{\delta}$. The relation depends mainly on the filter length. The dependence on the filter shape is minor. Again, we fitted a function to estimate the number of independent feature-vectors $\tilde{n}$ from the difference $\tilde{\delta}$ and the number of feature-vectors n:

$$\tilde{n} = 1 + (n-1)0.7\tilde{\delta}^3. \tag{6}$$

2.3 Enhanced Gaussian Mixture Models

In the case of GMM, the above described approach can be used for individual components. The optimality is not ensured since the overlap of Gaussian components is ignored. All the estimates are calculated incorporating the posterior probability of the individual Gaussian components. Instead of the number of feature-vectors n, the sum of posteriors is used in the equations (3), (4), and (5). Only in the equation (6) the sum of posterior square roots is used.

2.4 Training Procedure

We use a modified Expectation-Maximization (EM) algorithm. We modified the estimation equations as described above. The iterative training converges in most cases, but the convergence is not assured. Appropriate number of iterations is higher in comparison with the classic EM algorithm. Some model components may be found redundant during the iterations. This means that other components comprise most of the data from the redundant component. The redundant component should be discarded since its estimates become very inaccurate. Discarding the redundant component naturally produces a model with an optimal number of components.

The modified EM algorithm does not converge to a global optimum alike the classic EM. The initial seed is important. According to our experiments, starting with a single component followed by subsequent split of the component with highest weight results to a reliable final model. However, it is a very time-consuming method when dealing with large datasets, where a random initialization of all the components gives a model in much shorter time. In that case, we recommend to try out several random initializations and select the model witch gives the best value of the ML criterion. The individual random initializations may vary in the number of components, but the proposed algorithm provides (after a few iterations) their optimal subset. Hence, the redundant ones will be discarded.

3 GMM Estimator Software

We incorporated previous methods of the covariance matrix enhancement into our GMM estimator software. The GMM estimator supports diagonal and full covariance matrices and it is developed for processing of very large datasets. It uses CUDA GPU acceleration (if available) [11], [12] or multi-threaded SSE implementation in all other case. It

is free for academic use. More information is available at `http://www.kky.zcu.cz/en/sw/gmm-estimator`.

4 Speech Recognition Results

We employed the proposed GMM training approach into an acoustic model training. For this paper, we choose a ML trained triphone Hidden Markov Model (HMM) baseline where each tied-state has a uniform number of components with diagonal covariance matrices. We used this model to label all the training feature-vectors with tied-state labels. Each state was then trained as an independent GMM. On the end, we collected all the GMMs and constructed a new HMM.

We compared models with diagonal as well as full covariance matrices. Three variants were assumed. We tested various uniform numbers (each HMM state has same number of components) of components (i.e. 2, 4, 8, 16, 32) and kept the best performing ones: 16 components for diagonal covariance models and 8 components for full covariance models. We did not use Speaker adaptive training, discriminative training, nor adaptation, in order to keep the influence of the modelling approach evident. Summary of the compared models:

- Diagonal covariance HMM with uniform number of components per state, 16 components used (denoted as *Diag_16G*), trained by classical EM algorithm.
- Enhanced diagonal covariance HMM with the target number of components equal to 16, but in one third of states some of the components were marked as redundant and left out (marked as *Diag_16G_Enh*).
- Enhanced diagonal covariance HMM with variable number of components, the largest component was split until the ML criterion grew (marked as *Diag_Vari_Enh*).
- Full covariance HMM with uniform number of components per state equal to 8 (marked as *Full_8G*) trained by classical EM algorithm. 10% of the states was not able to be trained with full covariances because of ill-conditioning. These states was replaced by enhanced GMMs from the following model.
- Enhanced full covariance HMM with target number of components equal to 8, but one half of states had some components redundant (marked as *Full_8G_Enh*).
- Enhanced full covariance HMM with variable number of components, the largest component was split until the ML criterion grew (marked as *Full_Vari_Enh*).

4.1 Test Description

A corpus *Bezplatne Hovory* was chosen as a data source for the experiments. It is a *Switchboard* like telephone speech corpus, recorded at 8kHz in Czech language. It contains spontaneous speech with unlimited vocabulary, hence it is hard to get high recognition accuracy compared to some domain specific corpora.

280h of speech were selected for training and other 2h were selected for tests. Feature vectors were standard PLPs with delta and acceleration coefficients followed by the Cepstral Mean Subtraction. Total dimension of the vectors was 36.

Czech language belongs to flexible languages, therefore the vocabulary used needs to be extremely large in order to carry most of the pronounced words. Since less than 5 million words from the training set transcriptions do not suffice for such a task, a language model was trained from distinct text sources. Next, 200 million words mainly from internet forums and blogs were mixed with training set transcriptions in the ratio 1:3. Resulted trigram back-off language model contained 550k words (650k baseforms). Kneser-Ney smoothing was used. A perplexity of the test set was 102 with 1.17% of OOV words.

Table 1. Recognition results

Model	#States	#Gaussians	#Gaussians/#States	WER[%]
Diag_16G	4104	65,654	16.0	45.24
Diag_16G_Enh	4104	54,444	13.3	45.02
Diag_Vari_Eng	4104	108,422	26.4	44.69
Full_8G	4104	32,809	8.0	42.73
Full_8G_Enh	4104	28,246	6.9	41.89
Full_Vari_Enh	4104	29,870	7.3	40.81

The results of the speech recognition are shown in Table 1. Most interesting is the last column: Word Error Rate (WER). The full covariance models perform better than diagonal ones in this task. The full covariance models are also more sensitive to the selected training algorithm. The enhanced training procedure with variable number of components per state gave best results for both diagonal and full covariance models. The middle column with a total number of Gaussians is also of interest. It illustrates how many redundant Gaussians were present in the uniform models. The full covariance model with variable number of components *Full_Vari_Enh* performed better by 4.5% absolutely, when compared to the baseline diagonal model *Diag_16G*.

The overall WERs are somewhat high. This is caused by the difficulty of the task - spontaneous telephony speech with unlimited vocabulary, which contains also slang and expressive words. We needed to add more than a half million of words to the vocabulary in order to carry the speech variability. We also did not use any adaptation nor discriminative training techniques to keep the influence of the training method evident. Merging the enhanced covariances, discriminative training and adaptation is going to be the focus of our future research.

5 Conclusions

The approach of covariance matrix enhancement was proposed and described in this paper. It handles all the most problematic issues from the GMM training: optimal model complexity, unseen data, numerical stability. The key idea is to move the focus of ML criterion from the training data to the source distribution of the data. The covariance

matrix needs to be enhanced to get an optimal ML criterion. We performed a very large set of Monte Carlo simulations to get the optimal enhancement of a covariance matrix.

The proposed approach was incorporated into our high-performance GMM training software, which is free for use for research community. Finally, we successfully tested the new approach incorporating it to the training of acoustic models for ASR. The significant reduction of WER was achieved using the covariance matrix enhancements.

6 Acknowledgments

This research was supported by the Technology Agency of the Czech Republic, project No. TA01011264.

References

[1] S. Young et al.: The HTK Book (for HTK Version 3.4). In: Cambridge, 2006.
[2] Diehl, F., Gales, M.J.F., Liu, X., Tomalin, M., Woodland, P.C.: Word Boundary Modelling and Full Covariance Gaussians for Arabic Speech-to-Text Systems. In: Proc. INTERSPEECH 2011, p.p. 777-780.
[3] Bell, P., King, S.: A Shrinkage Estimator for Speech Recognition with Full Covariance HMMs. In: Proc. Interspeech 2008, Brisbane, Australia.
[4] Bell, P.: Full Covariance Modelling for Speech Recognition. In: Ph.D. Thesis, The University of Edinburgh.
[5] Lee, Y., Lee, K.Y., Lee, J.: The Estimating Optimal Number of Gaussian Mixtures Based on Incremental k-means for Speaker Identification. In: International Journal of Information Technology, Vol.12, No.7, pp. 13-21, 2006.
[6] Figueiredo, M., Leitão, J., Jain, A.: On Fitting Mixture Models. In: Proc. EMMCVPR 1999, pp. 54-69, Lecture Notes In Computer Science, Springer-Verlag London.
[7] Mclachlan, G.J., Peel, D.: On a Resampling Approach to Choosing the Number of Components in Normal Mixture Models. In: Computing Science and Statistics, Vol. 28, pp. 260-266, 1997.
[8] Paclík, P., Novovičová, J.: Number of Components and Initialization in Gaussian Mixture Model for Pattern Recognition. In: Proc. Artificial Neural Nets and Genetic Algorithms, pp. 406-409, Springer-Verlag Wien, 2001.
[9] Schwarz, G.E.: Estimating the dimension of a model. In: Annals of Statistics, Vol. 6 (2), pp. 461-464, 1978.
[10] Akaike, H.: On entropy maximization principle. In: Applications of Statistics, North-Holland, Amsterdam, pp. 27-41. 1977.
[11] Machlica, L., Vanek J., Zajic, Z.: Fast Estimation of Gaussian Mixture Model Parameters on GPU using CUDA. In: Proc. PDCAT, Gwangju, South Korea, 2011.
[12] Vanek J., Trmal, J., Psutka, J.V., Psutka, J.: Optimized Acoustic Likelihoods Computation for NVIDIA and ATI/AMD Graphics Processors. In: IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, 6, pp. 1818-1828, 2012.