UNSUPERVISED TRAINING OF HMMs WITH VARIABLE NUMBER OF MIXTURE COMPONENTS PER STATE

C. Martín, L. Villarrubia*, F.J.González**, L. Hernández***

*Telefónica I+D, Emilio Vargas 6 Madrid 28043, SPAIN
**ETSI Telecomunicación, Ciudad Universitaria s/n Madrid 28040, SPAIN
e-mail: cma@tid.es

ABSTRACT

In this work automatic methods for determining the number of gaussians per state in a set of Hidden Markov Models are studied. Four different mix-up criteria are proposed to decide how to increase the size of the states. These criteria, derived from Maximum Likelihood scores, are focused to increase the discrimination between states obtaining different number of gaussians per state. We compare these proposed methods with the common approach where the number of density functions used in every state is equal and pre-fixed by the designer.

Experimental results demonstrate that performance can be maintained while reducing the total number of density functions by 17% (from 2046 down to 1705). These results are obtained in a flexible large vocabulary isolated word recognizer using context dependent models.

1. INTRODUCTION

There are a number of studies about how to train Hidden Markov Models so that their performance in a Speech Recognition System [1] is increased. However, although the designer uses an algorithm that optimizes the models, there is still a point of non-optimality: the topology.

Traditionally, designers have had to rely on their experience to decide the topology of the models they used. Usually, the designer establishes the number of states and iteratively the models are trained with 1 gaussian per state, with 2 gaussians, etc. This kind of training gives the same number of density functions per state for all the models. Nevertheless, the designer can fix the number of gaussians used in each state according with *a priori* knowledge criteria.

Some research works have been reported looking for a method to train not only those parameters which are typically considered (means, variances, transition probabilities), but also to train the topology of the models. That is the case of the number of states and the number of density functions used in each state.

The problem designers have to deal with is to obtain a compromise between a good resolution in modeling the underlying distributions and a reasonable number of parameters such that they can be reliably estimated. There can also be the limitation in number of gaussian densities imposed by maximum computational perplexity allowed in the system. Even though there are enough data to train a huge number of density functions to finely fit the real distributions the final system can not afford so many functions due to real-time limitations.

Usually, to face this situation, the number of mixture components is increased up to a fixed number, which is common to all the states of the models. Another possibility is to use the "tied-mixtures" approach. In such a case, the same set of mixture components is shared by all the models, or by a subset of them.

Intuitively, we can see that the optimum situation would be to use so many mixture components as needed per state, but no more. The goal is to obtain the "correct" number of gaussians for each state of the set of models, instead of using the same number for all of them. This is the purpose of our present work.

There have been some approximations to this problem, as the one proposed by Normandin [2], or V. Valtchev [3], who used Maximum Mutual Information Estimation (MMIE) to successively split gaussians. Fissore [4] increased the number of mixture components until it reached a pre-set maximum value or the average likelihood of the observations in a evaluation subset decreased.

A new automatic method to obtain variable number of mixture components per state based on ML is presented in this work.

2. REFERENCE SYSTEM

All the experiments in this work were carried out on the VESTEL database [5]. We trained 288 left biphones, modeled with 3-state HMMs, which makes 864 different states. The training set included 5828 files containing digits, names, cities and commands. Four different sets of files were used to test. All of them contain only Spanish names and surnames with high acoustic similarity.

Set	Perplexity	Vocabulary independent	Files
DV448	448	no	2944
IV955	955	yes	1683
IV1573	1573	yes	3037
IV2000	2000	yes	3037

Table 1: Recognition sets

The states were all initialized to 1 gaussian, and reestimated. Then, they were tied by bottom-up data-driven clustering [6] in order to reduce the number of states to 341. This tying stayed fixed across all the different tests.

The reference system was trained with the classical procedure, that is, incrementing in every iteration the number of mixture components in all the states. Results for various numbers of mixture components were obtained.

The reference system results are summarized in Table 2. The recognition results improve as long as the number of mixture components is increased. It can be observed that the improvement is more significant when the number of gaussians is low. In fact there is almost no difference between models with 6 or 7 gaussians per state. This means that this number of components is sufficient to adequately train our system with the given training database. This point can be considered as an upper limit.

3. UNSUPERVISED MIX-UP TRAINING PROCEDURE

The basic flow of Unsupervised Mix-Up Training Procedure is represented in Figure 1 and can be explained as follows:

1. The mix-up method starts with a set of models, all of them with one gaussian distribution per state.
2. The models are trained with Maximum Likelihood (Baum-Welch) in order to estimate the parameters that define every model.
3. The training database is aligned using the current models so that boundaries and average frame probability are obtained for every state of the models in the database.
4. According to the selected mix-up criterion, some states increase the number of mixture components, that is, some gaussians are split. Besides, a minimum number of frames per mixture is required to avoid the problem of undertraining.
5. If a maximum number of gaussian distributions is reached stop the training, otherwise back to step 2.

The alignment is made on the state level, since they are the minimal elements we are considering to change the number of mixture components. This alignment is obtained by applying recognition to every file with a special grammar, which forces to recognize exactly what is said in the file being processed. Similar procedures can be applied to more complex parts of the models (e.g. the whole model), or to a set of them.

In previous works it has been shown that using a splitting criterion based on the local behavior of each mixture component (as the mixture weight count [2]) does not work properly. So a global behavior of the whole set of mixtures of each state is considered in the present contribution. The motivation of this approach is to obtain the best possible modeling of a state by using the correct number of gaussians.

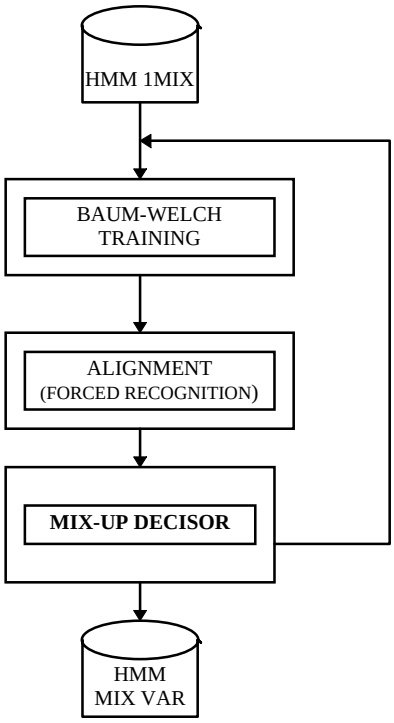

Figure 1. Variable mixture training procedure scheme

3.1 Frame Probability Average

The first and most intuitive splitting criterion is the standard *frame probability average (FPA)* of each state. To obtain this measure, the total probability given by each segment assigned to the considered state is normalized by the number of frames it lasts.

$$FPA = \frac{1}{NO_s} \sum_{\substack{\forall r: \\ O^r \text{ assigned to } s}} \frac{1}{length(\mathbf{O}^r)} \sum_{i=1}^{length(\mathbf{O}^r)} P\left(\mathbf{o}_i^r | s\right)$$

Eq. 1

Where NO_s is the number of examples of state s (times it appears), no matter the length of every one. $\mathbf{O}^r$ represents the r^{th} observation of the state s, which is composed of a number of frames, given by the Viterbi-based segmentation.

$$\mathbf{O}^r = \{\mathbf{o}_i^r\} \qquad i = 1,2,\ldots,length(\mathbf{O}^r)$$

Eq. 2

Then, according to this measure, in the K worst states the number of mixture components is incremented by one. The number K may be chosen so that a fixed percentage of the states is incremented or, on the other hand, it can be derived in each iteration from a threshold from the worst state.

In Table 2 results are presented for this criterion. It can be seen that the performance of the Frame Probability Average is worse than in the reference system.

A close analysis of this training procedure revealed that the main problem related to this splitting criterion is that this score does not distinguish between states with similar Frame Probability Average but very different number of training examples. This fact produces that states with not many examples can be chosen to be split instead of others with slightly higher score but much more examples. Another effect detected was that some states with similar *FPA* and number of occurrences showed very different improvements after increasing the number of gaussians.

Trying to overcome with these problems some different measures are proposed. The first idea is to use the improvement of the scores instead of the scores themselves. Besides, the average of all the training examples of one state is obtained, without normalizing by the number of frames.

3.2 Delta of Frame Probability Average

All the states are incremented in the first iteration. In successive iterations difference on Frame Probability Average on iteration *t* and *t-1* is taken. That is:

$$\Delta FPA_s(t) = FPA_s(t_u) - FPA_s(t_u - 1)$$

Eq. 3

Now, a new gaussian density function is added in the *K* states with best score. The delta score is obtained for a given state only when a new gaussian is added to it, and that is the quantity used until that ΔFPA_s becomes one of the *K* best scores. To remark this, the index t stands for the current iteration, while t_u stands for the last iteration in which the given state was modified.

The results for this measure are included in table 2. Although this training criterion outperforms the Frame Probability Average criterion, it does not reach the reference system. Therefore, we moved to the State Probability Average.

3.3 State Probability Average

Once the boundaries are obtained, for every state, the average probability per observation is calculated. That is, every time a specific state appears in the aligned training database, the total probability obtained for that segment is accumulated as Eq. 4.

$$SPA = \frac{1}{NO_s} \sum_{\substack{\forall r: \\ \mathbf{O}^r \text{ assigned to } s}} \sum_{i=1}^{length(\mathbf{O}^r)} P\left(\mathbf{o}_i^r \middle| s\right)$$

Eq. 4

Again, the *K* worst states are incremented, with the same considerations applicable to *K*. Table 2 shows that State Probability Average achieves a performance which falls below the reference system. However, it can be seen that it depends greatly on the initialization. When initialized with 2 gaussians per state (see * in table 2) the system obtains a significant improvement, although not being able to equal the reference system. This fact reveals that this measure relies excessively on the number of frames each state is assigned to. Thus, an important number of states with relatively few assigned frames remains unsplit. So, when forcing a minimum number of two gaussians, the performance significantly improves.

3.4 Delta of State Probability Average

In this case, the delta scoring is applied to State Probability Average, with the same considerations as in ΔFPA applicable to the method for obtaining the delta score. That is:

$$\Delta SPA_s(t) = SPA_s(t_u) - SPA_s(t_u - 1)$$

Eq. 5

In every iteration, the *K* states with best scores are incremented. This criterion, as ΔFPA does, imposes at least two gaussians in every state. The so trained system outperforms the reference system in almost all the cases. In fact, results show that performance can be maintained while reducing the total number of density functions by 17% (from 2046 down to 1705).

In Figure 2 and Figure 3 we can see the resulting distribution of gaussians among the states for the case of 4 and 7 gaussians per state ratio. As we might hope, there are a number of states that do not need more than 2 or 3 mixture components to be well defined, so it is more efficient to distribute those density functions among the rest of the states.

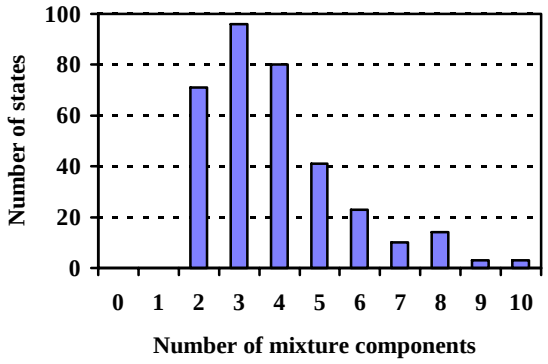

Figure 2 Distribution of gaussians. 4 gauss/state ratio

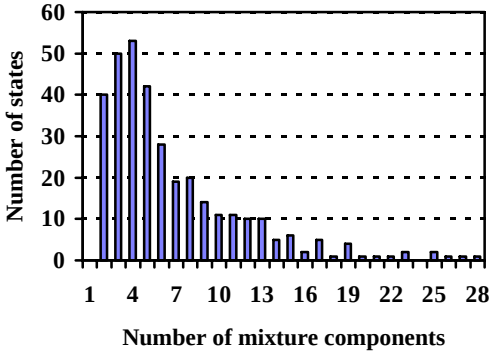

Figure 3. Distribution of gaussians. 7 gauss/state ratio.

4. SUMMARY AND FUTURE WORK

In this work, we have presented different approaches to train HMMs with variable number of mixture components per state. Four splitting criteria were used trying to efficiently model the state real distribution. All of these methods are based on the Maximum Likelihood Estimation criterion.

The experiments show that the criterion that best accomplishes with this task is the one based on Delta of State Probability Average, which manages to maintain the performance although the global number of mixtures of the whole set of models is reduced.

That is, the hint is to improve the match between the whole state and its real distribution, instead of improving each gaussian locally. We found also that it is important to increment the number of mixtures specially in those states that have improved most since the last increment. Anyway, minimum modeling detail is necessary to guarantee a good enough performance.

Now, we are in the process of evaluating this training method on triphones-based speech recognition system. Also, new measures are being studied to model more efficiently the state distributions.

5. REFERENCES

1. L. Rabiner and B-H Juang, "Fundamentals of Speech Recognition, Prentice Hall, 1993.

2. Normandin Y., "Optimal Splitting of HMM Gaussian Mixture Components with MMIE Training", Proc. ICASSP-95, pp. 449-452.

3. V. Valtchev, P.C. Woodland, S. J. Young, "Discriminative optimization of large vocabulary recognition system". In Proc. ICSLP'96. pp. 18-21.

4. L. Fissore, F. Ravera, P. Laface, "Acoustic-phonetic Modeling for Flexible Vocabulary Speech Recognition". In Proc. EUROSPEECH 95, pp. 799-802

5. D. Tapias, A. Acero, J. Esteve, J.C. Torrecilla, "The VESTEL Telephone Speech Database". In Proc. ICSLP'94. pp. 1811-1814.

6. S. J. Young, P. C. Woodland, "The use of state tying in continuous speech recognition". In Proc. EUROSPEECH'93, pp. 2203-2206.

Criterion	Total # of Mix.	ratio mix/state	Recognition rate per Set (%)			
			DV448	IV955	IV1573	IV2000
Fixed number of mix. comp.	1364	4 (all)	91.30	88.71	84.72	82.88
	1705	5 (all)	91.54	89.54	85.81	84.00
	2046	6 (all)	92.05	89.78	86.04	84.29
	2387	7 (all)	92.05	89.60	86.14	84.62
Frame Prob. Average	1364	4.00	90.90	88.35	83.73	81.89
Delta Frame Prob. Av.	1000	2.93	90.66	88.29	84.29	82.52
	1357	4	91.17	88.59	84.85	82.98
State Prob. Av.	1367	4.00	88.01	84.55	80.11	78.14
	1415	4.15	87.87	84.79	80.14	78.27
	1414 *	4.15	90.12	87.70	83.77	81.99
Delta of State Prob. Av.	1357	4.00	91.81	89.42	85.61	84.03
	1415	4.15	91.88	89.30	85.71	84.06
	1705	5.00	91.98	89.78	86.10	84.62
	2011	5.89	92.46	89.66	86.47	84.66
	2387	7	92.60	89.78	86.24	84.59

Table 2. Results for different criteria and number of mixture components. Asterisk (*) means that the system was initialized with 2 gaussian density functions, instead of one function.