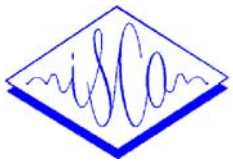
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Combining Agglomerative and Tree-based State Clustering for High Accuracy Acoustic Modeling

Zhaobing Han, Shuwu Zhang, Bo Xu

National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100080
{zbhan,ljia,swzhang,xubo}@hitic.ia.ac.cn

Abstract

Robust estimate of a large number of parameters against the availability of training data is a crucial issue in triphone based continuous speech recognition. To cope with the issue, two major context-clustering methods, agglomerative (AGG) and tree-based (TB), have been widely studied. In this paper, we analyze two algorithms with respect to their advantages and disadvantages and introduce a novel combined method that takes advantage of each method to cluster and tie similar acoustic states for highly detailed acoustic models. In addition, we devise a two-level clustering approach for TB, which uses the tree-based state tying for rare acoustic phonetic events twice. For LVCSR, the experimental results showed the performance could be highly improved by using the proposed combined method, compared with those of using the popular TB method alone.

10.21437/Interspeech.2004-166

1. Introduction

Most state-of-the-art large vocabulary conversational speech recognition systems use context-dependent models and continuous density HMMs to model speech data. The context-dependent (CD) models, typically triphones with their suitability for capturing co-articulation effects, are currently the most powerful subword units. However, constructing the acoustic models for triphones has not been an easy task due to the scarcity of data. It requires to maintain a good balance between model complexity and available training data [1]. In order to meet this requirement, most systems use distribution clustering to define a set of distributions which are shared among all possible context-dependent models.

In earlier implementations of distribution tying, an agglomerative state clustering [2] was widely used. Clustering follows two stages. Firstly, similar acoustic states are merged, and then any cluster with insufficient training data is merged with its nearest neighbor. The sparse or unseen context backs off to diphone or monophone. Moreover, to escape from the local optimal configuration, an element reshuffling step is performed after the merge step [3], because otherwise, the algorithm would become a greedy algorithm, where an earlier decision could not be undone. The advantage lies in that it is not restricted by the prior phonetic questions and it gains relatively accurate clustering results. But this method has no prediction power 1and backs off to the less CD models for the unseen triphones.

In more recent work, phonetic decision tree state tying [3][4] based acoustic modeling has become increasingly popular for modeling speech variations in LVCSR. The decision tree that is a binary tree classifies Markov states of triphones represented in the training data set by asking linguistic questions composed of conjunctions, disjunctions, and /or negations of pre-determined categorical linguistic questions [5]. The tree construction is a top-down process based on a one-step greedy tree growing algorithm. Its two major advantages are the abilities to synthesize unseen models and to maintain the model complexity that a number of parameters are robustly estimated. One disadvantage of the tree-based top-down classification is that once a split is made, individual elements cannot be moved to new positions in the tree. This may result in a sub-optimal clustering configuration. Another drawback is that the tree construction is constrained by the prior phonetic questions. To some extent the question set may not be coincident with the set that can be best captured by a given model. For example, in telephone speech, where much of the high-frequency information is lost, it may not be optimal to use the same variety of fricatives as used for full-bandwidth speech [6].

A combined method is introduced in this paper that incorporates the merits of the above approaches. It tries to achieve accurate clustering via data driven algorithm while attain solutions of unknown contexts via decision tree algorithm. In other words, we combine the agglomerative algorithm and tree-based algorithm to construct the CD models instead of using either of them alone. AGG and TB are applied to train the acoustic CD models separately. After that, the maximum likelihood (ML) optimization principle is used. The model with the better likelihood improvement in one context is selected. Additionally, some methods for AGG or TB state tying to improve the robustness and accuracy of acoustic modeling are described. At last, the splitting criteria of TB using the entropy-based and likelihood-based measures are investigated. Experimental results on large vocabulary Mandarin speech tasks indicate that the proposed approach outperforms the conventional single clustering algorithm, especially when the prior question sets are inappropriate (telephone and spontaneous speech) or might not exist at all, be scarce or expensive (multilingual speech).

In this paper, we will firstly introduce AGG and TB algorithm briefly, and then explain how to combine them to pick out more accurate model for a given context. Afterwards, extensive experiments will be carried out to test the modeling capability of the proposed method to model the co-articulation effects. Finally, some experimental conclusions will be drawn.

2. The Agglomerative Algorithm

The AGG state clustering is also referred as data-driven or bottom-up algorithm. The AGG algorithm works as follows:

1) Create a singleton cluster for each output distribution of

This work was partially supported by Digital Olympic Project of Beijing Committee of Science and Technology (H030130050430).

each triphone.

2) Use vector quantization (VQ) algorithm to cluster the acoustic space into VQ regions if the occupation count of each singleton context is less than the predefined threshold. Then, estimate all initial triphone HMMs.
3) The pair of clusters that are the most similar are merged. The dissimilarity measures based on multiple-mixture Gaussian distance will be explained in formula (1). Then, re-estimate the clustered triphone HMMs.
4) Re-shuffle: Move one element from one cluster to another if it results in a decrease in the total entropy of all clusters.
5) Repeat step (3) until some convergence criteria are met.

In [3], Huang pointed out that the probability density functions of HMM states should be shared using information divergence clustering. Following [7], we use the increase in the weighted-by-counts entropy of the mixture-weight distributions that is caused by the merging of the two states. Then, the distortion that occurs when state s1 and s2 are clustered together into the clustered states is defined as:

$$d(s_1, s_2) = (n_1 + n_2)H(s) - n_1 H(s_1) - n_2 H(s_2) \qquad (1)$$

Where n1, n2 represent the number of observations of states s1 and s2, H(s) denotes the entropy of the distribution.

There are two differences between the work here and the work in [2][3]. First of all, VQ cluster is used for the initial spare data context. Secondly, multiple-mixture Gaussian distance measure between two clusters is applied instead of the single mixture Gaussian. In the initial clustering when samples of most states are spare, VQ measure is more robust than Gaussian. During merging nodes, the single mixture Gaussian reduces the computational complexity, but it is very coarse and not consistent with the recognition because the multiple mixture Gaussian distribution is often used in the decoder process.

One of the most positive aspects of AGG clustering method is its high reliability of clustered classes whose similarity is evaluated from the training data set. In this sense, even though two context-dependent models exhibit similar linguistic rules, they may be clustered to a different class if the distance between them is not the nearest [8].

3. Decision Tree-based Clustering

Recently, a prominent method to build context-dependent acoustic model in LVCSR is the tree-based state cluster. TB cluster is also referred as rule-based or top-down algorithm. There are many attempts to improve the phonetic decision tree state tying based on the approaches in acoustic modeling [9], [10], [11]. Decision tree consists of three essential components, which substantially affect the performance of continuous speech recognition: (1) specification of question sets, (2) criterion for choosing the best splitting question and stopping the node splitting, and (3) criterion for pruning the over-large tree. We focus on (1) and (2).

The construction of decision tree is as follows:

1) Prepare a set of prior manual phonetic questions.
2) Initialized the decision tree from the root node containing all context-independent observation frames associated with the same phonetic unit. The root node is labeled non-expanded node.
3) Find the best question (A proper evaluation function is used to determine the “best” question for node splitting) for the non-expanded node. Then the node splits into two child nodes, corresponding to the answers ”Yes” and “No” respectively. At last, the node is marked expanded node.
4) Go to step 3 unless some stop-growing criteria are met.

3.1. The Question sets

In our Mandarin LVCSR, we adopt a commonly used approach in which a relatively large and collective set of questions is created and the questions to be used are determined ultimately by acoustic training data. Basically manner and place of articulation of the immediately adjacent phone are of primary consideration when classifying phonetic context. The resultant Mandarin question set contains 37 and 41 questions for left and right contexts respectively.

3.2. Evaluation Function

Since the entire potential question set has been prepared, the task is how to find the best question for a node splitting. The choice of the best question is equivalent to finding the best split for data samples of the node. For continuous PDF, the use of ML in divisive clustering can guarantee an increase in the likelihood of the data, so it is applied in TB method since there is no straightforward entropy measurement.

When the tree is continuously growing, it is necessary to adopt a stopping criterion to terminate the node splitting. Therefore, if the increase in the likelihood or the sample counts are less than certain thresholds, the node will be marked “LEAF” and no splitting will be executed.

4. Proposed Algorithm

As pointed out in the above sections, the data-driven method is unable to cluster unseen contexts though it is more accurate. On the contrary, the rule-based method has relatively less accuracy although it is able to cluster unknown contexts. The pros and cons of the bottom-up clustering versus top-down classification can be summarized as follows:

State clustering	quality for seen triphones	quality for unseen trriphones	question sets
AGG	nearly optimal	monophones or diphones	non-sensitive
TB	sub-optimal	decision-tree based senones	sensitive

In this section, we propose a new modeling construction method that combines two clustering methods to remedy the limitations of either of them. The practical algorithm execution steps in detail are described herein.

4.1. The AGG Clustering Process

Firstly, VQ algorithm is used to cluster the spare sample. Secondly, the pair of the most similar clusters is merged by exploiting the multiple mixture Gaussian measure. AGG can’t predict the unseen model and must back off to the less context dependent model that can be estimated more reliably. So the monophone and diphones are both trained. In our experiment, we directly back off to the less CD models, not using the deleted-interpolation or Bayesian prediction algorithm.

4.2 The Tree-based Clustering Process

In this section, we present a new robust two-level clustering method. It consists of an initial grouping of rare triphones and

a subsequent state tying, using decision tree clustering twice. Robust estimation of rarely seen triphones is another critical issue. The rarely seen tripones often provide less contribution to select the best questions. The best question selected during the node splitting may not be optimal for the rarely seen tripones. Consequently, we first cluster them into a low sample count threshold for robust estimation. In [9], AGG algorithm is applied to cluster the sparsity triphones. As we all know, when the data are spare, the rule-based method is often superior to the data-based approach. Our experimental results also proved that the TB algorithm is empirically more suitable to deal with data sparsity problem. At last, the comparison of three goodness-of-split evaluation criteria (ML, Entropy and MCE) is investigated. The experimental results will be given below.

4.3 Select the Most Optimal Model

When combining the AGG and TB, the criterion used to select the models is based on ML. Fig.3 illustrates the algorithm to find the best senone model for a given context. The shared parameter (i.e., the output distribution) associated with a cluster of similar states is called senone [3].

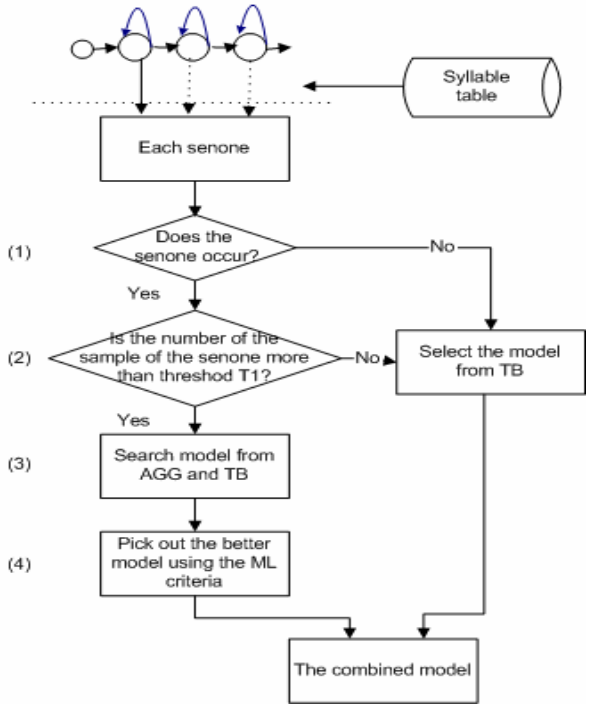


Fig1. The process of selecting the model

Two aspects should be noted for Fig.1:

1) At step (3), for the unseen triphone, we cannot use its sample to evaluate the models by ML, then directly select the TB or AGG models. In the literature, it was believed that with large amounts of training data, triphones outperform less CD units; but that with small to moderate amounts of training data less CD units perform better. So with large amounts of training data, the triphones of TB are picked out. Otherwise, step (3) will be modified to select the less CD units of AGG.
2) If the senones have very few training samples, only TB model is picked out instead of the combined model in order to deal with the over-fit problem of the data-driven method.

The hybrid approach [13] is analogous to our combined method. But it suggests that seen triphones obtain the best sharing structure with AGG, while unseen triphones traverse the decision-tree to locate the generalized allophone it belongs to, then share senones according to the state quantization result. Therefore, the mis-average problem associated with generalized allophones and state-quantization error for unseen triphones are unavoidable.

5. Experimental Results

The accuracy and efficiency of the proposed method are evaluated by the recognition task of clean reading, telephone quality and spontaneous speaker-independent continuous Mandarin speech.

There are three types of benchmark speech data that are trained and tested respectively. For the clean reading speech, the acoustic data is the standard set (863Train), which is provided by China National Hi-Tech Project 863 (CNHT863) and contains 48366 sentences from 83 different male. Its test set (863Test) consists of 240 sentences by 6 male speakers from CNHT863, 40 sentences per speaker. The telephone quality training set (863TelTrain) and test set (863Tel) are derived from disposing the first database with two methods: 1) resample the database to 8000Hz and u-law quantization; 2) pass the database through the real PSTN network by Dialogic telephone cards plugged in PCs. The telephone quality and spontaneous database is the Mandarin CALLHOME corpus. Its training database has 50 conversation sides, about 6 hours with 14,995 sentences. And the evaluation database has 8 conversations sides, about 1.2 hours with 2317 sentences. CALLHOME corpuses consist of many kinds of background sound, such as cough, laugh, Lombard, channel noise, and many English speeches. To cope with background sound, a garbage model is trained.

The acoustic features consist of energy, pitch, 12 mel-cepstral with delta and delta-delta features. Cepstral mean subtraction and variance normalization are performed. The pitch is extracted through the autocorrelation algorithm with DP algorithm to smooth. Tri-gram statistics is used for language modeling.

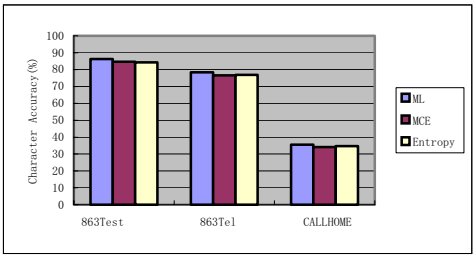


Fig2. Comparison of ML, MCE and Entropy criteria

5.1 Comparison of Several Criteria in TB

First of all, we evaluate the node splitting criteria using ML and discriminative methods (MCE and Entropy). In Fig2, we find ML consistently obtains higher recognition rates than those of MCE and Entropy for different speech data.

5.2 Improvement of Tree-based Clustering

In Tab1, Two-level_1 denotes the method in [9], and Two-level_2 to the proposed approach. The experimental results of

the two-level cluster confirmed that using the proposed two-level decision tree has slight improvement over Two-level_1. We think that the discriminative question sets can be used to cluster the sparsity triphone whose a few seen samples may be inadequate to represent the distributions of its own state. In this condition, the rule-based clustering is more effective than the data-based clustering.

Tab1 Character error rate (CER) for two-level optimal TB

	863Test	863Tel	CALLHOME
Baseline	14.09%	22.04%	65.06%
Two_level_1	13.82%	21.67%	64.79%
Two-level_2	13.57%	21.03%	64.35%

5.3 Comparison of Models

In this section, we evaluate the effectiveness of the proposed modeling method described in Section 4. Tab2 shows CER comparing the proposed method with the existing data-based and rule-based methods for three types of speech database. The performance of this method provides a 0.9% absolute improvement in CER for 863Test, 1.4% for 863Tel, 1.1% for CALLHOME in comparison with TB approach. It means that, when the prior question sets in the special cases (telephone quality, spontaneous speech and so on) used for tree-based clustering are not inappropriate, the data-based cluster can remedy the drawback. However, from Tab.2, the performance of CALLHOME is less than that of 863Tel. One key reason for this is the smaller amount of CALLHOME speech data available for acoustic training. In such case, the data-based clustering is not accurate and has less effect on the last combined models.

Tab2. CER for the combined model

	863Test	863Tel	CALLHOME
Agglomerative	14.43%	23.12%	64.82%
Tree-based	13.57%	21.03%	64.35%
Combined	12.65%	19.59%	63.27%

In the experiment of CALLHOME over AGG, the diphone outperforms the triphone. It proves that with small or moderate amounts of training data, the parameter estimation of less CD model is more robust.

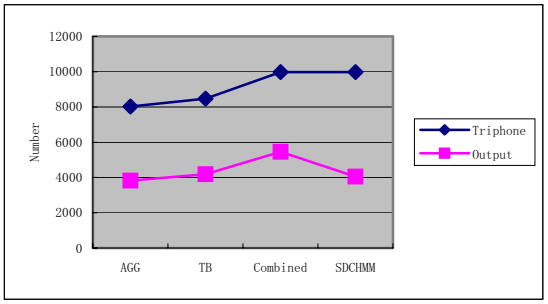


Fig3 The reduction in the number of model parameters

Of course, the acoustic parameters by our method are more than that of the previous approaches. But the experimental results show that it doesn't increase computational complexity. For instance, although 863test produced about 9.7% increment in the number of model parameters, it only adds 4.3% more computational load. This may be attributed to the pruning beam algorithm. We still use SDCHMM [12] to reduce the number of model parameters with almost no performance loss. The results of the clean reading speech are showed in Fig3.

6. Conclusions

In this paper, the presented algorithm takes advantage of both agglomerative and tree-based clustering to increase the accuracy of context-dependent large vocabulary continuous speech recognizers. Compared with the traditional single modeling methods, the combined method provides a flexible and feasible way to optimize the acoustic modeling between the prior linguistic knowledge and the available training data. The experimental results showed that the proposed method has significant effects on the case where the prior knowledge is not precise, such as telephone quality, spontaneous speech and so on. Additionally, the two-level decision tree strategy has been proposed as well as experiments, which have shown this method can be useful. Future work includes the application to multilingual acoustic modeling where the prior linguistic questions can't be obtained easily.

7. References

[1] S.J. Young, P.C. Woodland, "State Clustering in Hidden Markov Model-based Continuous Speech Recognition," Computer, Speech and Language, Vol.8, No.4, Oct 1994, pp. 369-384.

[2] P.C. Woodland et al., "Large vocabulary CSR using HTK," ICASSP'94, pp.125-128

[3] M.Y.Hwang, X.D.Huang, and F.Alleva, "Predicting unseen triphones with senones," ICASSP'93, pp.311-314.

[4] S.J. Young, J.J. Odell, P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," In Proc. Human Language Technology Workshop, March 1994, p. 307-312.

[5] X.D Huang, A. Acero, and H.W Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development," Prentice Hall PTR, 2001.

[6] J. M. Huerta and R. M. Stern, "Distortion-Class Modeling for Robust Speech Recognition under GSM RPE-LTP Coding," in Speech Communication, Vol. 34 (1-2), 2001, pp.213-225.

[7] V. Digalakis et al., "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers," IEEE Transactions on Speech and Audio Processing, Vol 4, No4, 1996, pp.281-289.

[8] J. Park and H. Ko, "Construction of Decision Tree from Data Driven Clustering," ICSLP 2002, pp.2657-2660.

[9] W. Reichl and W. Chou, "Robust Decision Tree State Tying for Continuous Speech Recognition," IEEE Trans. Speech and Audio Proc., Vol.8, No.5, 2000, pp.555-566.

[10] J.T Chien, C.H Huang, and S.J Chen, "Compact Decision Trees with Cluster Validity for Speech Recognition," ICCASP 2002, pp.2462-2465.

[11] S.Gao et al., "Weighted Graph Based Decision Tree Optimization for High Accuracy Acoustic Modeling," ICSLP2002, pp.1233-1236.

[12] E.Bocchieri, B.Make, "Subspace Distribution Clustering Hidden Markov Model," IEEE Trans. Speech and Audio Proc., Vol. 9, No 3, 2001, pp. 264-275.

[13] MY Hwang, "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition", Ph.D. Thesis, School of Computer Science, Carnegie Mellon University. Pittsburgh, PA 1993