Computer Speech and Language (1995) **9,** 107–121

Search and learning strategies for improving hidden Markov models

Renato De Mori, Michael Galler and Fabio Brugnara

McGill University, School of Computer Science, 3480 University Street, Montreal, Quebec H3A 2A7, Canada

Abstract

A speaker-independent automatic speech recognition system is developed using hidden Markov models (HMMs). Simulated annealing and randomized search are used to optimize discrete features of the system, including topologies, parameter ties, context clusters, and the sizes of mixture densities. Domain knowledge is used to initialize and to constrain the search, which optimizes recognition performance while reducing the number of model parameters. System performance results for new types of discrete and continuous HMMs measured on the TIMIT corpus are reported. The small set of context-independent phoneme HMMs produced is competitive with much larger systems of context-dependent models.

1. Introduction

Hidden Markov models (HMMs) are popular acoustic models used for automatic speech recognition (ASR). An HMM is a probabilistic finite state automaton which can describe a stochastic source with a good compromise between simplicity and generality (Huang, Ariki & Jack, 1990). The list of arc-connecting states, and the mapping which assigns to each of them a probability density are referred to as the model "topology". While rigorous mathematical methods have been developed for estimation of acoustic parameters, the choices made for the topology, for tying distributions among different arcs or states, and for the phonetic or phonemic events to be represented by an HMM, have been considered a design art.

Another important aspect to consider concerns training strategies, and in particular the possibility of performing successive refinements and simplification steps on topologies and distributions by initializing parameter estimation in a given step using the parameters estimated in a previous step.

Previous studies have provided experimental evidence that having different HMMs for the allophones of a given phoneme substantially improves the performance of ASR systems (Lee *et al.*, 1990; Kimball, Ostendorf & BechWati, 1992). It is also well-known that training a large number of allophone models requires a very large training corpus containing many samples of each allophone model to be trained.

Various training methods have been proposed to reduce the imprecision of parameter estimation if a limited amount of data is available for certain phonemes. One is to

0885–2308/95/020107 + 15 $08.00/0

train models at different levels of detail—isolated phonemes, and phonemes in the left and/or right context—and to interpolate the statistical parameters of these models in order to obtain the parameters of allophone HMMs (Chow *et al.*, 1986). Another method consists of clustering contexts, and performing a sort of generalization using classification and decision trees (Hon, 1992), such that enough training samples are available for each cluster, and clustering respects some mathematically-defined criteria about the "impurity" introduced when merging allophones.

In addition to interpolation and clustering, various types of tying the statistical distributions associated to different elements of HMMs have been studied (Young, 1992).

This paper proposes a methodology for a mathematically sound solution of the problems whose solutions until now have usually been considered an "art". The basic concept is that choosing a set of allophones, HMM topologies, and tying of distributions is a search problem. There are various methods and measures of success in directing a search toward some goal.

In Section 2, simulated annealing is proposed as a search method for the above-mentioned problems, guided by the recognition performance on a subset of the experimental corpus disjoint from the test set. In general, search complexity is prohibitive, so suitable heuristics have to be used in order to constrain it. Section 3 describes how search is used to derive topologies and distribution ties. Also, heuristics based on speech knowledge are proposed to constrain the choice of phonemic and phonetic contexts which characterize a cluster.

We report experimental results for phoneme recognition on the TIMIT corpus using the allophone models obtained with the above-described search procedure. The recognition language model for our experiments is a loop of allophone models similar to those described in the literature (Lee & Hon, 1989). The experiments show small improvements obtained with topology optimization, and substantial improvements with allophone models.

It is well-known that HMM parameter estimation depends on the initial values assumed for the statistical parameters. In principle it should be possible to redesign phoneme models using the topologies and the results of allophone-model training as starting conditions. Section 4 describes how allophone models corresponding to the same phoneme are merged into a single phoneme model and shows that performance of the new phoneme models is close to that of the allophone models, suggesting that the method proposed in this paper allows one to build a small and effective set of phoneme models with a moderately large number of distributions.

Further improvements are obtained by conceiving simple HMMs, one for each phoneme, with mixtures having a large number of Gaussian distributions only at the beginning and at the end of the model. This choice is supported by the conjecture that co-articulation effects produce large parameter variability at the boundaries between a phoneme and its neighbours. The initial values of the parameters of the Gaussian distributions are the ones of the trained distributions in corresponding initial and final arcs of the allophone models. Also, improvements can be obtained by eliminating similar Gaussian distributions and retraining the models with a reduced number of mixtures.

Section 4 also shows how performances can be improved by introducing simple acoustic parameters describing time and broad-band features not well characterized by mel-scaled cepstral coefficients and their derivatives.

2. Search strategies

2.1. Simulated annealing

In contrast with hill-climbing or gradient-descent optimization methods, randomized search does not set out to thoroughly explore the vicinity of a local extremum of the cost function, but employs instead a random solution generator to visit points all over the solution surface. One advantage of this approach is that it is easily applied to almost any optimization problem, continuous or discrete, regardless of non-linearities or discontinuities. A randomly generated set of solutions S_i is unlikely to contain the global or even a local extremum to a non-trivial cost function $f(S)$. However, there is a higher probability of yielding a solution S^* such that

$$|f(S^*)-f(S_{optimal})|<m, \tag{1}$$

where m is some acceptable margin for error in the solution.

The goal here is to find improved values for certain discrete parameters of an HMM speech recognition system. These parameters include the *topology* of the unit models. The cost measure to be optimized is a measure of the recognition performance of the system. As an example, consider the problem space represented by an HMM topology restricted to topologies with no more than seven states and seven output distributions. The state–transition matrix has 49 entries, each of which can contain nine values. This means there are $5{\cdot}7\times10^{46}$ distinct solutions. Clearly, exhaustive search is infeasible.

Although the computational cost of generating and testing new solutions prohibits more than a cursory search of the problem space, a pure Monte Carlo search can nevertheless have practical utility. Unless the initial solution is a good local optimum, any series of randomly generated candidates which are perturbations of the initial solution is likely to contain some candidates which improve the performance measure. Care must be taken with this approach, however, since the goal is to derive a speech model which generalizes well to new data.

A more directed kind of search is feasible, in which a method exists to escape from local minima. For the last 10 years researchers have applied the technique of *simulated annealing* (Kirkpatrick, Gelatt & Vecchi, 1983) to many areas where gradient and other hill-climbing methods were unavailable or inadequate. Simulating annealing is a method of randomized optimization which acts like a hybrid between Monte Carlo search and hill-climbing, beginning like the former and settling into the latter near the end of the search.

In *Boltzmann annealing*, the solutions to a given cost function are assumed to be distributed with the Boltzmann probability factor $P(S_i)=\exp(-E(S_i)/kT)$ where $E(S_i)$ is the cost-function value for solution S_i, k is Boltzmann's constant and T is a system parameter called *temperature*. A new solution is generated randomly, and accepted if the cost improves (i.e. $\Delta E<0$). Otherwise, the new solution is accepted according to the probability ratio

$$\begin{aligned}\frac{P(S_{i+1})}{P(S_i)}&=\exp(-(E(S_{i+1})-E(S_i))/kT)\\&=\exp(-\Delta E/kT).\end{aligned} \tag{2}$$

This conditional acceptance of non-improvements allows the search procedure to escape local minima. As T is reduced, the probabilities given by the Boltzmann distribution vanish for all but the lowest-cost solutions. It can be shown that, given the above assumptions, if the temperture is lowered in stages and enough solutions are sampled at each temperature, Boltzmann annealing will converge to a globally optimal solution. One "cooling" schedule that guarantees optimality is the logarithmic schedule

$$T_n = T_0 \frac{\ln n_0}{\ln n}, \tag{3}$$

where n is the temperature iteration and $\{T_0, n_0\}$ are some reasonable starting values. In practice faster schedules are used which, while sacrificing the theoretic property of convergence, tend to provide useful optimization results. The latter approach is called *simulated quenching*.

In this paper simulated annealing is used to optimize the structure of HMMs for English phonemes. Because of the prohibitive size of the search space, a variety of non-optimal schedules, including linear ones, were employed. During the annealing, new solutions were generated from old by randomly permutating some discrete value, such as the value of an entry in the matrix representing the HMM topology. The measured recognition accuracy of the HMMs provided the value for cost function E.

2.2. Knowledge-guided search

The solutions searched for are refinements to the structure and organization of HMMs for speech, things that are usually set by hand. The models arrived at are in turn fitted to training data in the form of acoustic speech samples. To ensure the things learned about HMMs are generalized improvements, the objective function that drives the annealing process must be evaluated accurately. This means retraining and re-evaluating the models on thousands of acoustic samples every time a new model structure is generated. The computational cost of this procedure effectively precludes exploring the search space very well. In topology experiments, for example, only a few hundred solutions were tested at each temperature. However, any measurable improvement on existing solutions is desirable.

Another interesting approach would be to let the amount of data used to train and evaluate the models serve as the system "temperature". This would allow more solutions to be tested during the early part of the annealing, with the higher error in the evaluation function serving as a probabilistic factor affecting the acceptance of new solutions. Then as the system "cooled", increased training would provide a more and more accurate measure of solutions, and direct the system more toward a better solution (in this case lower recognition error).

In any case, since the problem space is large and the ability to explore it limited, the best way to solve the problem is to use the methodology in tandem with knowledge of the problem domain. Search begins with knowledge-guided and empirically-proven solutions. Randomized search is used to develop new and better solutions. These are evaluated in the light of knowledge about the speech modelling problem, and adjustments are made. If necessary, the entire procedure can be repeated. In the following experiments, this approach of knowledge-guided random search has proved to be an effective compromise, and provided better models for speech.

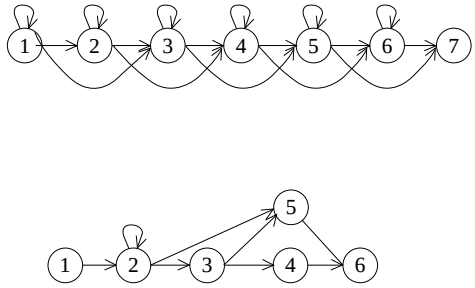

Figure 1. The initial (top) and optimized topology for model "b".

3. Applying search to HMM structure

3.1. Basic recognizer architecture

In each experiment, the original models are first evaluated by training them on a set of sentences, and testing their recognition performance on another set. The train/test suite is later used to measure the optimized speech system, to see if recognition performance improved. The sentences used to score the models during search belong to a third, separate test suite. All acoustic data were drawn from the TIMIT acoustic-phonetic speech corpus. Three thousand six hundred and seventy-nine sentences were used for training, and the core test of 192 sentences containing 7333 labelled phonemes was used for final evaluation.

Both discrete and continuous HMMs were trained and evaluated. For the continuous models, the HMM output distributions of 12 MEL-scaled cepstral coefficients, 12 difference cepstrum coefficients and energy and energy rate-of-change, were estimated from the training data using continuous-mixture densities composed of Gaussians. The covariance matrix is assumed to be diagonal. For the discrete models, three codebooks, one for each feature set, were introduced (256 entries were used for each of the first two sets and 64 for the energy features).

To speed up search, continuous models were abandoned in favour of discrete ones during optimization experiments. The HMMs were trained by Baum–Welch re-estimation. Recognition was performed using Viterbi maximum-likelihood decoding over a looped, phonemic finite state network.

3.2. Topologies and distribution ties

As mentioned in the introduction, the topology and tying of distributions in HMMs for speech are usually determined *a priori*, or based on statistical measures (Young, 1992). Some attempts are being made to evolve the topologies algorithmically (Casacuberta, Vidal, Mas & Rulot, 1990; Jouvet, Mauuary & Monne, 1991; Takami & Sagayama, 1992). In the method described here, the topology and distribution assignments of an HMM are represented as an integer array, and simulated annealing is used to optimize this data structure. Each perturbed solution is applied to the basic English phoneme models which are trained on one target set of sentences from the TIMIT database, using discrete parameters. The models are then tested on a separate test suite, with the measured unit accuracy used as the objective cost function to be maximized.

After a set of topologies adapted to particular phoneme classes have been derived, they are evaluated on a third data set of 192 sentences. Fig. 1 shows one such topology.

TABLE I. Comparison of initial and optimized topologies

Continuous models	No. distribution	Plosive errors	Vowel errors	Phone error rate
Seven-state topology	1440	402	924	44·06
Mixed topologies	1362	385	890	43·62

A mix of the best performing topologies are selected (based on individual phoneme accuracy), and finally evaluated by training a model set with continuous-mixture densities.

The searches are initialized with a proven, left-to-right model, and limit the growth of the model to seven states. Using domain knowledge in this way, the search is constrained enough to achieve a reasonable solution in reasonable time. Since the different phoneme classes are acoustically dissimilar, it is conjectured that different topologies should be developed for specific broad-phoneme classes.

Search began with the initial topology represented by the top of Fig. 1. The optimization proceeded by repeatedly perturbing values in the transition matrix, and evaluating the new topology represented by these perturbed values. Since the values represented both transitions and the distributions tied to them, the method optimizes simultaneously over both topology structure *and* parameter-tying.

Forty-eight or 53 units are modelled, depending on the experiment suite. Following Lee and Hon (1989), the original unit models are mapped onto a simpler set of 39 recognition units for classification. All performance results in this paper are for these 39 units. Recognition performance is given by the *phone error rate* (PER):

$$\mathrm{PER} = 100 \times \left(\frac{\text{no. of insertions} + \text{no. of deletions} + \text{no. of substitutions}}{\text{no. of units}} \right).$$

Topology optimization was performance-based, using 512 sentences to train and 96 sentences to evaluate each perturbation of the topology set. The success or failure of the experiments was determined by testing models on another set of 96 sentences. After many stages of optimization driven by large-sample performance, a best set of class-specific topologies was derived. This set contained seven new topologies for the phonetic classes: silence, closure, fricative, nasal, liquid, vowel and plosive. The models were trained as many iterations as possible until performance failed to improve, and then tested on a new dataset. A unit-by-unit study was made of how the new topology performed relative to the seven-state topology, counting the number of errors E associated in each case with a given phoneme p. E has four components: insertions, deletions, the number of times a p is misrecognized as something else, and the number of times something else is misrecognized as a p. Using the measure E, a new model set was constructed, a mixture containing the better-performing topology for each phoneme. This mixture showed an improved performance with a reduced number of Gaussian distributions (Table I). The mixture showed a 2% improvement in the discrete case, and 1% in the continuous. The mixture was even better when considered on a per-class

basis, significantly reducing the number of errors for vowels and hard-to-distinguish plosives.

3.3. Allophone context clusters

The same performance-driven optimization approach was applied to the problem of efficiently clustering contexts for allophone models. Since the problem space is large and expensive to search, acoustic-phonetic reasoning was used to determine a sensible initial grouping of contexts, and then simulated annealing perturbed these clusters so as to improve the recognition accuracy of models trained on these contexts. The output of the search procedure was also manually adjusted to conform to speech knowledge.

This approach is somewhat different from the tree-clustering algorithm of Bahl, deSouza, Gopalakrishnan, Nahamoo and Picheny (1991), or the state-splitting approach of Takami and Sagayama (1992). In this case, search is driven by performance and knowledge. In previous sections HMMs model the context-independent phonemes. Because of allophonic variability, better results are possible when phonemes are modelled in context (Schwartz *et al.*, 1985; Lee & Hon, 1989). However, there are 48 possible left contexts and 48 possible right contexts for each phoneme. If each context of each phone were modelled separately, it would be necessary to train $48^3 = 110\,592$ different models. In fact, the task would be impossible with existing speech corpora, since most of these contexts occur too rarely to be well represented in the training data (many in fact never occur).

The solution is to combine phonemic contexts into clusters which have similar contextual effects on the preceding or following phonemes. To model left-context units using any of 10 clustered contexts, the allophones may be represented with fewer than 500 models, a manageable number for which there are likely to be adequate training samples.

The problem is to choose the clusters appropriately. One approach would be to choose as fine-grained a context as can be well-trained from the available data. The more varied samples there are of a given speech unit, the more specialized models can be made for that unit. Kimball *et al.* (1992) suggest a distribution can be adequately estimated when the number of samples is about six times the dimension of the observation vector. If the vectors contain 12 cepstral coefficients (MEL or DME), at least 72 samples per model are needed (assuming at least one output distribution is not tied to multiple transitions). Depending on one's point of view, a rigid threshold like this may result in too many or too few models. (In practice, good results are achieved with some units trained on fewer samples.) In Lee, Giachin, Rabiner, Pieraccini and Rosenberg (1991) another data-driven approach is used to select clusters. Following a unit reduction rule based on the number of samples available for a unit in the training data, they build a set of models containing 47 context-independent phones, 134 diphones and 1101 triphones. While this approach guarantees the trainability of the units, it may produce a large number of models with similar distributions, in effect adding parameters without reducing system entropy.

In Lee *et al.* (1990) an initial set of context-dependent models is trained, and these models are merged into clusters, called generalized allophones, according to some algorithm. The first method, agglomerative clustering, is based on an entropy distance measure applied to the allophones. The second method is a heuristic decision tree, in which the root consists of the complete set of allophones corresponding to a particular

TABLE II. Initial right-context clusters for plosives

1.	ao aa ay aw ax ah
2.	ix
3.	ae
4.	ih
5.	uw uh
6.	er
7.	iy
8.	oy ow
9.	eh
10.	y
11.	ey

phoneme, and the leaves contain generalized allophones. At each node, the allophones are recursively divided into two subclusters based on the answer to a question provided by an expert linguist, a question designed to capture contextual effects. The recursion ends according to a metric of the distance between a parent cluster and its subclusters. In both above methods, the metric used to merge or split clusters is the "entropy increase" or information loss in the output distributions when two models are merged. Clustering is chosen so as to minimize entropy gain (Hon, 1992). A context-decision tree is also constructed in Bahl *et al.* (1991), but their subtrees are divided according to a Poisson-model likelihood of the possible splits measured against the training data. The generation of subtrees is statistically dependent on not just the adjacent (left or right) contexts, but on the several preceding and following phones as well. In De Mori, Laface and Piccolo (1976) two methods are suggested which proceed from the opposite direction: beginning with a generalized word model, they iteratively generate more HMMs to model that word, in which each HMM is re-estimated from different subsets of the training samples for that word. Although they do not address context, the same methods could in effect derive context clusters automatically, without appealing to acoustic-phonetic reasoning. However, the resultant context-sets would probably overlap. The state-splitting algorithm of Takami and Sagayama (1992) optimizes automatically along both topological and contextual axes, based on measure likelihood on the training data. A trivial Markov model is iteratively grown into a more complex model in which contexts are clustered and integrated. Jouvet *et al.* (1991) avoids the clustering problem, instead reducing the parameter space by integrating all left and right contexts into the topological structure of the allophone model. This can be seen equivalently as tying the distributions of the internal states of the allophone models.

In this paper, acoustic-phonetic reasoning is combined with the performance-driven randomized search described earlier in order to optimize the context clustering with respect to the available training data. In contrast with Takayama and Sagayama (1992), recognition accuracy rather than training-set likelihood serves as the objective function. We call this approach performance and knowledge-guided search.

Optimization was first applied to the problem of efficiently clustering right contexts for plosive allophones. Speech science suggests that the right-hand event, or succeeding phone, mostly affects plosive burst and transitions of relevant acoustic parameters. Intuition suggested an initial grouping of vowel right contexts as shown in Table II, in

TABLE III. Optimized right-context clusters

1.	ao aa ay aw ax ah
2.	ix ih iy y ey
3.	er ae
4.	uw uh oy ow
5.	eh
6.	f v
7.	s z
8.	l
9.	r
10.	w

TABLE IV. Initial left-context clusters for vowels

1.	p b f v m	13.	ax
2.	s z zh th dh jh t d n ch	14.	ix
3.	ng hh g k	15.	ih
4.	l	16.	ae
5.	w	17.	ah
6.	ao	18.	uh
7.	aa	19.	oy
8.	uw	20.	iy
9.	er	21.	ow
10.	ay	22.	eh
11.	ey	23.	sil epi bcl dcl gcl pcl tcl kcl
12.	aw		

which some phonemes starting with the same symbol were grouped together. Simulated annealing perturbed these clusters so as to improve the recognition accuracy of models trained in these contexts. The output of the search procedure showed a tendency toward grouping phonemes by place of articulation. Minor corrections to clustering were manually performed to finalize the trend, resulting in the grouping shown in Table III; some consonant clusters were also manually added.

Performance was used to drive the annealing. Since plosives were the model of interest, the number of errors on the plosive data alone served as the performance measure (i.e. how many plosive events were deleted or misrecognized). Almost all plosives in the training data are found in the contexts of Table III.

Optimization was next applied to the problem of efficiently clustering left contexts for vowels. The initial grouping (Table IV) puts all the unsonorant consonants having similar place of articulation in the same class. The clusters produced after search are presented in Table V. The number of contexts was reduced from 23 to 13.

TABLE V. Optimized left-context clusters

1.	p b f v m
2.	s z zh th dh jh t d n ch
3.	ng hh g k sil epi bcl dcl gcl pcl tcl kcl qcl
4.	l uh aw ow uw
5.	w
6.	eh er
7.	ao
8.	aa
9.	ay ey ih oy
10.	ax ix
11.	ae
12.	ah
13.	iy

TABLE VI. Development of the recognizer with discrete hidden Markov models. Phone error rate includes insertion errors

Discrete models	Plosive errors	Vowel errors	Misrecognized rate	Phone error rate
48 phonemes	199	508	42·88	47·06
359 models, left contexts	202	461	40·25	44·51
364 models, left and right contexts, seven-state topology	192	465	39·63	44·35

TABLE VII. Results for continuous models (no language model)

Continuous models	No. of parameters	Misrecognized rate	Phone error rate
48 phonemes	74 880	38·36	44·06
364 allophones, left and right contexts	473 304	34·32	40·76

A summary of results for discrete models using various types of context-dependent allophones is shown in Table VI. The solutions from all the searches were combined, and tested on a new set of 96 sentences using discrete-codebook HMMs. The left contexts were expanded to include other consonant classes, using the context clusters of Table V. Next, the left-context vowel and consonant models were combined with the right-context plosives. As hoped for, higher accuracy by class averaged out to a higher overall unit accuracy. Finally, these same context sets were used to train models with topologies based on the earlier optimization experiments. The results are in the last row of Table VI.

Results for the continuous-parameter models are summarized in Table VII.

TABLE VIII. Merging the discrete hidden Markov models

Discrete models	Plosive errors	Vowel errors	Misrecognized rate	Phone error rate
53 phonemes, merged distributions	173	489	38·90	43·37
53 phonemes, simplified	157	464	38·56	43·43

TABLE IX. Merged continuous context-independent models

Continuous models	No. of parameters	Misrecognized rate	Phone error rate
53 phonemes, simplified	356 304	34·8	40·0
53 phonemes, with additional features	397 416	33·8	39·7
53 phonemes, Fig. 2 topology	301 252	32·7	38·2
Same, with bigrams		30·8	35·5

4. Derivation of phoneme models from allophone models

4.1. Merging and retraining

To this point, the result of the search process was a set of unit model topologies and output distributions well trained for units in left or right context. The final experiment attempted to simplify the speech recognition system by merging the distributions of the allophones into phoneme units with parallel transitions. This was done for both discrete and continuous-distribution models.

In the discrete case, the n allophones for unit U were merged in a straightforward way. All allophones for U had the same topology **T**. A new topology was created with n parallel transitions for each single transition in **T**. Each of these parallel transitions was tied to the corresponding distribution in one of the allophones for U. Once these *merged*-model phonemes were built, they were then retrained for four iterations. These models had a better phone error rate than the best allophone models (Table VIII).

In order to simplify the models and further reduce the number of parameters, all the parallel transitions but one of the internal states of the merged models were removed, on the hypothesis that the extra distributions were only useful for modelling the contextual effects at the left or right end of the units. In fact, the simplified models showed improvements, after four training iterations, with respect to their predecessors.

The end result of the search process was to construct a set of discrete models with internal topological structure complex enough to model contextual effects of neighbouring units significant to the particular unit. Results are summarized in Table VIII. It appears that a substantial performance improvement can be obtained in context-independent models by just adding context-dependent distributions on the initial and final transitions of phoneme HMMs.

The same merge/simplify procedure was applied to the continuous models. Continuous-distribution HMMs already employ density mixtures. In this case the process of combining allophone distributions essentially means selecting mixture sizes for context-independent models and initializing the corresponding distributions with well-trained values. Results are summarized in Table IX.

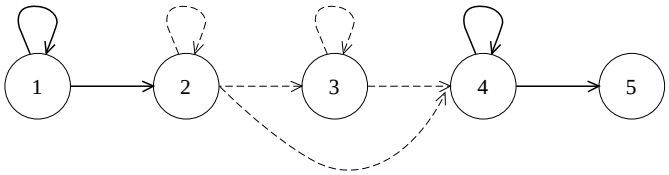

Figure 2. Final topology. (——), Large mixtures: modelling contextual effects; (– – –), small mixtures: modelling the internal states.

4.2. Further improvements

An interesting research avenue consists in using new acoustic parameters which are likely to contain information not well represented by MEL cepstral coefficients. A preliminary investigation has been conducted by considering simple measures in the time domain and in broad frequency bands.

Following experiments described in De Mori *et al*. (1976) the signal energy can be described in terms of peaks and valleys. Let an "event" be a peak or a valley. Let t_{b_j} be the beginning time of the j-th event $e(j)$. A temporal feature

$$temp(t) = t - t_{b_j} \tag{4}$$

is computed where t_{b_j} is the beginning time of $e(j)$ that covers a time interval including t. A value $\mu(t)$ is then computed as follows:

$$\mu(t) = \begin{cases} temp(t) & \text{if } temp(t) < \mu_0 \\ \mu_0 + \log_2 (temp(t) - \mu_0) & \text{otherwise} \end{cases} \tag{5}$$

where μ_0 is a constant chosen *a priori*. The feature $\mu(t)$ suggests the position of the t-th frame in the suprasegmental acoustic event t belongs to.

Two other features are obtained by introducing $e_1(t)$, the energy of the highest spectral value in the 100–900 Hz band at time t, and

$$\Omega_{31}(t) = \log_{10} \frac{e_3(t)}{e_1(t)} \tag{6}$$

where $e_3(t)$ is the highest spectral value at time t in the 3–5 kHz band. Performances were further improved by adding the new acoustic features, as shown in Table IX.

A final experiment was performed using the same model topology, shown in Fig. 2, for all complex phoneme models (vowels, liquids, nasals and plosives). In this topology, the transitions represented by thick lines are modelled with a moderately large number of Gaussian distributions, while the transitions represented by thin lines are modelled with small mixtures. The relative sizes of these mixtures reflect the number of contexts in which the allophones described earlier were trained.

The initial distributions were taken from the well-trained merged models described

in Subsection 4.1. These distributions were duplicated or reduced in number, so that each transition from the first state of the new topology could be tied to a mixture of 39 densities, each mixture tied to the fourth state could have 30 densities, and all the internal "thin" distributions could be mixtures of six probability density functions. The new models were then retrained for five iterations, and low-probability transitions were pruned. The results (third row in Table IX) confirm the importance of good initialization of the parameters before estimation.

Taking into account the comparatively modest feature set employed here, and the context-independence of the resulting models, these results compare favourably with those reported in the literature. Currently, the best reported phone-recognition rates on TIMIT (Robinson, 1991; Young & Woodland, 1994) are achieved using second-order derivatives and various kinds of context-dependency learning.

5. Discussion

Knowledge-guided, performance-driven randomized search has proven effective in optimizing structural features of a machine-learning model. This optimization proceeds along two coordinates: improving the performance measure, while reducing the number of parameters in the model. The search should be guided by knowledge about the problem domain, in order to improve its efficiency.

The search for an optimal HMM topology can be guided by measures of model likelihood or recognition accuracy. Average likelihood of the model set was found to be unpredictive. Another possibility would be to optimize a unit topology based on the *individual* model likelihood. It would be instructive to look for similarity in the topologies evolved through these methods, with those of Takami and Sagayama (1992) and Sanchis and Casacuberta (1991). All of these results could be compared another way: training an *ergodic*, fully connected model on sufficient data so as to reduce the probabilities of the unnecessary state-transitions to near zero. Once this model is edited to remove low-likelihood paths, it should possess a near-optimal structure. Although it would be difficult to train unit models sufficiently due to insufficient data, it is possible in this way to compare the methods in producing a general topology. Experiments reported in this paper do not encourage such efforts.

The problem of clustering allophone models has been summarized by Hon (1992) with three considerations: consistency, trainability and generalization. The first two considerations are a trade-off: to make sure the units selected are consistent (i.e. perform well), the training must be as context-specific as possible. The difficulty is acquiring a sufficient number of training samples to produce a consistent model. Consistency will also depend on building models that are sufficiently distant from each other. Thus, the different metrics that have been used to cluster models include training-sample size (Lee *et al.*, 1991), model likelihood (Bahl *et al.*, 1991; Takami & Sagayama, 1992), acoustic-phonetic reasoning, model cross-entropy (Lee *et al.*, 1990), and recognition performance. The method of this paper, randomized search through cluster-space driven by performance, implicitly optimizes all of these measures, though somewhat inefficiently.

The last consideration, generalizability, is that the trained units can provide good candidate models for new units not present in the training data. This becomes an important issue when an ASR system based on subword units is presented with new vocabularies to model. The clusters developed in this paper introduce a useful degree of generalization.

In this research allophone models were initialized with the distributions of context-independent models, and vice versa. For both consistency and generalization of subword units, it is better to perform an interpolation of new subclusters with the more general, better trained phoneme model from which they are derived. This may improve performance.

The results described in this paper indicate that a simple model for phonemes containing rich mixtures of Gaussian distributions is a serious competitor for context-dependent allophones. Choosing a variable number of Gaussians has a positive impact, especially if parameter estimation is initialized with the results of previous optimizations. Furthermore, this performance can be sustained even with a reduced number of mixtures if parameters are again initialized with the results of previous optimizations. Having a large number of Gaussians per mixture does not imply a high computation time if the number of different mixtures is not large. In fact, the probability of a distribution can be approximated with the probability given by the Gaussian that makes the biggest contribution to the mixture for a given observation. A given observation can be compared with intervals computed off-line, each interval corresponding to a Gaussian that will eventually give the highest contribution to the mixture. After a fast binary search, only the computation of a value in a single Gaussian distribution per mixture has to be performed.

Finally, simple broad-band acoustic parameters and temporal features appear to have a significant positive impact as well.

This research was sponsored by the National Science and Engineering Council of Canada, and the Institute for Robotics and Intelligent Systems. The authors are grateful to G. Antoniol, M. Omologo and D. Giuliani from IRST (Italy) for their part in developing the base HMM software.

References

Bahl, L. R., deSouza, P. V., Gopalakrishnan, P. S., Nahamoo, D. & Picheny, M. A. (1991). Decision trees for phonological rules in continuous speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 185–188.

Casacuberta, F., Vidal, E., Mas, B. & Rulot, H. (1990). Learning the structure of HMMs through grammatical inference techniques. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 717–720.

Chow, Y. L., Schwartz, R., Roucos, S., Kimball, O., Price, P., Kubala, F., Dunham, M., Krasner, M. & Makhoul, J. (1986). The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1593–1596.

Hon, H. W. (1992). Vocabulary-Independent Speech Recognition: The VOCIND System, PhD Thesis, CMU-CS-92-108, Carnegie Mellon University.

Huang, X. D., Ariki, Y. & Jack, M. A. (1990). *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh.

Jouvet, D., Mauuary, L. & Monne, J. (1991). Automatic adjustments of the structure of Markov models for speech recognition applications. *Eurospeech* September, pp. 927–930.

Kimball, O., Ostendorf, M. & BechWati, I. (1992). Context modeling with the stochastic segment model. *IEEE Transactions on Signal Processing* **40**, 1584–1587.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R. & Rosenberg, A. E. (1991). Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 161–164.

Lee, K. F. & Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**, 1641–1646.

Lee, K. F., Hon, H. W. & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing* **38**, 35–44.

Lee, K. F., Hayamizu, S., Hon, H. W., Huang, C., Swartz, J. & Weide, R. (1990). Allophone clustering for continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 749–752.

De Mori, R., Laface, P. & Piccolo, E. (1976). Automatic detection and description of syllabic features in continuous speech. *IEEE Transactions on Acoustics, Speech and Signal Processing* **24**, 365–379.

Rabiner, L. R., Lee, C. H., Juang, B. H. & Wilpon, J. G. (1989). HMM clustering for connected word recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 405–408.

Robinson, T. (1991). Several improvements to a recurrent error propagation phone recognition system. Technical report TINFENG/TR.82, Cambridge University Engineering Department.

Sanchis, E. & Casacuberta, F. (1991). Learning structural models of subword units through grammatical inference techniques. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 189–192.

Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M. & Makhoul, J. (1985). Context-dependent modeling for acoustic-phonetic recognition of continuous speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1205–1208.

Takami, J. & Sagayama, S. (1992). A successive state splitting algorithm for efficient allophone modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I573–576.

Young, S. J. (1992). The general use of tying in phoneme-based HMM speech recognizers. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I569–572.

Young, S. J. & Woodland, P. C. (1994). State clustering in hidden Markov model-based continuous speech recognition. *Computer Speech and Language* **8**, 369–383.

(*Received 7 July 1993 and accepted for publication 2 January 1995*)