Efficient Gaussian Mixture for Speech Recognition

Réduction de mélange Gaussien pour la reconnaissance de la parole

Leila Zouari
Gérard Chollet

2006D001

2006

Département Traitement du Signal et de l'Image
Groupe Perception, Apprentissage et Modélisation

Dépôt légal : 2006 – 1er trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Ecole Nationale Supérieure des Télécommunications

Groupe des Ecoles des Télécommunications - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.enst.fr

Département TSI

Réduction de mélange Gaussien pour la reconnaissance de la parole

Résumé

Ce document présente une technique de groupement hiérarchique pour déterminer le nombre optimal de composants dans un mélange gaussien. L'idée consiste à partir d'un mélange comportant un nombre important de gaussiennes à regrouper deux à deux dans une structure d'arbre, en utilisant une distance basée sur la similarité entre distributions. La coupure de cet arbre à un niveau donné permet d'obtenir un nouveau mélange dont les performances seront comparées à celle du mélange initial ainsi qu'à un mélange obtenu par regroupement hiérarchique en utilisant une métrique basée sur la perte en vraisemblance. Différents critères de coupure de l'arbre sont également étudiés. Les expériences réalisées sur la base de données d'émissions radio-diffusées ESTER ontt permis de noter une importante réduction absolue du taux d'erreur de l'ordre de 4.8% dans un intervalle de confiance de 1%.

Mots clefs : *Reconnaissance de la parole, mélange Gaussien, Modèles de Markov cachés.*

Efficient Gaussian Mixture for Speech Recognition

Leila Zouari and Gérard Chollet
GET-ENST/CNRS-LTCI
6 rue Barrault 75634 Paris cedex 13, France

zouari, chollet@tsi.enst.fr

Abstract

This article presents a clustering algorithm to determine the optimal number of components in a Gaussian mixture. The principle is to start from an important number of mixture components then group the multivariate normal distributions into clusters using the divergence, a weighted symmetric, distortion measure based on the Kullback-Leibler distance. The optimal cut in the tree, i.e. the clustering, satisfies criteria based on either the minimum amount of available training data or dissimilarities between clusters. The performance of this algorithm is compared favorably against a reference system and a likelihood loss based clustering system. The tree cutting criteria are also discussed. About an hour of Ester, a French broadcast News database is used for the recognition experiments. Performance are significantly improved and the word error rate decreases by about 4.8%, where the confidence interval is 1%.

1. Introduction

Nowadays, state of the art Hidden Markov Models based large vocabulary speech recognition systems make use of a important number of Gaussian distributions to improve their acoustic modeling accuracy. One of the disadvantages of this practice is the increase of the complexity of the system making it unsuitable for practical, embedded or even real time applications. Several criteria are generally used to stop growing the mixture, i.e. the number of Gaussian distributions. Mainly a tradeoff is to be found between the model precision and the ability to accurately estimate the model parameters. In the literature three classes of approaches are used to stop growing a mixture:

- when the amount of training data is insufficient (a threshold is placed on the number of frames used to estimate the mixture components),
- if no significant likelihood increase is observed,
- and when the Bayesian Information Criterion (BIC) gain becomes negative or below a threshold. This criterion controls the model complexity by penalizing the likelihood with the number of parameters.

An alternative proposed by Messina [1] is to grow a mixture only when it's distance to a frame is important. So, distances between a frame and Gaussians are computed and the minimum is selected. If this minimum distance is less than a threshold the component is updated with this frame otherwise a new mixture component is created. To decrease the number of mixture components in a phonetically tied mixture system, Digalakis [4] classifies the Gaussian distributions and re-estimates the obtained clusters. The clustering metric is based on the increase of entropy due to merging distributions. This way, the number of Gaussians is reduced to less than 40% with a little degradation of performance. Besides, this system is more accurate (+0.8%) than the reference one using the same number of Gaussians.
In the present work it is proposed to determine the optimal number of components in a Gaussian mixture models using a growing-clustering process, following the same principle of clustering as Digalakis, and we introduce the weighted cross entropy metric for better distributions classification. The idea driving this procedure is to explore a large set of components, then the set dimension is reduced by merging close elements. So, for each Gaussian mixture, distributions are grouped into a binary tree structure and every cut in the tree defines a possible clustering. To determine the optimal cut in the tree, two criteria are experimented: a data driven one and a dissimilarity based other. For each criterion, the weighted Kullback-Leibler divergence performance is compared to the initial system and also to a loss likelihood based clustering system.
The remainder of this paper is organized as follows: section 2 outlines the classification process, presents the proposed weighted Kullback-Leibler metric and details the tree cutting criteria, section 3 reports on tests protocols and results, the conclusions and prospective work are described in section 5.

2. Gaussian distributions classification

2.1. Clustering process

In order to build Gaussian trees, hierarchical bottom-up classification algorithm is applied to each mixture. It performs in many steps:

- Compute distances between all pairs of distribution.
- Merge the closest two distributions as follows:
  If $g_1(\omega_1, \mu_1, \sigma_1)$ and $g_2(\omega_2, \mu_2, \sigma_2)$ are merged into $g_3(\omega_3, \mu_3, \sigma_3)$ then :

  $$\omega_3 = \omega_1 + \omega_2$$
  $$\mu_3 = \frac{\omega_1}{\omega_1 + \omega_2}\mu_1 + \frac{\omega_2}{\omega_1 + \omega_2}\mu_2$$
  $$\sigma_3 = \frac{\omega_1}{\omega_1 + \omega_2}\sigma_1 + \frac{\omega_2}{\omega_1 + \omega_2}\sigma_2 + \frac{\omega_1\omega_2}{(\omega_1 + \omega_2)^2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

  g3 replaces (g1, g2) in the set whose size is reduced by one.
- If the number of Gaussians is greater than 1 go to step1.

2.2. Metrics

Two distances are used: likelihood loss based distance and weighted relative entropy based metric.
- Loss likelihood based metric: If g1 and g2 are merged into g3 then the likelihood loss (pv) is the difference between the likelihoods of g1 and g2 and the likelihood of g3 on the training data:

$$PV(g_1, g_2) = log\frac{\|\sigma_3^{(n_1+n_2)/2}\|}{\|\sigma_1^{n_1/2}\|\|\sigma_2^{n_2/2}\|}$$

This metric is somewhat similar to the loss of entropy based distance used by Digalakis [4]. It was successfully used in model adaptation [2].
- The weighted symmetric Kullback Leibler divergence (Klp): It is expressed as the distance between two probability density functions weighted by the amount of training data.

$$KLp(g_1; g_2) = \frac{1}{2}tr(\omega_1\frac{\sigma_1}{\sigma_2} + \omega_2\frac{\sigma_2}{\sigma_1}) + \frac{1}{2}(\mu_1 - \mu_2)^T(\frac{\omega_1}{\sigma_1} + \frac{\omega_2}{\sigma_2})(\mu_1 - \mu_2) - (\omega_1 + \omega_2)d$$

d is the dimension of the parameters vectors.
The use of the information provided by the amount of training data is advantageous if training and testing data have the same proportions otherwise it can be harmful.

2.3. Tree cutting

From the root of the tree to the leaves, different cuts can be defined allowing many classifications. Three cutting ways are proposed:

- Fixed: We consider a constant number of classes for each mixture. Beginning from the leaves, traverse each level till the number of nodes at the corresponding stage reaches the predefined value of classes and cut.
- Weight based: The number of classes depends on the amount of the available data to estimate the distributions of each class. Starting from the root, the tree is processed and we stop at node for which the children weight is less than a predefined threshold.
- Distance based: tree cutting is performed when the distance between two levels reach a maximum value. Considering only the maximum can lead to a very little (or large) number of clusters, besides many important distances can be close to the maximum value. for all these reasons, several cuttings per tree have been considered. Each cutting is operated in a particular level of the tree.

For weight and distance criteria, as the number of Gaussians per state can be different, a mean value is computed.
The resulting mixtures are re-estimated by means of Baum Welch algorithm.

3. Experiments and results

3.1. Resources

All the experiments are conducted using parameter vectors with 12 MFCC coefficients, energy, and their first and second derivatives. The 40 acoustic models are context independent with 3 states per model. For the training task, about 82 manually transcribed hours of the Ester train database [3] are used. The dictionary contains 118000 words (with 65000 distinct words). The language model is formed by 4 millions of bigrams and trigrams. Tests are conducted using an hour of Broadcast News extracted from the Ester test data set.
The initial system contains 256 Gaussians per state. For each mixture, the 256 Gaussians are classified by a bottom up hierarchical algorithm. Depending on the experiments, likelihood loss or weighted cross entropy based metric is used for clustering. Then classes are obtained by cutting the binary tree following a criterion: fixed, weight based or distance based number of clusters.
In order to compare the different systems, several reference systems 32, 64, 80, 128, 180, 256, 220 and 512 Gaussians per mixture are produced and evaluated.

3.2. Fixed classes

The number of classes is fixed and is the same for the reference (Ref), the loss likelihood (pv) and the weighted Kullback-Leibler (klp) based systems. After clustering, the obtained pv and klp models are trained. We find that au maximum two iterations are needed to estimate these models parameters. Results within a confidence interval of 1% are as follows:

Table 1. wer for Ref, PV, and KLp systems

Gaussians nbr	Ref (%)	PV (%)	KLp (%)
32	42.6	40.6	39.5
64	40.4	38.0	37.5
80	38.3	37.4	36.9
128	37.3	36.2	36.2
180	36.4	36.1	36.2
220	36.3	35.8	35.5
256	36.3	-	-
512	35.5	-	-

Table1 and figure1 show that both pv and klp systems outperform the reference one, with a little advantage for the latter. Especially, for the klp models with 32 and 64 Gaussians per state, the word error rate (wer) decreases by about 3% compared to the reference system. With a large number of clusters differences are less interesting.
Performance of the klp system using 220 Gaussians per state are similar to the 512 reference one.

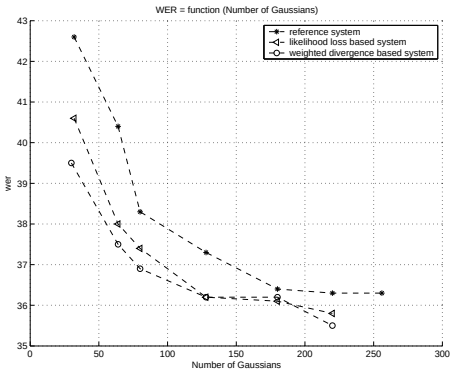

Figure 1. wer vs number of Gaussians, for ref, pv and klp systems

3.3. Weight based classes

By using this criterion, the number of clusters in the different mixtures is variable and depends on the acoustic variability of each state. Besides, this way we ensure that each cluster has sufficient amount of training data to estimate it. So, in each level of the tree, when a node reaches the global minimum of this level, we cut at his parent level. Results are as reported in table 2 and figure 2.

Table 2. pv and klp weight based cutting

Metric	Gaussians	wer (%)
KLp	28	40.0
	53	36.6
	150	35.9
	195	36.0
PV	53	39.5
	101	36.8
	156	36.5

We notice that using the weight criterion, the klp system outperforms both pv and the reference system. Especially, with only a mean of 53 Gaussians per state, it's performance is close to that of the reference system with 256 Gaussians. Besides, the wer decreases by about 4.8% compared to the initial system using the same number of Gaussians.
For klp system with 28 Gaussians per state, the wer is also better than the initial 64 Gaussians one. Finally, klp models with 150 Gaussians are quite performing as the 512 reference system.

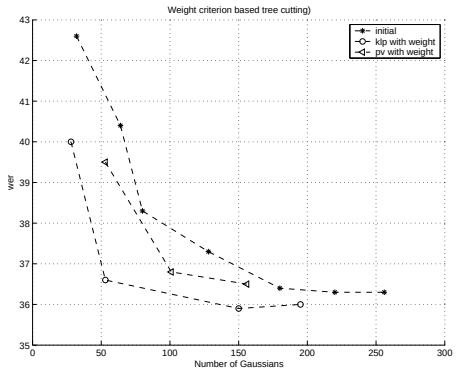

Figure 2. Weight based tree cutting

3.4. Distance based classes

This criterion prevents clustering too distant Gaussians:

- if distributions are too different in the case of klp distance
- or merging them leads to a big likelihood loss if pv based metric is employed

We consider several levels of the tree and cut when the distance between two clusterings is the maximum in this level. The obtained results are reported in table 3.

Table 3. pv and klp distance based cutting

Metric	Gaussians nbr	wer (%)
KLp	30	40.7
	59	37.7
	101	36.1
	196	35.9
PV	44	39.4
	94	36.7
	204	35.8

Once again, we see that pv and klp systems outperform the reference one, and that the klp divergence based system is the best. Applying klp or pv clustering process, we obtain globally the same performance as the reference system using only about 40% of the total number of Gaussians. These results are interesting but they remain less important than the previous experiments (53 Gaussians) in which this number is reduced to 20%.

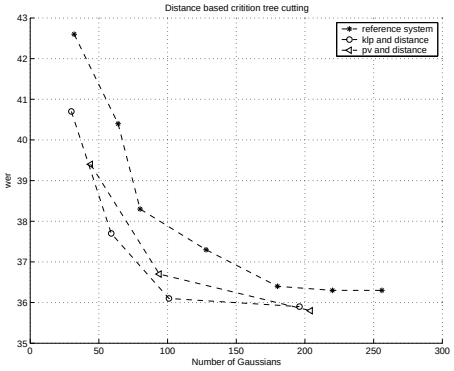

Figure 3. Distance based tree cutting

3.5. Weight versus distance

To compare distance and weight criteria we plot the correspondent curves using either klp or pv metric.

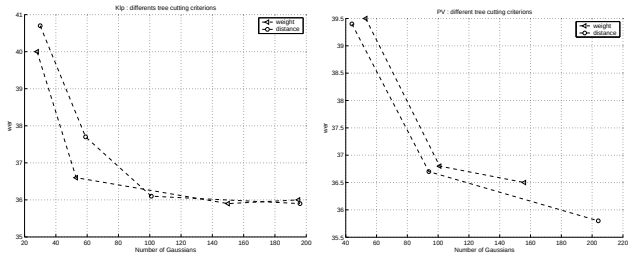

Figure 4. Weight and distance based tree cutting for respectively klp and pv systems

In the klp system, the weight criterion performs better than distance one, especially when the number of clusters is low. In the case of pv clustering, it is the opposite situation and the distance is better. These results can be interpreted as follows:
- when klp clustering metric is used, no particular attention is given to the amount of training data available for each cluster. Only resembling Gaussians are merged, ensuring that at each level clusters are as distant as possible. So in some levels many clusters could not have enough training data, and cutting at these levels is not interesting.
- In the case of pv based clustering, the loss of likelihood is minimum at each level. So the resulting clusters are as representative as possible of the training data. Knowing that no information about similarity of clusters to each others is taken into account, many resemblant clusters can be present in the same level. In this case the distance based cutting criterion can remove the redundant information.

4. Conclusion and discussion

An hierarchical Gaussians clustering algorithm for optimal mixture dimension determination based on a weighted Kullback Leibler distance (klp) is described. Experiments varying the tree cutting criterion show that in all the cases the proposed metric outperforms the loss likelihood based clustering (pv) system and the initial one. We also notice that the tree cutting criterion depends of the clustering distance. While the weight based tree cutting criterion is better for the klp system, the distance based cutting is more interesting for the pv system. In both cases, the good criterion of cut is that which brings information the distance does not take in consideration. As a perspective to this work, a linear discriminant analysis per state can be deduced from the Gaussian classification. This way, more separate and hence discriminant parameter vectors can be constructed et tested for better performance speech recognition systems.

References

[1] R. Messina and D. Jouvet. Sequential clustering algorithm for gaussian mixture initialisation. In *In proceedings ICASSP*, 2004.

[2] C. Mokbel. Online adaptation of hmms to real life conditions: A unified framework. In *IEEE Transaction on Speech and Audio Processing*, 2001.

[3] J. B. S. Galliano E. Geoffrois D. Mostefa, K. Choukri and G. Gravier. The ester phase ii campaign for the rich transcription of french broadcast news. In *In proceedings Eurospeech Interspeech*, Lisboa, 2005.

[4] H. M. V. Digalakis, P. Monaco. Genones : Generalized mixture tying in continuous hidden markov model- based specch recognizers. In *IEEE Transactions on Speech and audio Processing p 294-300*, 1996.