A SIMPLIFIED SUBSPACE GAUSSIAN MIXTURE TO COMPACT ACOUSTIC MODELS FOR SPEECH RECOGNITION

Mohamed Bouallegue, Driss Matrouf, Georges Linares

LIA, University of Avignon, France

mohamed.bouallegue@etd.univ-avignon.fr
driss.matrouf@univ-avignon.fr,georges.linares@univ-avignon.fr

ABSTRACT

Speech recognition applications are known to require a significant amount of resources (memory, computing power). However, embedded speech recognition systems, such as in mobile phones, only authorizes few KB of memory and few MIPS. In the context of HMM-based speech recognizers, each HMM-state distribution is modeled independently from to the other and has a large amount of parameters. In spite of using state-tying techniques, the size of the acoustic models stays large and certain redundancy remains between states. In this paper, we investigate the capacity of the Subspace Gaussian Mixture approach to reduce the acoustic models size while keeping good performances. We introduce a simplification concerning state specific Gaussians weights estimation, which is a very complex and time consuming procedure in the original approach. With this approach, we show that the acoustic model size can be reduced by 92% with almost the same performance as the standard acoustic modeling.

Index Terms— Compact Acoustic Models, Subspace Gaussian Mixture, Embedded speech recognition, Gaussian Mixture Models, Hidden Markov Models

1. INTRODUCTION

Most of the state-of-the-art Continuous Speech Recognition Systems are based on Hidden Markov Models that represent elementary speech units, typically phonemes or triphones. Usually, the state-level probabilities are estimated by using Gaussian Mixture Models (GMM) which offers several advantages : a well established mathematical formalism, automatic parameter training and improved performance. To achieve a good performace, the number of states and hence the number of parameters becomes more and more important (several tens of parameters). In this paper we deal with the problem of reducing the number of parameters while keeping good performances. Previous works deal with this problem, mainly with the purpose of reducing the memory footprint of acoustic models [1] [2]. Semi-continuous HMMs (SCHMM) are based on a Gaussian codebook that is shared between all HMM states, state models resulting from a specific weighting of the common Gaussian set [3]. This full Gaussian tying allows a significant reduction of model complexity but with a significant accuracy decrease [4]. Some authors extend this modeling by using compact transformation functions that map the Gaussian codebook to state-dependent GMMs [1].

In spite of the use of the state-tying technique, the size of the acoustic models stays large and certain redundancy remains between states. In this paper we propose to use the Subspace Gaussian Mixture (SGM) approach to allow a supplementary reduction of acoustic model size. All state GMMs are derived from a single GMM called GMM-UBM (Universal Background Model) with very small specific state parameters . This approach has some similarities to Eigenvoice [5] and cluster Adaptive Training [6]. In the SGM approach the specific state weights are estimated using a complex and very time consuming procedure. We replace this procedure using simple EM estimation and by keeping the N-best weights in each HMM state.

In Section 2, we recall the standard acoustic modeling [7]. In Section 3, we describe the proposed approach for compacting the acoustic models using the SGM approach sketching the difference to other similar approaches. In Section 4, we present some experimental results. And finally, in Sections 5 conclusions are proposed.

2. THE STANDARD ACOUSTIC MODEL STRUCTURE FOR SPEECH RECOGNITION

The baseline system adopted in this work uses a left-to-right HMM architecture of 10,002 context-dependent phonemes. To reduce the size of the acoustic models, we used the principle of state clustering, where we replace two or more HMM states having very similar data (GMM parameters are similar) with a single state by ”tying” these states together [8]. This technique allows us to model 10,002 context-dependent phonemes by 3327 states instead of 30,006, with 64 Gaussians per state and 39 PLP (Perceptual Linear Predictive) coefficients per frame (13 static with first and second deriva-

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

tives). The size of the acoustic models is still large and some redundancy remains between states even though state-tying technique has been used. In the following section we will describe the use of the SGM approach to reduce this redundancy and hence to reduce the size of the models.

3. SGM FOR COMPACT ACOUSTIC MODELING

In the SGM modeling approach all GMM-states are derived from the same global GMM, the Universal Background Model (UBM) : UBM=$(\alpha_g, m_g, \Sigma_g)$, where α_g, m_g and Σ_g are the weight, the mean and covariance matrix of the g^{th} Gaussian. let m be the mean super-vector obtained by concatenating all Gaussian means. In the SGM modeling the mean super-vector random variable of the state s is obtained as follow:

$$\mathbf{m}_{(s)} = m + \mathbf{U}\mathbf{x}_{(s)} \tag{1}$$

where $\mathbf{m}_s$ is the state dependent mean super-vector (random variable vector), $\mathbf{U}$ is the inter-states variability matrix (a $MD \times R$ matrix) of low rank R . Where M is the number of Gaussians in the UBM and D the is cepstral feature size. $\mathbf{x}_s$ are the state factors, an R vector. $\mathbf{x}_s$ are assumed to be normally distributed among $\mathcal{N}(0, I)$. In the training phase, the $\mathbf{U}$ matrix is estimated on all training data and the MAP point estimate $x_{(s)}$ of $\mathbf{x}_{(s)}$ is estimated on for each state [9].

In the SGM modeling [10], the specific Gaussian weights are obtained from the UBM as follow :

$$w_g^s = \frac{exp\ \mathbf{w}_g^T \mathbf{x}_s}{\sum_{g'=1}^{M} exp\ \mathbf{w}_{g'}^T \mathbf{x}_s} \tag{2}$$

where $\mathbf{x}_s \in \Re^R$ is the "state vector" with R the subspace dimention. $\mathbf{w}_g$ is weight vector depending on the Gaussian but not on the state.
This state specific weights estimation is difficult to perform. Its derivation is based on a combination of Jensen-type inequalities, local second-order Taylor-series expansions, and a modification to the resulting quadratic auxiliary function which ensures stability while maintaining the same local gradient [11]. As an alternative, we propose to re-estimate the gaussian weights by simple iteration EM (Expectation-maximization algorithm). Its derivation is further more simplified and allows us to gain computing time. Let w_g be the weight of the Gaussian g in the UBM. The weight w_g^s for that Gaussian in the state s is calculated as follows:

$$w_g^s = \frac{\sum_{x \in s} P(g|x)}{N_s} \tag{3}$$

where,

$$P(g|x) = \frac{w_g * f(x|g)}{\sum_{g'} w_g * f(x|g')} \tag{4}$$

$N_{(s)}$ is the number of frames of the to HMM-state (s), $P(g|x)$ is the a *posteriori* probability of the Gaussian g given the frame x and $f(x|g)$ is the likelihood for the frame x given the Gaussian g.

For further reduction of the number of parameters in a state GMM, for a state s we select the N-best Gaussians having the largest weights and we ignore the other ones. **N** is chosen in such a way that the sum of the weights of the selected Gaussians reach a predefined threshold. In our experiments, 333,000 Gaussians were selected , which represent an average of 100 selected Gaussians per state.

3.1. Estmation of U and latent variable $\mathbf{x}_s$

Let $\mathbf{N}_{(s)}$ and $\mathbf{X}_{(s)}$ be vectors containing the zero order and first order state statistics respectively :

$$\mathbf{N}_s[g] = \sum_{t \in s} \gamma_g(t);\ \{\mathbf{X}_{(s)}\}_{[g]} = \sum_{t \in (s)} \gamma_g(t) \cdot t \tag{5}$$

where $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g for the observation t. In the equation $\sum_{t \in s}$ means the sum over all frames belonging to the state s.

In the following, the estimation of the state vector are given. Let $\overline{\mathbf{X}}_{(s)}$ the state dependent statistics defined as follows:

$$\{\overline{\mathbf{X}}_{(s)}\}_{[g]} = \{\mathbf{X}_{(s)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{h \in s} \mathbf{N}_{(s)}[g] \tag{6}$$

Let $\mathbf{L}_{(s)}$ be $R \times R$ matrix, and $\mathbf{B}_{(s)}$ a vector of dimension R, both defined as:

$$\begin{aligned} \mathbf{L}_{(s)} &= \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(s)}[g] \cdot \{\mathbf{U}\}_{[g]}^t \cdot \boldsymbol{\Sigma}_{[g]}^{-1} \cdot \{\mathbf{U}\}_{[g]} \\ \mathbf{B}_{(s)} &= \sum_{g \in \text{UBM}} \{\mathbf{U}\}_{[g]}^t \cdot \boldsymbol{\Sigma}_g^{-1} \cdot \{\overline{\mathbf{X}}_{(s)}\}_{[g]}, \end{aligned} \tag{7}$$

By using $\mathbf{L}_{(s)}$ and $\mathbf{B}_{(s)}$, $\mathbf{x}_{(s)}$ can be obtained by using the following equation:

$$\mathbf{x}_{(s)} = \mathbf{L}_{(s)}^{-1} \cdot \mathbf{B}_{(s)} \tag{8}$$

The matrix $\mathbf{U}$ can be estimated line by line, with $\{\mathbf{U}\}_{[g]}^i$ being the i^{th} line of $\{\mathbf{U}\}_{[g]}$ then:

$$\mathbf{U}_{[g]}^i = \mathbf{LU}_g^{-1} \cdot \mathbf{RU}_g^i, \tag{9}$$

where $\mathbf{RU}_g^i$ and $\mathbf{LU}_g$ are given by:

$$\begin{aligned} \mathbf{LU}_g &= \sum_s \sum_{h \in s} (\mathbf{L}_{(s)}^{-1} + \mathbf{x}_{(s)} \mathbf{x}_{(s)}^T) \cdot \mathbf{N}_{(s)}[g] \\ \mathbf{RU}_g^i &= \sum_s \sum_{h \in s} \{\overline{\mathbf{X}}_{(s)}\}_{[g]}[i] \cdot \mathbf{x}_{(s)} \end{aligned} \tag{10}$$

The algorithm 1 presents the adopted strategy to estimate the state variability matrix with the above developments (the standard likelihood function can be used to asses the convergence).

Algorithm 1: Estimation algorithm of **U** and latent variable **x**.

```
For each state $s$ : $\mathbf{x}_{(s)} \leftarrow 0$, $\mathbf{U} \leftarrow random$ ;
Estimate statistics: $\mathbf{N}_{(s)}$, $\mathbf{X}_{(s)}$ (eq.5);
for $i = 1$ to $nb\_iterations$ do
    for all $s$ and $h$ do
        Center statistics: $\overline{\mathbf{X}}_{(s)}$ (eq.6);
        Estimate $\mathbf{L}_{(s)}$ and $\mathbf{B}_{(s)}$ (eq.7);
        Estimate $\mathbf{x}_{(s)}$ (eq.8);
    end
    Estimate matrix $\mathbf{U}$ (eq. 9 and 10) ;
end
```

3.2. Compact Acoustic Models: Subspace vs SCHMM

A semi-continuous HMM (SCHMM), which can be considered as a special form of continuous mixture HMM, is based on a Gaussian codebook that is shared between all HMM states, state models resulting from a specific weighting of the common Gaussian set [5]. This full Gaussian tying allows a significant reduction of model complexity but with a significant accuracy decrease [6]. The main difference between the SGM acoustic modeling and the SCHMM is that in the first case both, means and weights are state dependent, whereas in the second case only Gaussian weights are state dependent.

To compare the proposed SGM approach with the SCHMM, we have to construct two systems using the two different approaches having the same number of parameters and compare their performances. The SGM system is based on acoustic models with 500 Gaussians in the UBM and with the **U** matrix having a rank of 88. In the SCHMM system, the UBM contains 600 Gaussians. By retraining only the top-100 Gaussians, we have to estimate the same number of parameters for the two systems, which corresponds to a 89% reduction of the number of free parameter.

Table 1 shows baseline results compared to SCHMM and SGM compact system. Using the proposed approach, we can see a relative gain of about 37.4% compared to the SCHMM and almost the same performances as the standard acoustic models (3% relative loss).

Radio broadcasts	Baseline	SCHMM	SGM compact system
RTM (2h)	35.51	44.31	35.81
RFI (1h30mn)	25.72	38.36	25.77
INFO (2h)	25.97	39.60	28.37
CLASSIQUE (1h)	21.7	34.5	22.4
Total (6h30)	28.06%	39.97%	29.09%
Size reduction	-	89%	89%

Table 1. *The results of baseline, SCHMM and the SGM compact system, in % Word Error Rate.*

4. EXPERIMENTAL FRAMEWORK AND RESULTS

4.1. The LIA broadcast system

Experiments are carried out by using the LIA broadcast news (BN) system which was used in the ESTER evaluation campaign. This system relies on the HMM-based decoder developed at the Laboratoire Informatique d'Avignon (LIA), Speeral [12]. Acoustic models are HMM-based, context dependent with cross-word triphones. The language models are classical trigrams estimated on about 200M words from the French newspaper Le Monde and about 1M words from the ESTER broadcast news corpus. The lexicon contains 67K words. In these experiments, only one decoding pass is performed in 3x Real Time.

The training parts of the data set are based on the training corpus provided for the ESTER evaluation campaign. The ESTER corpus consists of French radio broadcasts of the Radio-France group. We test our approach on 7.5 hours of speech extracted from the ESTER test set.

4.2. Results

In this section we will present the results obtained by our compact acoustic model. First we give the equations to calculate the number of parameters for the baseline system and for those of our new models. The number of parameters of the baseline system remains considerable even after state clustering. This number is calculated by Equation (11):

$$nbs * [\ \underbrace{(2 * nbp + 1)}_{\text{one Gaussian}} * nbg\ ] \tag{11}$$

Where nbg is the number of Gaussians (set to 64), nbs the number of emitting states (set to 3300), and nbp the number of acoustic features (set to 39). For our model, we have the parameters of the GMM-UBM, the inter-state variability matrix **U**, matrix **X** of state factors and the new weights. The parameters are calculated by the following equation.

$$\underbrace{(2 * nbp+1) * nbg}_{GMM-UBM} + \underbrace{ndg * nbp * R}_{U\ matrix} + \underbrace{nbs * R}_{X\ matrix} + \underbrace{nbs * n\_best}_{new\ weights} \tag{12}$$

where n_best is the number of selected Gaussians per state. In our experiments n_best is set to 100.

To study the impact of the acoustic model size variation on the performance of model we can vary the size of the UBM-GMM or the rank of the U Matrix. In our experiences we have chose to fix the size of the GMM in 600 gaussian and test serval ranks of U Matrix (40,100,150,190, and 250).

In Table 2 we show the results in terms of Word Error Rate (WER) on our test corpus. In the first column we find the different radio stations to which belong the test files. The second column shows the WER of the Baseline System and in the last five columns we find the WER of five different compact models. The last line indicates the percentage of reduction on the number of parameters compared to those of the Baseline system.

Radio broadcasts	Base	40	100	150	190	250
RTM (2h)	35.5	36.0	35.7	35.7	35.6	35.5
RFI (1h30mn)	25.7	26.5	26.2	25.8	25.7	25.3
INFO (2h)	25.8	28.3	27.7	27.6	27.1	27.1
CLASSIQUE (1h)	21.7	22.6	22.1	22.2	21.8	22.1
CULTURE (1h)	34.0	36.8	36.0	35.6	35.3	35.1
Total (7h30)	28.8 %	30.4 %	29.9%	29.7 %	29.5 %	29.4 %
Size reduction	-	92.1%	83.8%	76.9 %	70.9 %	63.0 %

Table 2. *Word Error Rate in % for different radio.*

The Table 2 shows that for some radio stations (RTM, RFI), we keep almost the baseline performance, while we have a negligible loss for the others (the global loss is between 2% and 5% relative). This negligible loss, despite the very significant reduction in the number of parameters to be estimated, shows the the capacity of the modified SGM to reduce the size of models while maintaining almost the same performance as the standard modeling.

5. CONCLUSION

In this paper we explored the capacity of the Subspace Gaussian Mixture to reduce the number of acoustic model parameters. We also propose a simplification concerning the estimation of the state specific weights : we use simple EM algorithm estimation. In our experiments, we showed 92% relative reduction in models size, while maintaining a similar performance (between 2% and 5% relative loss). With the gain in terms of model size, we can also improve the computation time, because likelihood calculation for each test frame can be limited to a small number of Gaussians driven by the UBM.

6. REFERENCES

[1] Christophe Lévy, Georges Linarès and Jean françois Bonastre, *"Compact acoustic models for embedded speech recognition"*, EURASIP Journal on Audio, Speech, and Music Processing,Vol9, pp. 806-186, 2009.

[2] Christophe Lévy, Georges Linarès, Pascal Nocera, and J-F Bonastre, *"Reducing computational and memory cost for cellular phone embedded speech recognition system*, In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Montrál, Canada, Vol. 5, pp. 309-312, 2004.

[3] J.Duchateau, K. Demuynck, D. Van Compernolle, *"A Novel Node Splitting Criterion in Decision Tree Construction for Semi-Continuous HMMs"*, Eurospeech'97, Rhodes, pp. 183-1186, 1997.

[4] T. Vaich and A. Cohen, *"Comparison of continuous density and semi-continuous hmm in isolated words recognition systems"* , EUROSPEECH'99, pp. 1515-1518, 1999.

[5] J.-C. Junqua, P. Nguyen, and N. Niedzielski, *"Rapid Speaker Adaptation in Eigenvoice Space"*, IEEE Transactions on Speech and Audio, Speech and Audio Processingn vol. 8, no. 6,Nov.2000.

[6] M. J. F.Gales, *"Multiple-cluster adaptive training schemes,"*, in ICASSP,2001.

[7] L. R. Rabiner and B. H. Juang, *"An introduction to hidden Markov models"*. IEEE Acoustics, Speech & Signal Processing Magazine, pp. 416, 1986.

[8] S. J. Young, J. J. Odell, and P. C. Woodland, *"Tree-based state tying for high accuracy acoustic modeling"*, Proc. ARPA Human Language Technology Workshop, pp. 307-312, 1994.

[9] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, Jean-Franois Bonastre, *"A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification"*, INTERSPEECH Conference 2007, Antewerp, Belgium, pp. 1242-1245, 2007.

[10] D. Povey, et al., *"Subspace Gaussian mixture models for speech recognition"*, In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10), March 14-19, 2010, Dallas, Texas, U.S.A, pp.1144-1160 , 2010.

[11] D. Povey, et al., *"A Tutoriel Introduction to Subspace Gaussien Mixture Models for Speech Recognition"*, Tech. Rep. MSR-TR-2009-111, Microsoft Research,2009.

[12] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, *"The LIA's french broadcast news transcription system"*, in SWIM: Lectures by Masters in Speech Processing, 2004.