A NEW APPROACH TO GENERALIZED MIXTURE TYING FOR CONTINUOUS HMM-BASED SPEECH RECOGNITION

Daniel Willett, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: {willett,rigoll}@fb9-ti.uni-duisburg.de

ABSTRACT

In this paper we present a new approach for a generalized tying of mixture components for continuous mixture-density HMM-based speech recognition systems. With an iterative pruning and splitting procedure for the mixture components, this approach offers a very accurate and detailed representation of the acoustic space and at the same time keeps the number of parameters reasonably small in favor of a robust parameter estimation and a fast decoding. Contrary to other approaches, it does not require a strict clustering of the pdfs into subsets that share their mixture components, so that it is capable of providing more general and flexible types of mixture tying. We applied the new approach on a semi-continuous HMM (SCHMM)-system for the Resource Management task and improved its recognition performance by 12% and vastly accelerated the decoding because of a much faster likelihood computation.

1. INTRODUCTION

In continuous mixture-density HMM-based speech recognition systems the HMM states' pdfs are modeled as weighted sums (mixtures) of primitive basic functions like Gaussians or Laplacians.

$$p(\mathbf{x}|w) = \sum_{i=1}^{C_w} d_{wi} f_{wi}(\mathbf{x})$$

In the equation above w denotes the HMM state, d_{wi} are w's mixture weights and C_w resembles the number of basic functions used in the pdf of state w. In this context the basic functions $f_{wi}(\mathbf{x})$ are called mixture components. Mixture components that are used in more than one pdf are said to be tied among those pdfs. The reason for tying parameters in general is to find a good tradeoff between the system's acoustic resolution and the robustness of the estimated parameters. Although there are a lot of other possible ways of parameter tying [5], the tying of mixture components has been one of the major points of discussion in the field of Hidden Markov Model-based speech recognition during the last years. The two extrema of having no tying of mixture components at all (continuous HMM [1]) and of having a single set of components that are used by all pdfs (semi-continuous HMM [1, 2]) are the most popular choices of tying. Nevertheless, other approaches like phonetically tied HMM [3] or Genones [4] proved that the optimal tying is located somewhere within "the continuum between fully continuous and tied-mixture HMMs" [4].

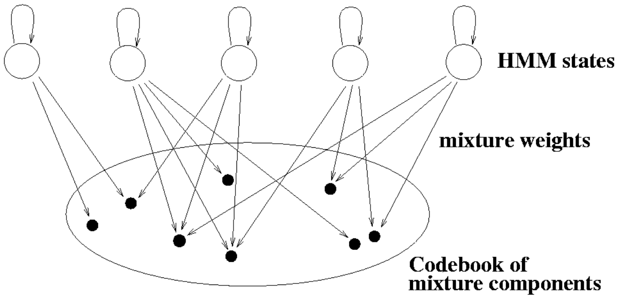

Figure 1. Generalized tying of mixture components

Of course, however, this optimal tying is largely task-dependent. The more training data is available for each pdf, the better a more independent realization of the pdfs like in continuous HMM (CHMM)-systems turned out to be.

As far as the decoding speed is concerned the tying of mixture components usually has positive effects, as evaluated mixture components can be used in several likelihood computations. However, once the tying is as exaggerated as in semi-continuous systems the extremely large number of weights per HMM state (which equals the total codebook size) results in an expensive likelihood computation as well.

Thus, in respect of the decoding speed, it seems to be important on the one hand to have a small total number of mixture components and on the other hand to have as few mixture components per pdf as possible. From this point of view CHMM and SCHMM are no good choices of tying at all. Because of that, a lot of publications like [4, 6, 7] deal with the problem of a fast likelihood approximation for such with regard to the decoding speed inconveniently tied systems. Unfortunately though, a fast likelihood approximation always comes along with at least a slight degradation in recognition performance. And on the whole, it does not seem to be very reasonable to build and train a system with a very detailed acoustic resolution and then in the recognition procedure to give up the accuracy in favor of a fast decoding.

And even 'Genones', as proposed by Digalakis et. al. [4], that have several codebooks of mixture components with each HMM state assigned to exactly one of them, are only suboptimal with respect to the structure of tying. The strict clustering of the states into those that do share mixture components and those that do not is certainly somewhat arbitrary. This way of tying lacks the possibility of having states that have some independent

mixture components to describe very individual acoustic features while sharing with other states some components that describe more common features.
In order to overcome the inconveniences of the common types of mixture tying, the following section will describe the procedure that we developed to create an optimized tying of mixture components, that achieves a very good tradeoff between acoustic resolution and robustness and that does not require a likelihood approximation for a fast decoding.

2. PROCEDURE TO CREATE AN OPTIMIZED TYING

As a superset of all common and all other possible types of mixture tying one can think of a global set of mixture components like in SCHMM systems that each pdf is using its own subset of. This general structure is illustrated in figure 1. As explained in the introduction the aim is to find such a general structure that has a small total number of mixture components (a small global set of mixture components) and a small number of used components per pdf (small subsets), that nevertheless provides a high acoustic resolution.

2.1. Basic procedure

The basic approach for the construction of a tying with a high acoustic resolution and a reasonably small number of free parameters is displayed in figure 2. In short, the idea is to start from a (trained) SCHMM system that has a small codebook and to remove those tyings with small weight values and then to moderately raise the number of mixture components by splitting components in order to maximize the observations‘ likelihood, until for instance the improved acoustic resolution does not improve the recognition performance on an independent reference test anymore.
These sub-procedures are described in more detail in the following sections.

2.2. Removal of tyings

When looking at a fully tied SCHMM systems, the tyings that can be cut with a minimum loss of recognition accuracy are those with the smallest weight values. In a system of 200 Gaussian mixture components, we observed that up to 90% of the tyings can be removed without dramatically reducing the recognition performance. A further re-estimation of the remaining weights can provide even more robust estimates for the remaining weights and improved recognition performance as the number of parameters has drastically been reduced.
The removal process can be accomplished in many ways. We found it to be most practical not to remove a fixed number of tyings per pdf, but to remove them in a way that the removed weights' sum does not exceed a certain limit. After removing tyings the remaining weights have to be re-normalized to sum up to unity again.

2.3. Split of mixture components

Using large codebook sizes in SCHMM-systems poses several problems. On the one hand it increases the computational effort of the parameter estimation process, on the other hand the more mixture components there are, the less robust are the estimated weights especially for

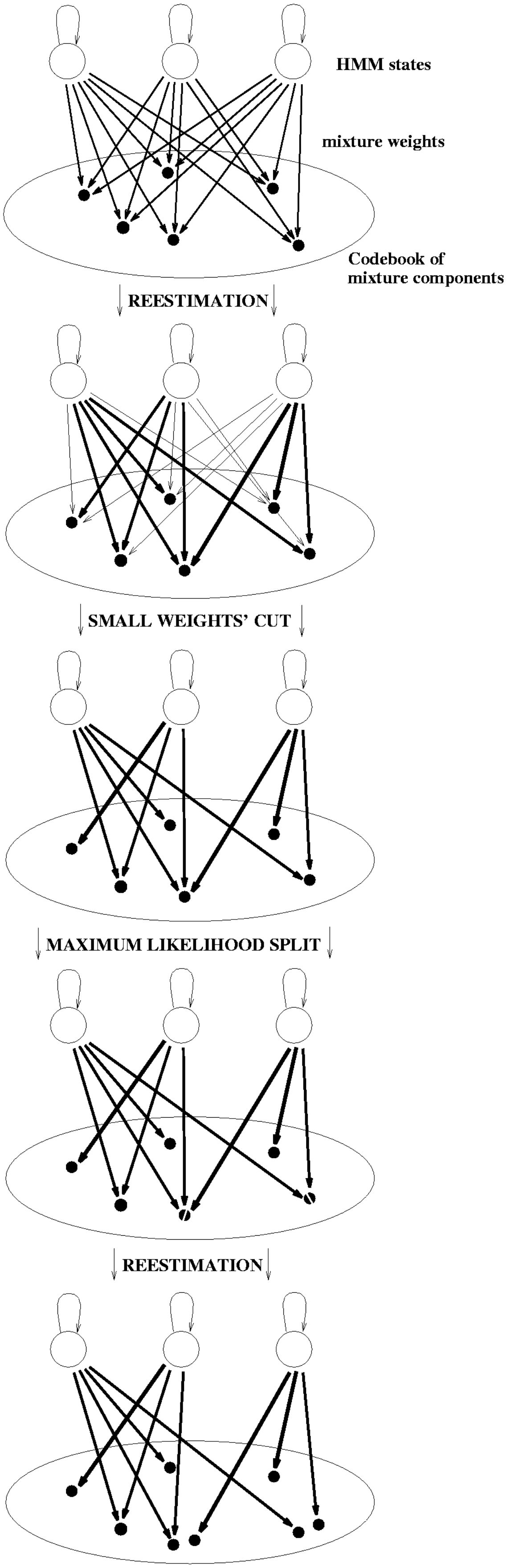

Figure 2. Steps to construct an optimized tying structure from a SCHMM system

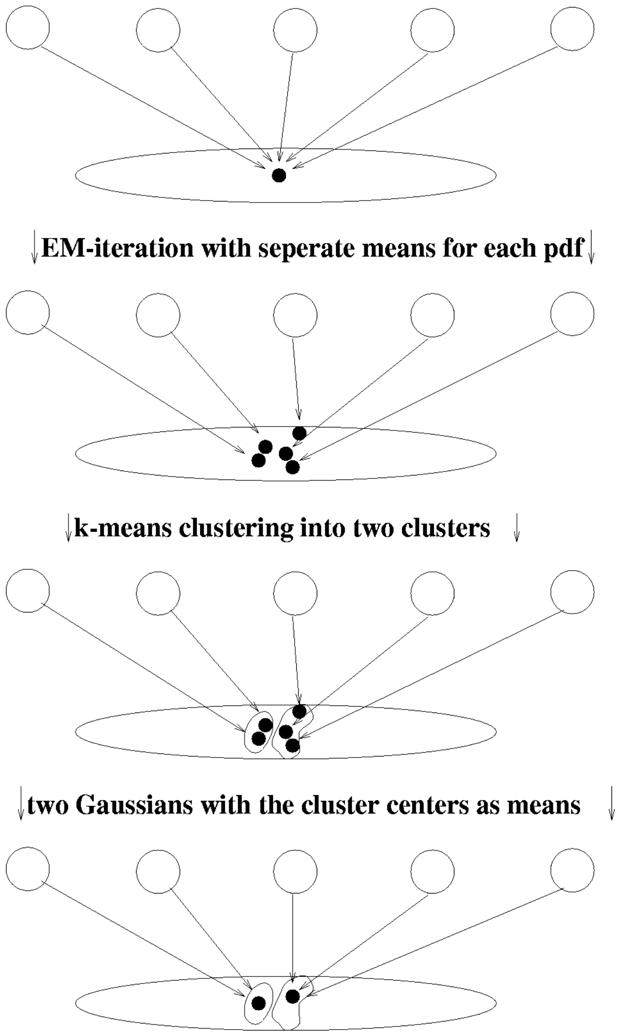
EM-iteration with seperate means for each pdf
k-means clustering into two clusters
two Gaussians with the cluster centers as means

Figure 3. Splitting a Gaussian mixture component

less represented HMM states. For these reasons we propose to have a baseline SCHMM system of moderately few mixture components and to increase this number by splitting mixture components in an additional procedure. The splitting of mixtures, too, can be performed in various ways. We chose the following Maximum Likelihood approach:

The Gaussians' means are re-estimated separately for each pdf and each mixture component with the EM-algorithm. For each Gaussian these means are clustered into two clusters with the k-means algorithm, and then those splits are performed that offer the maximum likelihood increase for the training observations.

Each of the involved pdfs is tied to exactly one of the two resulting Gaussians. This way it is possible to increase the total number of mixture components without increasing the number of tied mixture components per state.

The splitting procedure is illustrated in figure 3. A more discriminative splitting procedure would be possible as well. This procedure would split the Gaussians in order to maximize a discriminative Maximum Mutual Information (MMI) criterion like one of those formulated in [8]. Unfortunately though, this would forbid a further Maximum Likelihood re-estimation in order not to lose this discrimination again.

	baseline SCHMM	generalized tying of mixture components
no. of Gaussians per stream	200	400
av. no. of weights per pdf and stream	200	20
no. of streams	4	4
no. of HMM states	3500	3500
no. of parameters	2,800,000	310,000

Table 1. System comparison

	baseline SCHMM	generalized tying of mixture components
February'89	5.0%	4.3%
October'89	5.4%	5.2%
February'91	4.3%	3.8%
September'92	8.6%	7.2%
Average	5.8%	**5.1%**
Error reduction		**12.1%**

Table 2. Word error rates on the RM database

3. EXPERIMENTS AND RESULTS

We applied the proposed approach to improve a SCHMM system for the Resource Management task. The system uses linear word-internal triphones of three emitting states each, 39 features per 10ms-frame (a 12-value Cepstrum, log energy, and these values' first and second derivatives) in four independent streams and the standard wordpair-grammar of perplexity 60. The HMM-states are clustered in a tree-based phonetical clustering procedure. We reduced the number of tied components per pdf in each stream from 200 to an average of 20 and then enlarged the codebook of each stream from 200 to 400 components by mixture splitting as explained in section 2. The recognition results were obtained with a standard Viterbi beam-search and a linear lexicon organization. We measured the decoding times on a PentiumPro-200 PC. The word error rates and the average decoding time per test-set can be found in tables 2 and 3.

The system with the optimized tying achieves remarkably lower error rates and its decoding time is less than half of the conventional SCHMM system. (The decoding was performed using a Viterbi decoder with linear lexikon organization and a beam width of 150.) In a separate experiment we measured that the pure likelihood computation with the optimized tying is about four to five times faster compared to the SCHMM system. The average

	baseline SCHMM	generalized tying of mixture components
likelihood computations	90 min	21 min
relative speed up		**430%**
decoding time	114 min	45 min
relative speed up		**250%**

Table 3. Time consumed for likelihood computations and decoding

error rate of 5.1% achieved with the optimized tying is among the best ever observed on the RM database. We believe, it is the first system with such a small number of parameters to achieve such remarkable results.

4. CONCLUSION

The paper has demonstrated how to find an appropriate structure for tying the mixture components in continuous mixture-density HMM-based speech recognition systems. The presented experiment has proven that the new approach is capable of providing an optimized tying that achieves high recognition accuracy while keeping the computational costs of the likelihood computations reasonably low. Contrary to other forms of mixture tying this new type of tying does not require a likelihood approximation for a fast decoding, so that it is capable of providing an accurate acoustic resolution without the need for giving up this resolution in favor of a fast decoding. With the small number of parameters that is needed to achieve a very good recognition accuracy, we believe that the optimized tying as proposed in this paper is the best choice of tying with respect to the recognition accuracy as well as with respect to the computational complexity caused by the likelihood computations.

ACKNOWLEDGMENTS

This work was partly sponsored by the DFG (German Research Foundation) under contract number Ri 658/3-1.

5. REFERENCES

[1] X. D. Huang, Y. Ariki, M. A. Jack: "Hidden Markov Models for Speech Recognition", Edinburgh University Press, 1990.

[2] X. D. Huang, K. F. Lee, H. W. Hon: "On Semi-Continuous Hidden Markov Models", Proc. ICASSP'90, pages 689-692.

[3] D. B. Paul: "The Lincoln Robust Continuous Speech Recognizer", Proc. ICASSP'89, pages 449-452.

[4] V. V. Digalakis, P. Monaco, H. Murveit: "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognition", IEEE Transactions on ASSP, vol. 4, 1996, pages 281-289.

[5] S. J. Young: "The General Use of Tying in Phoneme-Based HMM Speech Recognisers", Proc. ICASSP'92, pages 569-572.

[6] E. Bocchieri: "Vector Quantization for the Efficient Computation of Continuous Density Likelihoods", Proc. ICASSP'93, pages 692-695.

[7] T. Watanabe, K. Shinoda, K. Takagi, K. Iso: "High Speed Speech Recognition using Tree-Structured Probability Density Function", Proc. ICASSP'95, pages 556-559.

[8] D. Willett, C. Neukirchen, J. Rottland: "Dictionary-Based Discriminative HMM Parameter Estimation for Continuous Speech Recognition Systems", Proc. ICASSP'97, pages 1515-1518.