IMPROVED PARAMETER TYING FOR EFFICIENT ACOUSTIC MODEL EVALUATION IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

*Jacques Duchateau, Kris Demuynck, Dirk Van Compernolle * and Patrick Wambacq*

Katholieke Universiteit Leuven - E.S.A.T.
Kardinaal Mercierlaan 94
B-3001 Heverlee, Belgium
E-mail: Jacques.Duchateau@esat.kuleuven.ac.be

ABSTRACT

In an HMM based large vocabulary continuous speech recognition system, the evaluation of - context dependent - acoustic models is very time consuming.

In Semi-Continuous HMMs, a state is modelled as a mixture of elementary - generally gaussian - probability density functions. Observation probability calculations of these states can be made faster by reducing the size of the mixture of gaussians used to model them.

In this paper, we propose different criteria to decide which gaussians should remain in the mixture for a state, and which ones can be removed. The performance of the criteria is compared on context dependent tied state models using the WSJ recognition task. Our novel criterion, which decides to remove a gaussian in a state if it is based on too few acoustic data, outperforms the other described criteria.

1. INTRODUCTION

In Continuous Density HMMs (CD-HMMs), a state is modelled as a mixture of a state specific set of elementary probability density functions (*pdfs*). All parameters in this model (both the densities and the weights for the densities in the mixture) have to be estimated with the - limited - acoustic data available for that state. This limits the number of states and/or the mixture size for each state that can be used in CD-HMMs.

A problem in CD-HMMs is that, as states overlap in the acoustic space, essentially the same density will be reestimated for different states. This can be avoided by tying the densities. In Semi-Continuous Density (or tied density) HMMs (SC-HMMs) [1, 2], all states are modelled as a mixture of a single large set of elementary *pdfs*. The distinction between states is made by the weights for each density in the mixture. As only these weights have to be modelled with the acoustic data available for the state, more different states or a larger mixture size can be used in SC-HMMs, this way providing more accurate acoustic modelling.

The evaluation of *full* (context dependent) SC-HMMs however, in which all gaussians are used in all states, is prohibitively time consuming. Therefore, different methods are proposed in literature to create HMMs with a density tying degree between the two extremes (CD-HMMs and full SC-HMMs).

- **Phonetically tied models.** Densities are tied over all states of the allophones of the same phone [3]. For each phone, a different set of densities is created, and this set is used for all context dependent variants.
- **Untying based on similarity between states.** In [4], automatic clustering of states determines which states have to share their densities. The proposed algorithm progressively unties the densities, enlarging the number of density sets and decreasing the set size. This method can be combined with phonetically tied modelling (the previous strategy).
- **Tying based on similarity between densities.** In this method, the models are progressively tied using density merging algorithms on the total set of densities in all states, as proposed in [5] and [6]. By merging densities from different states, the densities are automatically tied. Again this method can be combined with phonetically tied modelling.
- **Untying based on mixture weights.** In [7, 8, 9], the size of the mixture of densities for a state is reduced based on some selection criterion for the densities in the state. This way, the models are untied automatically.

Throughout the design of tied density models, we follow the last tying strategy. This paper focuses on the different options in the selection of the densities for a state. In section 2, the state modelling with SC-HMMs is briefly reviewed together with the methods we use to speed up the evaluation of the states. The selection criteria to reduce the mixture size in a state are described in section 3. In section 4, the experimental setup is described for the experiments on the Wall Street Journal (WSJ) recognition task.

* At Lernout & Hauspie Speech Products, Belgium.

Section 5 compares the different criteria and section 6 gives reference results on the task.

2. SC-HMM STATE MODELLING

Semi-Continuous HMM systems use a mixture of - generally gaussian - *pdfs* to model a state. The observation probability of state s for frame $\bar{X}$ in a *full* SC-HMM - in which a state is a mixture of all gaussians - is given by:

$$\mathcal{F}_s(\bar{X}) = \sum_{i=1}^{N} \lambda_{s_i} \times \mathcal{N}_i(\bar{X}) \qquad (1)$$

with N the size of the gaussian set, λ_{s_i} the weight for gaussian i in state s and $\mathcal{N}_i(\bar{X})$ the probability of gaussian i.

We use two methods to speed up these calculations:

- **The construction of *reduced* SC-HMMs.** Here for each state s the M_s most important gaussians are selected. The *pdf* of a state thus is a mixture of only M_s gaussians. As the total number of gaussians N has to be large for accurate acoustic modelling, M_s can be substantially smaller than N, so observation probability calculations are far more efficient for the reduced SC-HMMs than for the original full SC-HMMs. In the next section 3, different criteria are proposed for the selection of gaussians for each state.
- **The FRoG (Fast Removal of Gaussians) system.** The evaluation of the total set of gaussians is also very time consuming. Therefore we implemented the FRoG (Fast Removal of Gaussians) system. It decides in a very fast way which gaussians are expected to have a low probability for the current frame avoiding thus their exact evaluation. For a gaussian set of size 10000, the number of fully evaluated gaussians can be reduced to 500 (5%) without degradation in recognition performance. As FRoG is a scalar method, the overhead for the system is small, it is comparable to the cost of evaluating 2% of the gaussians (for $N = 10000$). For more details on the FRoG system, the reader is referred to [8, 9].

3. REDUCING THE SC-HMM

Reduction of the number of gaussians used to model a state will make the observation probability calculations more efficient. Different selection criteria are possible to decide which gaussians to use in each state. We investigated the following criteria:

- *fixed set size*: a fixed number of gaussians per state is selected, only the gaussians with the highest weights are retained. We used this method in [8, 9].
- *flooring on the weights*: an absolute flooring value for the weights is used, gaussians with a smaller weight are omitted for that state. This method was proposed in [7].
- *fixed probability percentage*: the gaussians with the highest weights are selected up to the point where the sum of these weights reaches a predetermined percentage. We obtained the results with context dependent SC-HMMs in [10, 9] with this method.
- *flooring on the occupancy*: the occupancy of a state is defined as the number of observations assigned to that state. The occupancy for a gaussian in a state then is the number of observations with which the weight for that gaussian is estimated (this is the weight for the gaussian multiplied by the state occupancy). Gaussians with an occupancy in a state smaller than a predefined value are removed for that state.

Note that in reduced SC-HMMs, the number of states in which a certain gaussian is used, is not controlled. Therefore it is possible that by reducing an SC-HMM, some gaussians are not used in any state. In this (rare) case, these gaussians are removed from the set.

4. EXPERIMENTAL SETUP

We evaluated the different selection criteria for reduction of the number of gaussians in a state on the speaker independent Wall Street Journal (WSJ) recognition task.

Standard bigram and trigram language modelling provided by Lincoln Laboratory for the 5k word closed vocabulary and the 20k word open vocabulary is used. The results - word error rate (WER) - are given on the November 92 evaluation test sets with non verbalised punctuation. They contain 330 sentences for the 5k word task, and 333 sentences for the 20k word task.

The signal processing gives mean normalised Mel scale cepstrum (12 parameters) and log energy, all of them with first and second time derivative. This results in 39 parameters in total.

Our acoustic modelling is gender independent and based on a phone set with 45 phones, without specific function word modelling. In the experiments below, no cross-word phonetic rules are used to adapt phonetic descriptions depending on the neighbouring words.

A time-synchronous beam search algorithm is used. In the experiments which compare the different selection criteria (section 5), the number of hypotheses in the beam is rather small. This means that the result can be improved when the thresholds in the beam controller are chosen conservatively as to avoid search errors. Results with a large search beam are reported in section 6 as reference for our acoustic modelling with SC-HMMs.

5. COMPARING CRITERIA

5.1. Design of the SC-HMMs

For the experiments comparing the selection criteria, the acoustic modelling is based on the standard SI-84 (WSJ0) training set. This training set consists of data from 84 different speakers, 7240 sentences in total.

First context independent acoustic models are created. The reduced SC-HMM is designed using the methods described in [8, 9]. Each state is modelled as a mixture of 256 gaussians out of a total set of 10296 gaussians. Up to this point, only the *fixed set size* method is used to select gaussians in a state.

Then cross-word context dependent models are designed using phonetic decision trees (and our node splitting criterion described in [10, 9]). In total 5139 tied states are estimated for 19677 different models. Each tied state is modelled as a mixture of 256 gaussians. These context dependent models are further trained with three Viterbi training steps, without any reduction in the number of gaussians per state.

5.2. Comparing Selection Criteria

In table 1, the different selection criteria are compared on the November 92 test set.

fixed set size				
reduced to	256.0	128.0	90.0	
5k words	7.29%	8.71%	10.93%	
20k words	14.51%	15.40%	17.24%	
flooring on the weights				
reduced to	92.9	81.2	67.8	61.1
5k words	7.75%	7.94%	8.89%	10.05%
20k words	14.97%	14.90%	16.00%	17.10%
fixed probability percentage				
reduced to	94.3	79.7	70.3	60.4
5k words	7.47%	7.70%	8.16%	8.63%
20k words	14.55%	14.51%	14.80%	15.88%
flooring on the occupancy				
reduced to	89.6	77.9	70.3	58.2
5k words	7.17%	7.55%	7.88%	8.11%
20k words	14.30%	14.41%	14.50%	15.31%

Table 1: Results (WER) with different reduction methods

To obtain these results, the models with 256 gaussians per state (described above) are reduced in various degrees and with different selection criteria, and these models are tested without any further training. In the table, the average number of remaining gaussians per state is given, together with the result on the 5k word and the 20k word task, using the bigram language models.

The difference in performance between the four methods can be seen clearly from the table. The method with *fixed set size* behaves the worst, then comes reduction with *flooring on the weights*, reduction with *fixed probability percentage* is the second best and the best method is our new criterion, based on *flooring on the occupancy*.

On the 20k word task, the models with *flooring on the occupancy* are as good as the original models (with 256 gaussians per state) up to a reduction to 70 gaussians per state. But on the 5k word task, some detail seems to be lost.

5.3. Further Model Training

The performance of the reduced models can be improved by further training. We did one more Viterbi training step on the reduced models with on the average about 70 gaussians per state. The results are shown in table 2. The worse the model, the more gain there is due to further training, but the results with *flooring on the occupancy* are still the best.

flooring on the weights, reduced to 67.8				
	no training		1 training step	
	5k words	20k words	5k words	20k words
	8.89%	16.00%	8.26%	15.01%
fixed probability percentage, reduced to 70.3				
	no training		1 training step	
	5k words	20k words	5k words	20k words
	8.16%	14.80%	7.83%	14.89%
flooring on the occupancy, reduced to 70.3				
	no training		1 training step	
	5k words	20k words	5k words	20k words
	7.88%	14.50%	7.77%	14.62%

Table 2: Results (WER) with further model training

6. REFERENCE RESULTS

For the reference experiments, the acoustic modelling is based on the (larger) SI-284 (WSJ1) training set, which consists of data from 284 different speakers, 37516 sentences in total.

In the decision tree based cross-word context dependent acoustic modelling, 7792 different tied states are estimated to construct 28438 models, using in total 20242 gaussians.

The number of gaussians per state is reduced with the *flooring on the occupancy* method to 142.0, 86.6 and 62.0. The results on the November 92 test set both for the bigram and the trigram language model are given in table 3, in this case a large search beam is used in order to avoid search errors. Reduction to 86.6 gives only a small deterioration of about 0.2% on average, the results of models with 62.0 gaussians per state are an additional 0.7% worse.

Next, some statistics about the experiment with 20k open vocabulary, bigram language model and the models with 86.6 gaussians on average per state are given.

nr. gauss.	5k closed		20k open	
per state	bigram	trigram	bigram	trigram
142.0	4.91%	3.29%	11.11%	9.18%
86.6	5.17%	3.21%	11.38%	9.66%
62.0	5.57%	3.57%	12.42%	10.53%

Table 3: Reference results (WER), November 92 test set

Using our FRoG system, on the average only 920.0 of the total set of 20242 gaussians have to be evaluated fully (4.54%). The average number of open hypotheses for the large search beam is 22106. Due to the beam search on the total of 7792 tied states on average 2807 states have to be evaluated. This number however depends on the beam size: for a small search beam (on average 8073 open hypotheses), only 1695 states are evaluated, resulting in a 0.4% drop in recognition performance.

On a 167 Mhz Sun Ultra-1, the evaluation of the gaussians takes about 150% of real time. For the small search beam, the calculation of the weighted products to evaluate the mixtures of gaussians that model the states costs about 250% of real time, the whole evaluation of the acoustic models thus being four times slower than real time.

Given these speed indications, and given that the error rate increases rather quickly by further reduction of the number of gaussians per state, it is clear that for a (considerably) faster acoustic model evaluation, not only the number of gaussians per state but also the number of gaussians and the number of states should be reduced.

One should also note however that in our current system, the search takes more time than the evaluation of the acoustic models, even for small search beams. Therefore reducing further the time spent on acoustic model evaluation may increase the total recognition time as less accurate models make the search more difficult.

7. CONCLUSIONS

Our current research in the field of acoustic modelling for continuous speech recognition focuses on the development of SC-HMMs for large vocabulary speaker independent systems. In this paper, we extended our previous work with SC-HMMs in order to make the evaluation of the observation probabilities for the states faster.

In particular, the number of gaussians used to model an SC-HMMs state is reduced. We compared four selection criteria to decide which gaussians will remain in the mixture for a state. Experiments on the WSJ recognition task show that our novel selection criterion, which selects gaussians for a state if their weight is based on sufficient acoustic data, outperforms the other criteria we investigated.

Further research can be conducted in order to solve the main problem in all four proposed criteria, namely that they always remove a mixture component because the weight for the component is small. Only the definition of *small* differs. In fact, a mixture component should be removed if the component is useless for that state, even if the weight is large. An (extreme) example: if a state is modelled with three gaussians, twice the same gaussian with a weight of 40% each, and one other gaussian with a weight of 20%, then all four proposed criteria will first remove the third gaussian. The loss in likelihood by dropping a mixture component in a state is a better criterion to decide if the component should be removed. But the algorithm based on this idea (or an approximation) is more complex than the four methods described above.

8. REFERENCES

1. J.R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter models for large vocabulary isolated speech recognition. In *Proc. of ICASSP*, volume I, pages 13–16, Glasgow, May 1989.
2. X.D. Huang and M.A. Jack. Semi-continuous hidden Markov models for speech signals. *Computer Speech and Language*, 3(3):239–252, July 1989.
3. C.H. Lee, L.R. Rabiner, R. Pieraccini, and J.G. Wilpon. Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4(2):127–166, April 1990.
4. V.V. Digalakis, P. Monaco, and H. Murveit. Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers. *IEEE Transactions on Speech and Audio Processing*, 4(4):281–289, July 1996.
5. C. Dugast, P. Beyerlein, and R. Haeb-Umbach. Application of clustering techniques to mixture density modelling for continuous-speech recognition. In *Proc. of ICASSP*, volume I, pages 524–527, Detroit, May 1995.
6. J. Simonin, S. Bodin, D. Jouvet, and K. Bartkova. Parameter tying for flexible speech recognition. In *Proc. of ICSLP*, volume II, pages 1089–1092, Philadelphia, October 1996.
7. D.B. Paul. The Lincoln tied-mixture HMM continuous speech recognizer. In *Proc. of ICASSP*, volume I, pages 329–332, Toronto, May 1991.
8. K. Demuynck, J. Duchateau, and D. Van Compernolle. Reduced semi-continuous models for large vocabulary continuous speech recognition in Dutch. In *Proc. of ICSLP*, volume IV, pages 2289–2292, Philadelphia, October 1996.
9. J. Duchateau, K. Demuynck, and D. Van Compernolle. Fast and accurate acoustic modelling with semi-continuous HMMs. *Speech Communication*, 24(1):5–17, July 1998.
10. J. Duchateau, K. Demuynck, and D. Van Compernolle. A novel node splitting criterion in decision tree construction for semi-continuous HMMs. In *Proc. of EUROSPEECH*, volume III, pages 1183–1186, Rodos, September 1997.