Fast and Accurate Acoustic Modelling
with Semi-Continuous HMMs

Jacques Duchateau, Kris Demuynck and Dirk Van Compernolle

November 7, 2001

Katholieke Universiteit Leuven
Department of Electrical Engineering,
Kardinaal Mercierlaan 94,
B-3001 Heverlee, Belgium

Address for correspondence
Jacques Duchateau
ESAT - PSI - Speech
Kardinaal Mercierlaan 94
B-3001 Heverlee
BELGIUM
phone: +32 16 321861
fax: +32 16 321986
E-mail: Jacques.Duchateau@esat.kuleuven.ac.be

* At Lernout & Hauspie Speech Products, Belgium.
Number of pages : 23
Number of tables : 3
Number of figures : 4

Keywords
Large vocabulary speech recognition, acoustic modelling, semi-continuous Hidden Markov Models, parameter tying, decision trees
Abstract

In this paper the design of accurate Semi-Continuous Density Hidden Markov Models (SC-HMMs) for acoustic modelling in large vocabulary continuous speech recognition is presented. Two methods are described to improve drastically the efficiency of the observation likelihood calculations for the SC-HMMs. First, reduced SC-HMMs are created, where each state does not share all the - gaussian - probability density functions (pdfs) but only those which are important for it. It is shown how the average number of gaussians per state can be reduced to 70 for a total set of 10000 gaussians. Second, a novel scalar selection algorithm is presented reducing to 5% the number of gaussians which have to be calculated on the total set of 10000, without any degradation in recognition performance. Furthermore, the concept of tied state context dependent modelling with phonetic decision trees is adapted to SC-HMMs. In fact a node splitting criterion appropriate for SC-HMMs is introduced: it is based on a distance measure between the mixtures of gaussian pdfs as involved in SC-HMM state modelling. This contrasts with other criteria from literature which are based on simplified pdfs to manage the algorithmic complexity. On the ARPA Resource Management task, a relative reduction in word error rate of 8% was achieved with the proposed criterion, comparing with two known criteria based on 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, for instance gaussians. However there is an important difference between CD-HMMs and SC-HMMs. In a CD-HMM, each state is modelled by means of a specific (small) set of gaussians while in SC-HMMs, one large set of gaussian pdfs is shared for all states which are distinguished only by the weights for the gaussians 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 parameter space, not the overlapping subspaces for the different HMM states. In this way re-estimating essentially the same gaussian for different HMM states can be avoided.

- In an SC-HMM the tied gaussians are trained using data frames from many states, only the mixture weights are estimated with data for the state itself. 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 either 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 and transform for instance context independent into context dependent models.

This paper presents the use of SC-HMMs in context dependent acoustic modelling for large vocabulary continuous speech recognition. In section 2 an overview of a prototypical training procedure for SC-HMMs is given as a guideline for the remainder of the paper. The two sections thereafter highlight some specific topics we investigated in SC-HMM modelling.

Section 3 describes how we improved the efficiency of SC-HMM observation likelihood calculations in two ways. First the number of gaussians used to model a state is strongly reduced, and second a system is implemented that decides in a very fast way which gaussians in the total set are expected to have a low likelihood for an input frame so that their exact evaluation can be avoided.

Section 4 focuses on phonetic decision tree based context dependent modelling. First it is shown how the node splitting criterion that is used to construct decision trees can be adapted to SC-HMMs. Further we discuss the relative importance of phonetic questions related to phone identity, state number and phonetic contexts during the construction of a single global decision tree for all acoustic data.

In section 5, results are given on the ARPA Resource Management task, comparing different splitting criteria for tree construction. Finally, in section 6 some conclusions are given.

2 Training procedure for SC-HMMs

This section gives an overview of a prototypical training procedure for SC-HMMs. The different training steps are enumerated, without going into every detail. The main purpose is to situate the topics in SC-HMM modelling which are described in detail in the next sections. For that reason these sections are referred to in this overview.

1. Training of context independent SC-HMM models

1.1. A phonetic alignment is obtained for the acoustic data (e.g. a Viterbi alignment based on older acoustic models)
1.2. A large set of gaussian densities (diagonal covariance) is created with which all acoustic data can be modelled. This can for example be done by merging small sets of gaussians for each of the phones, where a set of gaussians for a phone is obtained by some clustering procedure on the acoustic data for that phone (as indicated by the phonetic alignment). Whenever the likelihoods of the gaussians in this large set have to be calculated either during training or recognition, the FRoG system (see section 3.2) is used to speed up the calculations.

1.3. A full SC-HMM (in which each state is modelled as a mixture of all gaussians) is estimated based on the fixed alignment obtained in step 1.1.

1.4. A reduced SC-HMM (see section 3.1) is created, in which small weights are set to zero and hence each state is modelled with only a small subset of the full set of gaussians. Re-estimation of the parameters in the successive training steps is done either with the fixed alignment or with a new Viterbi alignment based on the current models (Viterbi training). Note that during training both the mixture weights and the gaussians (mean and covariance) are re-estimated.

2. Training of context dependent SC-HMM models

2.1. All (for decision tree construction) necessary context dependent models are estimated from the acoustic data as a weighted mixture of the set of gaussians, and the state occupancies (number of data frames aligned to a state) are stored as explained at the end of section 4.1.2. This is easily done with a Viterbi training where each context dependent model is initialised with the corresponding context independent model.

2.2. A (global) phonetic decision tree is constructed (see section 4).

2.3. Based on the decision tree, the context dependent models are defined and the tied states are estimated as combination of the contributing contexts (which are estimated themselves in step 2.1.).

2.4. The context dependent models are further trained with Viterbi training steps.

3 State Modelling in an SC-HMM

Semi-Continuous (or tied mixture) HMM systems ([Bellegarda et al., 1989], [Huang et al., 1989]) use a mixture of - generally gaussian - pdfs to model a state. The observation likelihood of state $s$ for frame $\hat{X}$ is given by:

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

(1)

with $N$ the size of the gaussian set, $\lambda_{s_i}$ the weight for gaussian $i$ in state $s$ and $\mathcal{N}_i(\hat{X})$ the likelihood 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 then 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 likelihood calculations are far more efficient for the reduced SC-HMMs than for the original full SC-HMMs.
• The FRoG (Fast Removal of Gaussians) system.
This novel algorithm decides in a very fast way which gaussians are expected to have a low likelihood for the current frame so that their exact evaluation can be avoided. Note that this system can be used for any set of gaussians, CD-HMM based recognisers included.

Both methods are described in detail in sections 3.1 and 3.2 respectively.

3.1 Reduced SC-HMMs
3.1.1 Density tying strategies
In literature, different methods are proposed to create HMMs with a density tying degree between the two extremes, this of CD-HMMs on the one hand and that of the full SC-HMMs on the other.
In phonetically tied models [Lee et al., 1990], the densities are tied over all states of the allophones of the same phone. This is a deterministic way of tying. In [Dialakis et al., 1996], untying of densities is based on similarity between states: automatic clustering of states determines which states have to share their densities. The proposed algorithm progressively unites the densities, enlarging the number of density sets and decreasing the set size. Density tying based on similarity between densities is proposed in [Dugast et al., 1995] and [Simonin et al., 1996]. As densities from different states are merged, they are automatically tied. Finally, density untying can be based on mixture weights. In [Paul, 1991] and [Demuynder et al., 1996], less important components in the mixture that models a state are removed based on the fact that they have a low weight. This method automatically unites the models.

3.1.2 Reducing the mixture size in a state
Throughout the design of reduced SC-HMMs, we follow the last density tying method, selecting the gaussians in a state based on the mixture weights. Different selection criteria are possible.
A first method uses an absolute flooring value for the weights: gaussians with a smaller weight are omitted for that state. A second criterion in based on a fixed set size: a fixed number of gaussians per state is selected, only the gaussians with the highest weights are retained. A third method uses a 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.
The first criterion is used in [Paul, 1991], here we only use the second and the third.

3.1.3 Training of reduced HMMs
To reduce a full SC-HMM a sequential procedure has been adopted. The reason is that the omission of a number of gaussians in a state implies that the remaining ones have to model more frames. So the weights associated to the remaining gaussians may change, and this may influence the result of the further selection. Therefore successive, smaller reduction steps are used to guarantee well-trained weights and consequently a good selection of gaussians for each state.
In practice a large reduction step is performed first (many weights in a full SC-HMM are not important). Subsequent reduction steps up to the point where the target reduction is reached are smaller. Note that in reduced SC-HMMs, the number of states in which a certain gaussian is used is not controlled. If after reduction, a gaussian is not used in any state, then this gaussian is removed from the set. But this case is rare, for instance during the whole design of the context dependent models used for the experiments in section 5, only 12 gaussians were lost on a total set of about 10000 gaussians.
3.1.4 Reduction rates and recognition results

In this paragraph, some indicative numbers are given for the reduction rates. For a small system with 2000 gaussians in total, the SC-HMM can be reduced to 64 gaussians per state without increase in error rate. For a system with 10000 gaussians in total - as used for the experiments in section 5 - selection with fixed set size can reduce the SC-HMM to 256 gaussians per state. Selection with fixed probability percentage brings the average number of gaussians per state further down to 70.

From these figures it is clear that the observation likelihood calculation for reduced SC-HMMs is far more efficient than the calculation for full SC-HMMs. Note that for context dependent SC-HMMs the reduction of the number of gaussians per state is inevitable: the calculation of large full SC-HMMs is prohibitively expensive given the large set of gaussians.

In [Demuynck et al., 1996] it is shown that the reduced SC-HMMs are as fast as CD-HMMs whilst outperforming them on large vocabulary recognition tasks when context independent models are used with the same total number of gaussians. The main reason is the better state modelling in reduced SC-HMMs due to the larger mixture size in the SC-HMM states.

3.2 The Evaluation of the Gaussians

The exact calculation of a large set of gaussian pdfs is very time consuming. In this section, a novel algorithm is described to speed up these calculations without affecting the performance of the recogniser.

3.2.1 Fast gaussian evaluation methods

In literature, different algorithms are described to speed up the calculations of the pdfs for a given frame. Some of them ([Beyerlein, 1994], [Fritsch et al., 1995]) approximate the calculation of a mixture of pdfs to the best pdf, and propose fast algorithms to find the nearest neighbour for a frame among the gaussians in the mixture.

Others look at the whole set of gaussians for all states together, and attempt to reduce computation time irrespective (at least in principle) of the specific states the gaussians are related to:

- **VQ method.** This method, proposed in [Bocchieri, 1993], is based on vector quantisation of the acoustic space. The VQ area to which an input frame belongs identifies a subset of neighbouring gaussians that should be calculated for that frame. The overhead of this system is the calculation of the VQ. More sophisticated variants based on this idea can be found e.g. in [Digalakis et al., 1996] and [Knill et al., 1996].

- **Tree method.** In [Watanabe et al., 1995], all gaussian pdfs in all models are structured in a tree with the original gaussians as leaves and with envelope gaussians for the children pdfs in each node. The envelope gaussians are used to decide if full evaluation of the subtree is necessary and to approximate likelihoods in not evaluated subtrees.

- **Scalar method.** In [Demuynck et al., 1996], we proposed a scalar method for selecting promising gaussians: each parameter in a given input frame determines independently a set of gaussian pdfs that should not be evaluated fully.

  The independency between the parameters is an important restriction on the shape of the regions in the parameter space in which a gaussian should - or should not - be evaluated. But on the other hand, the scalar approach results in a fast selection and very detailed regions. In our method each axis in the high dimensional parameter space can be divided in many parts,
whereas this is impossible in the above mentioned methods. Our scalar algorithm, which we called Fast Removal of Gaussians (FROG), is described in detail below.

3.2.2 The FROG algorithm

In our scalar approach, the axes in the parameter space are treated independently. The situation for one axis is depicted in figure 1. The axis is divided linearly in a large number of cells (only a small number is chosen in the figure for clarity). A gaussian is marked as improbable for a cell if its likelihood in the centre of the cell is less than a specified fraction of the average likelihood in that cell over the whole gaussian set. The same fraction - which corresponds to a threshold when log likelihoods are considered - is used for all axes. The number of gaussians that are removed is controlled by varying this threshold. Before recognition, for each cell and each axis the list of improbable gaussians is stored. Note that other criteria can be defined to determine the cells where a gaussian is improbable, for instance based on the variance of the gaussian. Here we only use the one explained above. Whenever the set of gaussians has to be evaluated for an input frame, the FROG system is first used to decide which gaussians have to be evaluated fully. In order to do this, all gaussians are first marked being probable for the current input frame. Then, axis by axis, the cell on the axis is determined to which the current frame belongs, and the gaussians in the list corresponding to that cell are marked improbable. After all axes have been checked, only those gaussians that are still marked probable, are evaluated. Thus a gaussian will only be evaluated if it is not marked as improbable for any of the axes.

3.2.3 Storing the information

A good representation of the lists of improbable gaussians should allow for both a fast access to the lists and for efficient memory use. The trivial solution to store all information needed during recognition - i.e. listing all improbable gaussians for each division and axis - requires too much memory. An efficient method is proposed instead whereby the regions are represented as the exclusive union of simple binary regions as the bottom part of figure 1 illustrates for one of the axes. So the gaussian in the figure will only be stored in the 4 shaded binary regions, not in all 11 cells where it is improbable. To each binary region a list is attached containing the gaussians that are improbable over the complete range of the binary region. During recognition, the list of improbable gaussians for the current frame will be found for one axis as the union of all lists for the binary regions to which the frame belongs (5 binary regions in the figure).

Using this efficient storing method, memory requirements are considerably reduced while maintaining easy access to the lists of improbable gaussians.

3.2.4 Combining the FROG system with reduced SC-HMMs

When the FROG algorithm is used in conjunction with reduced SC-HMMs it is possible that an undesirable situation occurs in which for some states, none of the gaussians is evaluated. Since in general it is not a good idea to remove a promising option from the search beam just because of a bad match in a single frame, a non-zero observation likelihood must be associated to these states. A fixed lower bound on the gaussian likelihoods or observation likelihoods is not an option, as we cannot anticipate the range of these likelihoods in advance. Therefore we add a relative lower bound to the gaussian likelihoods. This lower bound is estimated as a fraction of the average likelihood over
all gaussians for the current frame (symbols as in formula 1):

$$\beta \times \sum_{i=1}^{N} \mathcal{N}_i(\bar{x})/N$$

(2)

assuming a zero likelihood for the non evaluated gaussians. The exact value of the fraction ($\beta$) is not critical and was made dependent on the (average) number of gaussians selected per state $M_s$ according to the following formula:

$$\beta = 10 \log_2(M_s)$$

(3)

3.2.5 Results

This section presents the choice for the parameters in the FRoG system as well as the obtained results given a 39-dimensional space.  

Memory requirements. It was found experimentally that 512 is a good value for the number of cells per axis. This large value illustrates that the possibility to define accurate regions per axis (the advantage of a scalar method) is useful indeed. Furthermore when 10000 gaussians were considered, the index of each one of them was stored on average 7.2 times per axis. In other words, the size of the information required by the FRoG system is somewhat less than 2 times the size of the set of gaussians.

Reduction ratio. The number of gaussians that have to be fully evaluated when the FRoG system is used depends on the size of the set of gaussians. Although for larger sets more gaussians have to be calculated, the ratio between the number of fully evaluated gaussians over their total number drops. More precisely, in the case of a small HMM system with only 2000 gaussians, only 300 gaussians have to be fully evaluated which represents a reduction by factor 7. For a gaussian set of size 5000, 400 gaussians have to be evaluated (factor 13). And for 10000 gaussians in total, no degradation in recognition performance was noticed when for each input frame, only 500 gaussians are evaluated. This corresponds to a reduction by factor 20.

Overhead of the selection algorithm. For a set of 10000 gaussians, the selection algorithm itself costs about one 50th of the time needed to evaluate all gaussians. Thus this overhead corresponds (in time) to the calculation of 200 gaussians per frame.

4 Decision Trees for SC-HMMs

The use of phonetic decision trees [Bahl et al., 1991] 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.

4.1 The Criterion in Decision Tree Construction

4.1.1 Overview

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 of gaussian pdfs, the proposed splitting criteria are
all based on simplified pdfs because the algorithms for the mixtures of gaussian pdfs are prohibitively complex.
[Bahl et al., 1991] and [Kuhn et al., 1995] use Poisson models, [Odell, 1995] and [Bahl et al., 1994] base their criterion on a single gaussian pdf, [Hwang et al., 1993] and [Boulianne et al., 1996] 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 simplified - thus poor - state modelling in the criterion 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.
In [Duchateau et al., 1997], we introduced a node splitting criterion based on a distance measure between the mixtures of gaussian pdfs that are used to model SC-HMM states. It is explained in detail below.

4.1.2 Criterion for mixtures of gaussian pdfs

The combination of mixtures of gaussian pdfs and maximum likelihood or entropy optimisation used in the criteria cited above results in computationally unmanageable algorithms. To enable the use of a node splitting criterion with mixtures of gaussian pdfs, an alternative measure regarding the effectiveness of a division of the data by a question is to be optimised.
Here a question for a node is selected if it minimises the overlap between the two child nodes, both modelled with a mixture of gaussian pdfs. The overlap between two mixtures of gaussian pdfs is defined as the average likelihood of a vector from the first pdf evaluated by means of the second pdf. So the overlap between two mixtures of gaussian pdfs \( F_1 \) and \( F_2 \) is the \( M \)-dimensional (with \( M \) the dimension of the parameter space) integral of the product of both pdfs as formulated in the following equation:

\[
O_M(F_1(\tilde{X}), F_2(\tilde{X})) = \int_{\tilde{X}} F_1(\tilde{X})F_2(\tilde{X})d\tilde{X}
\]  

(4)

where \( \tilde{X} \) is an \( M \)-dimensional vector. Note that this is a symmetric measure.
In the case of two distinct mixtures \( F_1 \) and \( F_2 \) of one large set of \( N \) gaussians \( \mathcal{N}_i \) with weights \( \lambda_{1_i} \) and \( \lambda_{2_i} \), respectively, the above equation may be expressed as:

\[
O_M(F_1, F_2) = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{1_i} \lambda_{2_j} O_M(\mathcal{N}_i, \mathcal{N}_j)
\]  

(5)

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:

\[
O_M(\mathcal{N}_i, \mathcal{N}_j) = \prod_{k=1}^{M} O_1(\mathcal{N}_{i_k}, \mathcal{N}_{j_k})
\]  

(6)

with \( \mathcal{N}_{i_k} \) the \( k^{th} \) dimension of gaussian \( \mathcal{N}_i \).
Using basic integral calculations and by means of a renormalisation to guarantee a unit overlap between a gaussian and itself the following is obtained:
\[ O_{1}(N_{i}, N_{j}) = \frac{1}{\sqrt{(\sigma_{i}^2 + \sigma_{j}^2)/2\sigma_{i} \sigma_{j}}} \exp \left( -\frac{1}{2} \frac{(\mu_{i} - \mu_{j})^2}{\sigma_{i}^2 + \sigma_{j}^2} \right) \]  

(7)

where \( \mu_{i} \) and \( \sigma_{i}^2 \) are the mean and variance for the \( k^{th} \) dimension of gaussian \( N_{i} \).

In practice, for SC-HMMs, the pdfs for the nodes on the different levels in the decision tree are all a mixture of the global set of gaussians. So only the mixture weights have to be calculated, which can be done easily by combining the mixture weights of all contexts that contribute to that node. To do this, both mixture weights and state occupancy (number of data frames aligned to the state) have to be stored beforehand for the most specific contexts that were to be modelled.

Note that, by using our node splitting criterion, one approximation with respect to SC-HMM modelling and training is exchanged for an other approximation. On the one hand the pdfs involved in our criterion are the ones applied in SC-HMM modelling, but on the other hand the idea of optimising the maximum likelihood - as used during the SC-HMM training - is lost. A property of maximum likelihood optimisation in a node splitting criterion is that automatically subtrees of (nearly) equal size will be generated: a smaller improvement in likelihood for a lot of data frames will be preferred over a large improvement for few data frames. In our criterion as outlined above, there is no reason to prefer subtrees of equal size, and therefore unbalanced trees are created. In order to solve this problem (and to move somewhat closer to the behaviour with maximum likelihood optimisation), an automatic tree adjustment was adopted by dividing our above measure by the square root of the product of both node occupancies. The better the subtrees are balanced, the larger this product is, so more balanced trees will be preferred.

4.2 The Global Decision Tree

Using a decision tree, a set of states - corresponding to a part of the acoustic data - can be tied together based on their acoustic similarity. One has different choices as for the size of this set of states, and thus for the number of decision trees that cover all acoustic data.

Often, one tree is constructed for each context independent phone state. This way the states with the same state number (for instance the first states) of any allophone of a phone can be shared. One can also decide to allow state sharing between all states of the allophones of a phone, or between the states with the same state number in allophones of different phones.

Or one can even construct one single, global decision tree for all acoustic data. In this case, the phonetic questions in the nodes on the different levels of the tree can concern the phone identity, the state number in the phone, or the phone context.

In this section, we show that states with the same state number belonging to different phones are more similar than the different states of the allophones of the same phone. And therefore it is more important in decision tree construction to be able to share states with the same state number from different phones than states with different state number but from the same phone. First two facts substantiating the above and taken from the context independent modelling case are presented. Thereafter a global decision tree generated by means of the new splitting criterion is considered and this leads us to the same conclusion.

4.2.1 Arguments from context independent modelling

A first argument in favour of the above statement is illustrated in figure 2 which shows in two dimensions the mean of the states (modelled as mixture of gaussians) of context independent models for
the plosives (\(d\#1\) depicts the first state of \(d\), in a 3-state model). As example, the second cepstral parameter and its first time derivative are given. It is clear that grouping by state number gives a better clustering than grouping by phone. Note that this is not only true for the derivative parameter, where a difference between the states of the same phone can be expected, but it happens with the cepstral parameter too.

This effect can be even more obvious for the vowels, certainly when for some vowels both a stressed and an unstressed version are included in the vowel set. But this of course strongly depends on the chosen phone set.

The second justification is derived by considering the amount of sharing of gaussians in the reduced SC-HMM models for the (context independent) phone states. Each of the states is modelled with a different subset of the total set of gaussians (see section 3.1).

Hence a simple measure for the resemblance between two states in reduced SC-HMM modelling is the number of shared gaussians. For each state, all other states were sorted according to this measure with the state that shares most gaussians at the top. For 85% of the states, at least one state with the same state number in an other phone was found closer to the top than the other two states of the same phone. The rank of the first state with the same state number but from a different phone is on the average 1.3, whereas the rank of the nearest of the two other states of the same phone is on the average 4.6.

The above again indicates that different states of the same phone are further apart than states from different phones with the same state number.

### 4.2.2 Arguments from a global decision tree

In a global decision tree, the phonetic questions at the nodes on the different levels of the decision tree may concern the phone identity, the state number in the phone, or the phone context. Therefore conclusions about the relative importance of these types of questions can be derived by simply considering the global tree: the further from the root of the tree a question is found, the more similarity there is between the contributing states.

The tree used as example here is constructed based on the new node splitting criterion as described in section 4.1.2. However the same conclusions can be derived when alternative splitting criteria are used as discussed in section 5.3. The phonetic questions which are employed here can be found in [Odell, 1995].

In figure 3, a part of the tree is shown. In the top levels of the tree (near the root), phonetic questions are asked about the phone identity up to the point where the remaining phones are the plosives (/pbdkg/) and the labiodental and dental fricatives (/fTD/). From the same figure it can be seen that although the phone identity is not known yet, a question regarding its state number has been asked. This indicates that for these 10 phones, it is more relevant to cluster states from the different phones but with the same state number than to cluster states from the same phone but with different state number.

The fifth node at the lowest level in figure 3 is further expanded in figure 4. Here again a question about the state number is asked while four phones are still remaining (fricatives /fTD/). Moreover we can see at bottom right that a question about the (left) phone context is more important than the final question about the phone identity (a question that disambiguates between the first state of /f/ and /T/).

Note that the nodes at the lowest level in figure 4 are not the leaves of the tree. In the complete decision tree, more questions (mainly) about the phone contexts need to be addressed before the final leaves are reached.
5 Experimental Results

We evaluated our context independent and context dependent SC-HMMs on the speaker independent 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 calculates mean normalised Mel scaled 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 phone set with 46 phones, without specific function word modelling. In the experiments below, no cross-word phonological rules are used to adapt phonetic descriptions depending on the neighbouring words, though the acoustic modelling relies on cross-word context dependent phone models.

The SC-HMMs are modelled with a total set of 10000 gaussians. The FROG system is always used and on the average only 5% of the gaussians need to be calculated.

For all context dependent experiments, a single global decision tree for all acoustic data is constructed. The construction is based on our node splitting criterion for mixtures of densities (except in the experiments where different node splitting criteria are compared). The context length in the trees is limited to one (immediate left and right context).

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

5.1 Reference Results

The best results obtained with our SC-HMMs are summarised in table 1. The context independent models consist of 139 states (46 3-state left to right models for the phones and 1 noise state). A fixed number of 256 gaussians per state have been selected. The total set consists of 10740 gaussians. Thanks to the FROG system, only 552.9 (5.1%) of the gaussians had to be calculated on the average over the frames of the development test set feb89-SI.

As for the context dependent modelling, in total 15502 cross-word context dependent models are created with 3454 different tied states. On the average, 69.2 gaussians are selected per state with the fixed probability percentage method (see section 3.1.2). The characteristics of the gaussian set are similar to these of the context independent models (10698 gaussians in total, on the average 543.7 gaussians evaluated per frame).
5.2 Effect of the Decision Tree Size

The stop criterion which has been used to limit the decision tree is quite simple at the current stage. A threshold on the number of frames aligned to each of the nodes is defined. Any question for node splitting that results in a subnode occupancy (number of data frames aligned to the node) below a predefined threshold can not be selected. If no question can be selected, a leaf in the tree is reached. These leaves correspond to the tied states.

In table 2, the effect of the minimal occupancy of a leaf on the recognition performance is given. For different minimal occupancies, the number of distinct context dependent models, the number of tied states and the recognition result on feb89-SI with Word Pair grammar are presented. The tied states are modelled with a mixture of a fixed number (256) of gaussians.

It seems from this and other experiments which have been conducted, that the optimal value for the minimum number of frames aligned to a leaf is about 200. With these 200 data frames, the 256 weights in the mixture for the state have to be estimated (but no state specific gaussians).

5.3 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 [Odell, 1995] and [Boulianne et al., 1996] respectively. The third is our own criterion as described in section 4.1.2.

The difference between the three criteria we want to emphasise here, is the type of node 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 of gaussian densities used in the final tied state HMM modelling.

In table 3, 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 their type of node modelling. All three models are of comparable size, the models for our criterion are the same as those used for the reference results given above.

On the average over the three test sets, our node splitting criterion results in a 7.4% relative drop in WER compared with the criterion based on a single gaussian density, and a 9.7% relative drop in WER compared with the criterion based on a `discrete density`. When mixtures of pdfs are used to model the states, the design of an appropriate criterion seems to be necessary in order to get optimal results.

6 Conclusions

In this paper, we showed that Semi-Continuous HMMs can be used for accurate acoustic modelling in large vocabulary speaker independent speech recognition.

First, the SC-HMM state modelling is described. The aim is to speed up the observation likelihood calculations. In reduced SC-HMMs, a state is modelled with a mixture of only a subset of all gaussians instead of all of them. Our FRoG system on the other hand is a fast, scalar method to determine the most promising gaussians for the input data frame. With this method, only 5% of the gaussians have to be evaluated on a total set of 10000.

Secondly, these reduced SC-HMM states are used in tied state context dependent modelling based on phonetic decision trees. We proposed a novel node splitting criterion for the decision tree construction that is consistent with the complex mixtures of gaussian pdfs used in the SC-HMMs. Recognition results show that this criterion outperforms other criteria that refer to more simple pdfs.
Acknowledgement

This research was supported by IWT Research Contract 940044, entitled *Continue spraakherkenning van grote vocabularia in het Nederlands*, from the Flemish government.
References


List of Figures

1. A one-dimensional solution to the problem of detecting improbable gaussians . . . . 19
2. Parameter mean for the states in the plosives . . . . . . . . . . . . . . . . . . . . . . . . . 20
3. Global decision tree - part about phones /pbdkgfvtD/ . . . . . . . . . . . . . . . . . 21
4. Global decision tree - part about states 1 and 2 of phones /vfTD/ . . . . . . . . 22

List of Tables

1. Reference results (WER) with SC-HMM modelling . . . . . . . . . . . . . . . . . . . . . . . 23
2. Effect of the leaf occupancy threshold . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
3. Comparison between node splitting criteria (WER given) . . . . . . . . . . . . . . . . . 23
Figure 1: A one-dimensional solution to the problem of detecting improbable gaussians
Figure 2: Parameter mean for the states in the plosives
Figure 3: Global decision tree - part about phones /pbtdkgfTD/
Figure 4: Global decision tree - part about states 1 and 2 of phones /v_TD/
Table 1: Reference results (WER) with SC-HMM modelling

<table>
<thead>
<tr>
<th></th>
<th>feb89-SI</th>
<th>oct89-SI</th>
<th>feb91-SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>independent</td>
<td>24.3%</td>
<td>26.3%</td>
<td>22.9%</td>
</tr>
<tr>
<td>Word Pair</td>
<td>4.8%</td>
<td>8.8%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dependent</td>
<td>15.9%</td>
<td>17.5%</td>
<td>17.6%</td>
</tr>
<tr>
<td>Word Pair</td>
<td>2.9%</td>
<td>2.9%</td>
<td>2.7%</td>
</tr>
</tbody>
</table>

Table 2: Effect of the leaf occupancy threshold

<table>
<thead>
<tr>
<th>Threshold</th>
<th># models</th>
<th># states</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>3779</td>
<td>913</td>
<td>3.4%</td>
</tr>
<tr>
<td>500</td>
<td>7165</td>
<td>1680</td>
<td>3.3%</td>
</tr>
<tr>
<td>200</td>
<td>15502</td>
<td>3454</td>
<td>3.0%</td>
</tr>
<tr>
<td>100</td>
<td>22017</td>
<td>5668</td>
<td>3.1%</td>
</tr>
</tbody>
</table>

Table 3: Comparison between node splitting criteria (WER given)

<table>
<thead>
<tr>
<th>Criterion based on</th>
<th>feb89-SI</th>
<th>oct89-SI</th>
<th>feb91-SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single gaussian density</td>
<td>3.1%</td>
<td>3.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Discrete density</td>
<td>3.4%</td>
<td>3.4%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Mixture of gaussian densities</td>
<td>2.9%</td>
<td>2.9%</td>
<td>2.7%</td>
</tr>
</tbody>
</table>