BOTTOM-UP AND TOP-DOWN STATE CLUSTERING FOR ROBUST ACOUSTIC MODELING

C. Chesta ⋆ and P. Laface ⋆ and F. Ravera ◇

⋆ Dipartimento di Automatica e Informatica – Politecnico di Torino
Corso Duca degli Abruzzi 24 – I–10129 Torino Italy
E-mail laface@polito.it chesta@polito.it
◇ CSELT- Centro Studi e Laboratori Telecomunicazioni

Via G. Reiss Romoli 274 – I–10148 Torino, Italy
E-mail ravera@cselt.it

ABSTRACT

In this paper we describe our experience with bottom-up and top-down state clustering techniques for the definition and training of robust acoustic-phonetic units. Using as a test-bed a speaker-independent telephone-speech isolated word recognition task with a vocabulary including 475 city names, we show that similar performances are obtained by tying the HMM states both with an agglomerative or a decision-tree clustering approach. Moreover, better results are obtained by a priori selecting the set of states that can be clustered, rather than relying solely on their acoustical similarity. In the bottom-up approach a stopping criterion for the furthest neighbor clustering procedure is proposed that does not require a threshold. In the top-down approach we show that a careful selected impurity function allows lookahead search to outperforms the classical decision tree growing algorithm.

1. Introduction

It is well known that an important issue in acoustic modeling is to select a set of basic units that can be accurately modeled with the available training data, but that are also robust to phonetic contexts rarely or never appeared in the training database. In the last few years, we have proposed as an alternative to the classical acoustic modeling with biphones and triphones, a set of stationary/transitory state units [2, 3]. The relationships between these units and the triphones are given in Table 1. The central state of a triphone <(l)p(r)> is tied to the central state of all the other triphones of the same phone <p>. The final state of phoneme <p> is connected to the first state of the next phoneme <q> leading to a two-state diphone-transition unit <pq>.

We have reported in [2] that a recognition system employing these new units favorably compares with respect to a recognizer with Continuous Density Hidden Markov Models of context-dependent biphones and triphones, selected through a minimal occurrence criterion within the training database.
Since the new units are obtained by a process of tying that is purely based on a priori acoustic-phonetic knowledges and assumptions, in this work we have studied and compared the effects of including bottom-up or top-down acoustic driven tying [7, 1, 6].
State tying is an intermediate step in our standard training procedure that proceeds as follows:

- A set of 3 state left-to-right triphones is trained using the Viterbi segmental K-means algorithm: every state has associated its continuous density mixture with a variable number of Gaussians selected according to the technique presented in [2]. The observations contributing to the occupation count of each state are also recorded.
- Using these segmentations, new models are estimated with a single Gaussian associated to each state.
- For each set of triphones with the same base phone, the corresponding states are clustered by means of the procedures described in the following.
- New mixtures are re-estimated for each tied state.

2. A priori tying

It is worth noting that an a priori tying decision is made when the set of base phones is selected. We decided on the basis of preliminary experiments to not discriminate a priori between the stressed and unstressed vowels, or between the single and geminate consonants in

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Table 1: Context-independent and diphone-transition units

Phoneme sequence	$\ldots xpy \ldots$								
Triphone States	x_l	x_c	x_r	p_l	p_c	p_r	y_l	y_c	y_r
Diphone-transitions	$\ldots$		$< xp >$			$< py >$		$\ldots$	
Context-independent phonemes	$\ldots$	$< x >$			$< p >$			$< y >$	$\ldots$

our 52 base phone set. The clustering procedures, however, may account for these discriminations whenever they are suggested by the training data.
On the other hand, it is worthwhile, on the basis of the considerations given in [2] and of the experimental evidence, to constrain the set of states that can by tied. In particular, we allow the central states of the same allophones to be tied, while lateral states can only be tied within the same diphone-transition. Thus, if $(*)$ stands for every context, the first state of triphone $< (l)p(r) >$ can be tied with the first state of triphone $< (l)p(*) >$, while this is not the case for triphone $< (l')p(*) >$, where $l \neq l'$. Similarly, the final state of triphone $< (l)p(r) >$ can only be tied with the corresponding state of the set of triphones $< (*)p(r) >$.

3. Bottom-up clustering

We use a classical bottom-up clustering process which merges states that are similar according to the furthest neighbor criterion. This procedure is typically controlled by two thresholds: the maximum distance between the members of a cluster, and the minimum number of observations associated to the cluster.
In order to have a fair comparison with the results obtained with different distance measures, different constraints, or with top-down approaches, we use as a stopping criterion a threshold on the total number of clusters, rather than on the maximum cluster size. As an alternative, we propose a stopping criterion that does not require a threshold. The "optimal" number of clusters is automatically derived observing the behavior of the current best distance between clusters as a function of the number of merge operations that have been performed: as shown in Fig. 1, a rapidly increasing distance is a clear indication that further clustering is inappropriate. To automatically locate this threshold we search the minimum of the difference between the distance and a linear function joining the average distance values after the 10 first and 10 last merges respectively. Since the maximum number of merges is reached when all states are clustered within the a priori clusters, their number is given by the difference between the total number of states and the number of initial clusters. The total number of states and the number of initial clusters for the experiment illustrated in Fig. 1

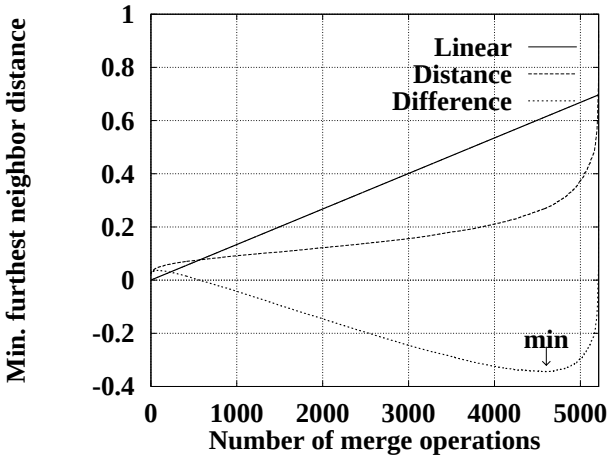

Figure 1: Minimum furthest neighbor distance as a functions of the number of merge operations

are 5862 and 575 respectively.

4. Top-down clustering

In the decision tree techniques, a tree is usually built for each state of every phone. The linguistic questions asked typically refer to the identity of the surrounding contexts.
In our approach, the questions asked for the first state of a phone refer primarily to its left context, while they refer to the right context for the last state. For the central state, the questions include both phone contexts. This asymmetry approximates our restriction requiring state tying within the same diphone-transition.

Lookahead Search

The log likelihood $L_{P_A}^{P_A}$ that a tree node (P), modeled by means of a k-dimensional Gaussian, generates the n_{P_A} training observations in set A that have been used for evaluating its parameters can be effectively computed as follows:

$$L_{P_A}^{P_A} = \sum_{n=1}^{n_{P_A}} \log \mathcal{N}(\overline{x}_{An}; \overline{\mu}_{P_A}, \overline{\sigma}_{P_A})$$

$$= -\frac{n_{P_A}}{2} \sum_{i=1}^{k} \left[\log\left(2\pi\sigma_{P_A i}^2\right) + 1\right] \qquad (1)$$

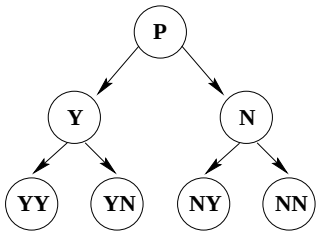

Figure 2: Yes-No nodes decision tree

Let's as refer to this "self" log likelihood of a parent node as L_P, and to L_Y and L_N to the corresponding likelihoods of its two children.
According to the standard technique, a tree is grown selecting the question that maximizes the increase of log likelihood at a node $\Delta L_P = L(Y) + L(N) - L(P)$. Since the procedure is driven by a local decision, it is not optimal. To alleviate this problem, in [5] it has been proposed to extend the locality of the optimization, comparing, for each question, the impurity of a node with that of its grandchildren. This technique did not improve the performance of their system. Our results confirm their findings, but not their conclusions: an analysis of the behavior of the decision tree growing algorithm for acoustic models suggests that the lookahead search must be advantageous at the beginning of the divisive process, otherwise the entire sequential decision approach would be ineffective. Going on with the divisive process, however, it is better to rely upon local optimal decisions to reduce the risk of generating trees with reduced generalization capabilities. In our lookahead approach, thus, the contribution of the grandchildren nodes is reduced as a function of their level in the tree, by selecting at each node the question that maximizes the increase of log likelihood:

$$\Delta L = \Delta L_P + \frac{(\Delta L_Y + \Delta L_N)/2}{tree\_level(P)}$$

where $tree\_level(root) = 1$, and $\Delta L_P, \Delta L_Y$, and ΔL_N are defined, with reference to Fig. 2, as follows:

$$\begin{aligned} \Delta L_P &= L(Y) + L(N) - L(P) \\ \Delta L_Y &= L(YY) + L(YN) - L(Y) \\ \Delta L_N &= L(NY) + L(NN) - L(N) \end{aligned}$$

Pruning

Using two training sets A and B, we compared the iterative growing and (bottom-up) pruning algorithm [4] with a pure top-down pruning approach. In both cases, pruning decisions are made on the basis of the performance of the tree models on new data.
If set B includes the n_{X_B} data that have been used to train the parameters of node X, it can be shown that Equation (1) can be generalized to evaluate the prediction accuracy of node P (trained with the data in set A) on set B in terms of the log likelihood:

$$\begin{aligned} L_{P_A}^{X_B} &= \sum_{n=1}^{n_{X_B}} \log \mathcal{N}(\overline{x}_{Bn}; \overline{\mu}_{P_A}, \overline{\sigma}_{P_A}) \\ &= -\frac{n_{X_B}}{2} \sum_{i=1}^{k} \log\left(2\pi\sigma_{P_A i}^2\right) + \frac{\sigma_{X_B i}^2 + \mu_{X_B i}^2 - 2\mu_{X_B i}\mu_{P_A i} + \mu_{P_A i}^2}{\sigma_{P_A i}^2} \end{aligned}$$

This evaluation is effective because it is solely based on the statistics computed during the previous Viterbi alignment procedure.
Rather than pruning, as usual, the children of a node P if its log likelihood is better than the sum of the log likelihoods of its (Y and N) children, i.e., if

$$L_{P_A}^{P_B} > L_{Y_A}^{Y_B} + L_{N_A}^{N_B} \tag{2}$$

to avoids the generation of offsprings with largely unbalanced likelihoods, we prune the children if one of them does not predict the new data better than its parent node, i.e.,

$$L_{P_A}^{Y_B} > L_{Y_A}^{Y_B} \quad \text{or} \quad \mathrm{L}_{\mathrm{P_A}}^{\mathrm{N_B}} > \mathrm{L}_{\mathrm{N_A}}^{\mathrm{N_B}} \tag{3}$$

Two strategies that do not make use of bottom-up pruning have also been experimented. The first one grows *in parallel* the state decision trees of every phone. It selects the question that maximizes the increase of log likelihood and stops expanding nodes when their number reaches a preset threshold. The second strategy, similarly to [4], grows a tree at a time, iterating between the alternative training sets. Tree expansion and pruning, however, is performed "top-down" since the children of a node must satisfy condition (3) and a minimum occupation count for terminal nodes.

5. Results

The above mentioned methods have been tested on a isolated word recognizer with a vocabulary of 475 city names, that has been exploited to develop a telephone service that provides information about the main railway connections. The training database includes a total of 21000 words pronounced by 2101 speakers, another 14400 utterances that were collected by 1050 speaker are used for testing. The vocabulary words are transcribed using 2024 triphones of 52 base phones. Table 2 summarizes the baseline results and the improvements obtained using bottom-up clustering. The first two columns refer to the baseline system using triphones and the diphone transition units defined in [2]. The importance of a priori constraining the set of states that can by tied is evident comparing the results in the

Table 2: Bottom-up clustering results

Set of units	Triphones	Transitions	No A Priori Selection	A Priori Selection	Retrain	No thresholds
Num. of states	5862	644	1287	1287	1287	1185
% error	5.44	5.11	5.08	4.44	4.32	4.21

Table 3: Top-down clustering results

Method	No A Priori Sel.	A Priori Selection	Lookahead	Weighted Lookahead	Top-down pruning
Num. of states	1272	1272	1272	1272	1324
% error	4.74	4.56	4.88	4.33	4.33

third and fourth column. It is worth noting that the reported results have been obtained without retraining, i.e., relying on the initial segmentation. Slightly better results are obtained after retraining, as shown in the fifth column. The results reported in the rightmost column refer to the stopping criterion for the furthest neighbor clustering that does not require a threshold: in this experiment the clustering procedure automatically stops after 4603 merges have been performed and gives better results than the classical stopping criterion based on cluster size.

As can be seen in Table 3, the best results obtained with the divisive and with the agglomerative clustering approaches are similar. These results refer to our proposed state decision trees growing strategies. For the results reported in the first four columns the stopping criterion is obviously a total number of 1272 nodes, while last columns refers to the stopping criterion related to condition (3). It is worth noting that these criteria give almost the same results, outperforming the classical bottom-up pruning approach that gives instead, even with weighted lookahead search, a set of 1810 tied states and an error rate of 4.56%.

The first two columns allow one to assess the importance of a priori selecting the states that can be tied. This selection is implicitly obtained because the questions that are asked at the beginning of the expansion process refer, for lateral states, to their related context. Comparing the result reported in the "Lookahead" columns it can be seen that lookahead search is most effective for the nodes at the first levels of the tree, and that the impurity function must be chosen accordingly.

6. Conclusions

We have proposed and compared some procedures for agglomerative and divisive state clustering that achieved an the error rate of 4.3% on a 475 city names recognition task with a set of models including 1300 states. The use of tying allowed the error rate of the system to be improved both with respect to 5.44% obtained with the set of most occurring triphones including 5800 states, and to 5.11% of the diphone-transition units with 644 states.

7. References

[1] L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, M. Picheny, "Decision Trees for Phonological Rules in Continuous Speech", Proc. ICASSP 91, pp. 185–188.

[2] L. Fissore, F. Ravera, P. Laface, "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", Proc. EUROSPEECH 95, pp. 799–802, 1995.

[3] L. Fissore, P. Laface, G. Micca, F. Ravera, "Vocabulary Independent Acoustic-Phonetic Modeling For Continuous Speech Recognition", Proc. EUSIPCO 96, Trieste, Italy, 1996, pp. 1615–1618.

[4] S. Gelfand, C. Ravishankar, E. Delp, "An Iterative Growing and Pruning Algorithm for Classification Tree Design", IEEE Trans. PAMI, vol. 13, n. 2, 1991, pp. 163–174.

[5] A. Lazaridès, Y. Normandin, R. Kuhn, "Improving Decision Tree for Acoustic Modeling", Proc. ICSLP 96, Philadelphia, 1996, pp. 1053–106

[6] S. Young, J. Odell, P. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling", Proc. ARPA Workshop on Human Language Technology, 1994, pp. 307–312.

[7] S. Young, P. Woodland P, "The Use of State Tying in Continuous Speech Recognition", Proc. EUROSPEECH 1993, Berlin, 1993, pp. 2207–2210.