A NOVEL NODE SPLITTING CRITERION IN DECISION TREE CONSTRUCTION FOR SEMI-CONTINUOUS HMMS*

Jacques Duchateau, Kris Demuynck and Dirk Van Compernolle †

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

ABSTRACT

In [1], we described how to improve Semi-Continuous Density Hidden Markov Models (SC-HMMs) to be as fast as Continuous Density HMMs (CD-HMMs), whilst outperforming them on large vocabulary recognition tasks with context independent models.

In this paper, we extend our work with SC-HMMs to context dependent modelling. We propose a novel node splitting criterion in an approach with phonetic decision trees. It is based on a distance measure between mixture gaussian probability density functions (*pdfs*) as used in the final tied state SC-HMMs, this in contrast with other criteria which are based on simplified *pdfs* to manage the algorithm complexity.
Results on the ARPA Resource Management task show that the proposed criterion outperforms two of these criteria with simplified *pdfs*.

1. INTRODUCTION

Both in Continuous Density HMMs (CD-HMMs) and in Semi-Continuous Density HMMs (SC-HMMs), a state is modelled as a mixture of elementary *pdfs*, gaussians in our case. But there is an important difference. In a CD-HMM, for each state a specific (small) set of gaussians is modelled. In SC-HMMs, one large set of gaussian *pdfs* is shared for all states. The distinction between states is made with the weights for each gaussian in the mixture.

We opted for SC-HMMs because they offer some advantages over CD-HMMs:

- The set of gaussian *pdfs* in the SC-HMM case directly models the whole feature space, not the overlapping subspaces for the different HMM states. In this way reestimating essentially the same gaussian for different HMM states can be avoided.
- In an SC-HMM the tied gaussians are trained using data points from many states, only the mixture weights are estimated with data for the state itself. On the other hand for a CD-HMM state both mixture weights and (state specific) gaussians have to be estimated. Therefore far less data is needed to estimate an SC-HMM state. In other words, using the same amount of data one can model more states, or improve state modelling by increasing the number of mixture components.
- In an SC-HMM the number of gaussians and the number of states can be chosen independently. So it is easy to change the number of states in the SC-HMM case, for instance to transform context independent models into context dependent models.

Modelling with SC-HMMs however has an important drawback. The total number of gaussians needed for accurate modelling of the feature space is quite large. Therefore the calculation of the output probabilities of a full SC-HMM - for which the *pdf* of each state is a mixture of *all* gaussians - is prohibitively time consuming.

In our previous work, we described two methods to speed up drastically the output probability calculation of full SC-HMMs. They are briefly reviewed in section 2.

With this efficient Semi-Continuous state modelling, tied state context dependent acoustic models are constructed using phonetic decision trees. In section 3, the node splitting criterion used in the decision tree construction is discussed. The criteria proposed in literature are based on a simplified modelling for the nodes in the tree to manage the algorithm complexity. We adapted the criterion for SC-HMM modelling and propose a novel node splitting criterion that is consistent with the complex mixture gaussian *pdfs* used in the SC-HMM.

In section 4, results are shown on the ARPA Resource Management task. Both reference results for context independent and context dependent SC-HMMs and a comparison between three different splitting criteria for tree construction are given.

* This research was supported by IWT Research Contract 940044, entitled Nerex.
† Currently with Lernout & Hauspie Speech Products, Belgium.

2. STATE MODELLING IN AN SC-HMM

For SC-HMMs, the output probability of state s for frame $\overline{X}$ is given by

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

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

First *reduced* SC-HMMs are constructed to decrease the computational complexity of the output probability calculations. For each state only the M gaussians with the highest weights are selected. Using large gaussian sets (as needed for accurate modelling), M can be very small with respect to N: the other gaussians do not give an essential contribution to the mixture *pdf* for the state. A typical value for $N = 10000$ is $M = 100$.

The evaluation of the total set of gaussians is also very time consuming. Therefore we implemented the FRG (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 FRG is a scalar method, the overhead for the system is small, it is comparable to the cost of evaluating 2% of the gaussians (for $N = 10000$). Note that this system can be used for any set of gaussians, even in CD-HMM based recognisers.

For more details on the methods and algorithms, the reader is referred to [1]. Experiments described there also show for context independent models that our SC-HMMs are as fast as CD-HMMs, and that they outperform them for both isolated word and continuous speech recognition tasks.

3. DECISION TREES FOR MIXTURE DENSITIES

The use of phonetic decision trees [2] is a known solution for maintaining the balance between model complexity and available training data in large vocabulary cross-word context dependent modelling. HMMs with tied states are created by successively splitting acoustic data based on phonetically motivated questions. The main advantage over data-driven approaches is the ability to provide a mapping not only for the contexts that occur in the training set, but for unseen contexts too.

3.1. Proposed criteria

One of the research items in the construction of decision trees is the node splitting criterion that evaluates the effectiveness of the division defined by a question. Although recent systems in literature that use decision trees model a state with a mixture gaussian *pdf*, the proposed splitting criteria are all based on simplified *pdfs* because the algorithms for the mixture gaussian *pdfs* are prohibitively complex.

In [2] and [3], Poisson models are used, [4] and [5] base their criterion on a single gaussian *pdf*, [6] and [7] calculate the criterion using discrete models. In fact the last paper works with SC-HMMs, however since only the closest mixture component is taken into account for likelihood calculations, the criterion boils down to the one for discrete models.

The use of these simplified - thus poor - output probability models has an important drawback: the inability of the poor models to represent the complex *pdfs* in the nodes influences the score of the different questions. In other words, if the criterion was calculated with the true *pdfs* used in the models, different - more correct - decision trees would be found.

3.2. Criterion for mixture gaussian *pdfs*

The combination of mixture gaussian *pdfs* and maximum likelihood or entropy optimisation used in the criteria cited above results in computationally unmanageable algorithms. To be able to base our novel node splitting criterion on mixture gaussian *pdfs*, we optimise an other measure for the effectiveness of a division of the data by a question.

In our approach, a question for a node is selected if it minimises the overlap between the two child nodes, both modelled with a mixture gaussian *pdf*. We defined the overlap between two mixture gaussian *pdfs* as the average probability of a point from the first *pdf* evaluated by means of the second *pdf*.

So the overlap between mixture gaussian *pdfs* $\mathcal{F}_1$ and $\mathcal{F}_2$ is the M-dimensional (with M the dimension of the feature space) integral of the product of both *pdfs*. In formula, this gives:

$$\mathcal{O}_M(\mathcal{F}_1(\bar{X}), \mathcal{F}_2(\bar{X})) = \int_{\bar{X}} \mathcal{F}_1(\bar{X}) \mathcal{F}_2(\bar{X}) d\bar{X}$$

with $\bar{X}$ a vector in the M-dimensional feature space. Note that this is a symmetric measure.

Working it out for $\mathcal{F}_1$ and $\mathcal{F}_2$ mixtures of one large set of N gaussians $\mathcal{N}_i$ with weights λ_{1_i} and λ_{2_i} respectively, one finds

$$\mathcal{O}_M(\mathcal{F}_1, \mathcal{F}_2) = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{1_i} \lambda_{2_j} \mathcal{O}_M(\mathcal{N}_i, \mathcal{N}_j)$$

For gaussians with diagonal covariance, the axes are independent, and the overlap between two M-dimensional gaussians can be written as the product of M one-dimensional overlap measures:

$$\mathcal{O}_M(\mathcal{N}_i, \mathcal{N}_j) = \prod_{k=1}^{M} \mathcal{O}_1(\mathcal{N}_{i_k}, \mathcal{N}_{j_k})$$

with $\mathcal{N}_{i_k}$ the kth dimension of gaussian $\mathcal{N}_i$.

Doing basic integral calculations and with a renormalisation to have a unit overlap between a gaussian and itself (dividing by the square root of the product of the inner overlap of both gaussians), we get

$$\mathcal{O}_1(\mathcal{N}_{i_k}, \mathcal{N}_{j_k}) = F \times \exp\left(-\frac{1}{2}\frac{(\mu_{i_k} - \mu_{j_k})^2}{\sigma_{i_k}^2 + \sigma_{j_k}^2}\right)$$

with μ_{i_k} and σ_{i_k} the kth parameter of the mean and sigma of gaussian $\mathcal{N}_i$, and factor F given by

$$F = \frac{1}{\sqrt{(\sigma_{i_k}^2 + \sigma_{j_k}^2)/2\sigma_{i_k}\sigma_{j_k}}}$$

In practice, for SC-HMMs, the *pdfs* for the nodes on the different levels in the decision tree are all a mixture of the total set of (fixed) gaussians. So the weights for the mixture gaussian *pdf* that models the training data in a node can easily be calculated by combining the mixture weights of all contexts that contribute to that node. To do this, both mixture weights and state size (number of data points) have to be stored beforehand for the most specific contexts that were to be modelled.

In contrast with criteria based on maximum likelihood, the outlined criterion does not automatically prefer subtrees of (nearly) equal size. In fact the algorithm with the criterion as described above will split off each time a rather small part of the remaining data, this way creating unbalanced decision trees.
In principle, this is not really a drawback for context dependent acoustic modelling as the tree can be re-organised afterwards. We nevertheless decided to re-organise the trees automatically by dividing the overlap measure above by the square root of the product of both node sizes (number of data points). The better the subtrees are balanced, the larger this product will be.

4. EXPERIMENTAL RESULTS

We evaluated our context independent and context dependent SC-HMMs on the speaker independent 991-word ARPA Resource Management (RM) task.

- Standard SI-109 train set for acoustic modelling. This train set consists of data from 109 different speakers, 3990 sentences in total.
- Test set feb89-SI is used for system development, oct89-SI and feb91-SI for evaluation tests. All three test sets consist of 300 sentences from 10 speakers.
- Standard Word Pair and No Grammar for language modelling. The Word Pair grammar gives a branching factor of about 60 on the test sets.
- The reported results are obtained with the NIST scoring programmes, allowing homophone errors for the No Grammar, but not for the Word Pair. The word error rate (WER) is given (sum of substitutions, insertions and deletions).

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.

The gender independent acoustic modelling is based on a phoneme set with 46 phonemes, without specific function word modelling. In the experiments below, no inter-word phonological rules are used to adapt phonetic descriptions depending on the neighbouring words.

In each of the context dependent experiments, the models are derived from a single *global* decision tree for all acoustic data. So the phonetic questions in the nodes on the different levels of the decision tree can concern the phoneme identity, the state number in the phoneme, or the phoneme context.

A time-synchronous beam-search algorithm is used. As we want to evaluate acoustic models, the thresholds in the beam controller are chosen fairly conservative to avoid search errors.

4.1. Reference results

The optimal results obtained with our SC-HMMs are summarised in table 1.

Context independent	feb89-SI	oct89-SI	feb91-SI
No Grammar	24.3%	26.3%	22.9%
Word Pair	4.8%	5.8%	5.2%
Context dependent	**feb89-SI**	**oct89-SI**	**feb91-SI**
No Grammar	15.9%	17.5%	17.6%
Word Pair	2.9%	2.9%	2.7%

Table 1: Reference results (WER) with SC-HMM modelling

The context independent models consist of 139 states (46 3-state left to right models for the phonemes and 1 noise state). Per state 256 gaussians are selected out of a set of in total 10740 gaussians. Using the FRG system, on

the average over the frames of the development test set feb89-SI only 552.9 (5.1%) of the gaussians had to be calculated.

As for the (cross-word) context dependent models, in total 15502 context dependent units are created using 3454 different tied states with on the average 69 gaussians selected per state. The decision tree construction was based on our node splitting criterion as described in section 3.2. The characteristics of the gaussian set are about the same as for the context independent models (10698 gaussians in total, on the average 543.7 gaussians evaluated per frame).

4.2. Comparing node splitting criteria

In this section, three different node splitting criteria are compared. The first two are based on maximum likelihood optimisation and can be found in in [4] and [7] respectively. The third is our own criterion, it uses a distance measure between densities as described in section 3.2.

The difference between the three criteria we want to emphasise here, is the type of node (or state) modelling on which they are based. Both criteria from literature use simplified node *pdfs*, the first a *single gaussian density*, the second a '*discrete density*' (which corresponds to discrete models). Our criterion on the other hand is derived for the *mixture gaussian density* used in the final tied state HMM modelling.

In table 2, the results (WER) with all three criteria are given on the development test set feb89-SI and on both evaluation test sets oct89-SI and feb91-SI (Word Pair grammar). The names for the three criteria correspond to the type of node modelling for the criterion as explained above. All models use a total set of 10000 gaussians, of which by means of our FRG system on the average only about 5% is evaluated.

Criterion based on	feb89-SI	oct89-SI	feb91-SI
Single gaussian density	3.1%	3.5%	2.5%
Discrete density	3.4%	3.4%	2.5%
Mixture gaussian dens.	2.9%	2.9%	2.7%

Table 2: Comparison between node splitting criteria

On the average, our node splitting criterion performs better than both other criteria. When mixture *pdfs* are used to model the states, it seems to be worth to design a specific criterion for these mixture densities.

5. 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 to context dependent modelling. In an approach with phonetic decision trees, we adapted the node splitting criterion to the specific state modelling (with mixture gaussian *pdfs*) in SC-HMMs.

Experiments on the ARPA Resource Management task show that the recognition performance improves when using decision trees that are constructed with a node splitting criterion based on the complex *pdfs* rather than on simplified *pdfs* as proposed in literature before.

REFERENCES

1. 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.

2. L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Decision trees for phonological rules in continuous speech. In *Proc. of ICASSP*, pages 185–188, Toronto, May 1991.

3. R. Kuhn, A. Lazaridès, Y. Normandin, and J. Brousseau. Improved decision trees for phonetic modeling. In *Proc. of ICASSP*, volume I, pages 552–555, Detroit, May 1995.

4. J.J. Odell. *The Use of Context in Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, U.K., March 1995.

5. L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc. of ICASSP*, volume I, pages 533–536, Adelaide, April 1994.

6. M.-Y. Hwang, X. Huang, and F. Alleva. Predicting unseen triphones with senones. In *Proc. of ICASSP*, volume II, pages 311–314, Minneapolis, April 1993.

7. G. Boulianne and P. Kenny. Optimal tying of HMM mixture densities using decision trees. In *Proc. of ICSLP*, volume I, pages 350–353, Philadelphia, October 1996.