ACOUSTIC MODELING WITH CONTEXTUAL ADDITIVE STRUCTURE FOR HMM-BASED SPEECH RECOGNITION

Yoshihiko Nankaku, Kazuhiro Nakamura, Heiga Zen and Keiichi Tokuda

Department of Computer Science and Engineering
Nagoya Institute of Technology, Nagoya 466-8555, Japan
E-mail: {nankaku, nkazu, zen, tokuda}@sp.nitech.ac.jp

ABSTRACT

This paper proposes an acoustic modeling technique based on an additive structure of context dependencies for HMM-based speech recognition. Typical context dependent models, e.g., triphone HMMs, have direct dependencies of phonetic contexts, i.e., if a phonetic context is given, the Gaussian distribution is specified immediately. This paper assumes a more complex structure, an additive structure of acoustic feature components which have different context dependencies. Since the output probability distribution is composed of additive component distributions, a number of different distributions can be efficiently represented by a combination of fewer distributions. To automatically extract additive components, this paper presents a context clustering algorithm for the additive structure model in which multiple decision trees are constructed simultaneously. Experimental results show that the proposed technique improves phoneme recognition accuracy with fewer number of distributions than the conventional triphone HMMs.

Index Terms— Hidden Markov models, Decision trees, Context clustering, Additive structure, Distribution convolution

1. INTRODUCTION

Context dependent models, e.g., triphone HMMs (hidden Markov models) are widely employed in recent speech recognition systems. It is well known that spectral features are affected by contextual factors, and extracting the context dependencies is a critical problem for acoustic modeling. One of the difficulties in the context dependent modeling is to find a good balance between the model complexity and the availability of training data. Although increasing the model complexity makes it possible to capture the accurate variations of spectral features, the reliability of parameter estimation is degraded due to decrease of the number of training data for each model. Furthermore, since it is difficult to prepare training data covering all context dependent models, there are a lot of unseen models which are not observed in training data but required in recognition phase.

To avoid this problem, the decision tree based context clustering have been proposed [1]. In the clustering, HMM states of context dependent models are grouped into "clusters," and all states belonging to the same cluster are assumed to have the same distribution. A binary tree is constructed based on the maximum likelihood criterion by applying a phonetic question to each node and splitting the cluster into two child clusters iteratively. By limiting the number of possible splitting using prior knowledge, linguistic and articulatory information can be reflected in the clustering results. Instead of the maximum likelihood criterion, the Minimum Description Length (MDL) criterion can also be adopted to determine the optimal number of clusters automatically without setting a threshold [2].

In the decision tree based context clustering, the context space is divided into the clusters by the contextual factors and the distributions of acoustic features are individually estimated for each cluster. This means that the effects of a particular factor are completely dependent of the other factors within clusters. On the other hand, the linear regression model [3] is another approach for modeling the spectral variations in which all the contextual factors independently affect acoustic features. Since the combination of contextual factors determines the spectral feature and it can efficiently represent variety of distribution. However, the dependency among contextual factors is ignored and it is difficult to determine the contextual factors which should additively affect to acoustic features.

To represent a more moderate dependencies between contextual factors and acoustic features, we propose an additive structure of acoustic feature components which have different context dependencies. The proposed approach includes intermediate structures of decision tree based context clustering and linear regression models as special cases. Since the output probability distribution is composed of the sum of the mean vectors and covariance matrices of additive components, a number of different distributions can be efficiently represented by a combination of fewer distributions. However, it is unknown which kinds of contexts have additive dependencies to acoustic features. Therefore, this paper proposes a context clustering algorithm for the additive structure model. The proposed algorithm automatically extracts additive components by constructing multiple decision trees simultaneously. Moreover, it can determine an appropriate number of additive components automatically.

1-4244-1484-9/08/$25.00 ©2008 IEEE

The rest of this paper is organized as follows. Section 2 describes additive structure models and the EM algorithm for the proposed model is derived in Section 3. In section 4, the multiple decision-tree based context clustering algorithm is described. Results of continuous speech recognition experiments are shown in Section 3. Concluding remarks and future plans are presented in the final section.

2. ADDITIVE STRUCTURE MODELS

In the context clustering, all states in the same cluster are assumed to have the same Gaussian distribution. This means that the states have direct dependencies of phonetic contexts. In this paper, we consider a more complex structure, an additive structure of acoustic feature components. An acoustic feature vector $\boldsymbol{o}_t$ at time t is generated by the sum of additive components:

$$\boldsymbol{o}_t = \boldsymbol{o}_t^{(1)} + \boldsymbol{o}_t^{(2)} + \cdots + \boldsymbol{o}_t^{(N)} \tag{1}$$

where $\boldsymbol{o}_t^{(n)}$ denote the n-th additive component. If each component is independent and generated according to a Gaussian distribution, the probabilistic density function of acoustic features is represented by the convolution of the additive components [5], so that,

$$\begin{aligned} P\left(\boldsymbol{o}_t \mid c_t, \lambda\right) = \int \Bigg[& \mathcal{N}\Big(\boldsymbol{o}_t - \sum_{n=1}^{N-1} \boldsymbol{o}_t^{(n)} \,\Big|\, \boldsymbol{\mu}_{c_t}^{(N)}, \boldsymbol{\Sigma}_{c_t}^{(N)}\Big) \\ & \times \prod_{n=1}^{N-1} \mathcal{N}\big(\boldsymbol{o}_t^{(n)} \,\big|\, \boldsymbol{\mu}_{c_t}^{(n)}, \boldsymbol{\Sigma}_{c_t}^{(n)}\big)\Bigg] d\boldsymbol{o}_t^{(1)} \boldsymbol{o}_t^{(2)} \cdots \boldsymbol{o}_t^{(N-1)} \\ = \ & \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{c_t}, \boldsymbol{\Sigma}_{c_t}) \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{c_t}^{(n)}$ and $\boldsymbol{\Sigma}_{c_t}^{(n)}$ are respectively the mean vector and covariance matrix of the n-th component given a context c_t. The output probability distribution is a Gaussian distribution whose mean vector and covariance matrix are given as

$$\boldsymbol{\mu}_{c_t} = \sum_{n=1}^{N} \boldsymbol{\mu}_{c_t}^{(n)}, \ \ \boldsymbol{\Sigma}_{c_t} = \sum_{n=1}^{N} \boldsymbol{\Sigma}_{c_t}^{(n)} \tag{3}$$

Since each additive components $\boldsymbol{o}_t^{(n)}$ have different context dependencies, we assume that each component has a different decision tree which represents tying structures of model parameters $\boldsymbol{\mu}_{c_t}^{(n)}, \boldsymbol{\Sigma}_{c_t}^{(n)}$.

Although it is unknown which kinds of contexts have additive dependencies on acoustic features in practice, we show an example of additive structure of triphone HMMs to explain effectiveness of the proposed technique. Here, we assume that the left, center and right phone are the contexts of additive components. Figure 1 shows the generative process of the triphone feature. The generative process of acoustic features is as follows: first, the component of a given monophone (center phone) context is generated from corresponding distribution obtained by descending the tree. Then, the additive components of left and right contexts are also generated independently from each distribution and then added to the monophone feature.

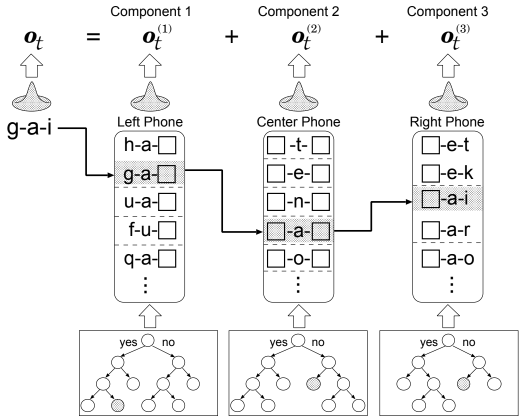

Fig. 1. An Example of A Contextual Additive Structure

The effectiveness of the proposed technique depends on whether acoustic features really have additive structures of contexts. When the acoustic features have the additive structure, a number of different distributions can be efficiently represented by a combination of fewer distributions. Furthermore, it is also effective to predict the acoustic features of unseen contexts: even though in the conventional method, unseen models are assigned to one of the clusters in the decision tree, the proposed method can construct the distribution for unseen contexts, which are different from any distribution of observed contexts.

2.1. EM algorithm for additive structure models

The Maximum Likelihood (ML) parameters of additive component distribution can be estimated by the EM algorithm. In the E-step, since the convolved output probability distribution becomes a Gaussian distribution, the standard forward-backward algorithm and the Viterbi algorithm can simply be applied as in standard HMMs. However, there is a difficulty in the M-step due to the dependencies among additive component distributions.

Using the statistics obtained by the E-step, the $\mathcal{Q}$-function with respect to the output probability distribution can be written as

$$\begin{aligned} \mathcal{L} &= \sum_{t=1}^{T} \sum_{c \in C} \gamma_t(c) \log P(\boldsymbol{o}_t \mid c_t = c, \lambda) \qquad (4) \\ &= \sum_{c \in C} \tilde{T}_c \Bigg[K \log 2\pi + \log \boldsymbol{\Sigma}_c \\ &\qquad + \mathrm{Tr}\big\{\boldsymbol{\Sigma}_c^{-1}\big(\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^{\top}\big)\big\}\Bigg] \qquad (5) \end{aligned}$$

where K is the dimensionality of feature vectors and C denotes all contexts observed in the training data. The statistics with respect to a context c is represented by $\tilde{\cdot}_c$ and each statistics is calculated as follows:

$$\tilde{T}_c = \sum_{t=1}^{T} \gamma_t(c), \quad \tilde{\boldsymbol{\mu}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c) \boldsymbol{o}_t \tag{6}$$

$$\tilde{\boldsymbol{\Sigma}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c)(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c)^{\top} \tag{7}$$

where $\gamma_t(c)$ is the state occupancy probability. For simplicity of notation, the state index is ignored. To represent the tree structure, a function $f^{(n)}(c)$ is introduced which gives the index of Gaussian distribution (leaf number of the decision tree) of n-th additive components for c. Using this function, the mean vector and covariance matrix of the convolved distribution is given by

$$\boldsymbol{\mu}_c = \sum_{n=1}^{N} \boldsymbol{\mu}^{(n)}_{f^{(n)}(c)}, \quad \boldsymbol{\Sigma}_c = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{(n)}_{f^{(n)}(c)} \tag{8}$$

Here, we focus on updating the parameters of a particular additive component. The derivative of the $\mathcal{Q}$-function with respect to the mean vector and covariance matrix can be written as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}^{(n)}_{m^{(n)}}} = \sum_{c \in \phi^{(n)}_{m^{(n)}}} \tilde{T}_c \boldsymbol{\Sigma}_c^{-1} (\tilde{\boldsymbol{\mu}}_c - \boldsymbol{\mu}_c) \tag{9}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}^{(n)}_{m^{(n)}}} = -\frac{1}{2} \sum_{c \in \phi^{(n)}_{m^{(n)}}} \tilde{T}_c \Big[& \boldsymbol{\Sigma}_c^{-1} + \boldsymbol{\Sigma}_c^{-1} \tilde{\boldsymbol{\Sigma}}_c \boldsymbol{\Sigma}_c^{-1} \\ & + \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^{\top} \boldsymbol{\Sigma}_c^{-1} \Big] \end{aligned} \tag{10}$$

where $\phi^{(n)}_{m^{(n)}}$ denotes the contexts which are included in the $m^{(n)}$-th cluster, i.e., $\phi^{(n)}_{m^{(n)}} = \{c \mid f^{(n)}(c) = m^{(n)}\}$. It can be seen from the equations that the update of $\boldsymbol{\mu}^{(n)}_{m^{(n)}}$ and $\boldsymbol{\Sigma}^{(n)}_{m^{(n)}}$ requires the parameters of the other additive components (decision trees). Hence, all parameters of all trees have dependencies on each other to compose of the output probabilities, therefore all parameters of all trees should be estimated simultaneously. Although there are many algorithms for this optimization problem, we approximate the simultaneous estimation by iterating the update of each parameter while keeping the other parameters fixed. If the other parameters are fixed, the mean vectors $\boldsymbol{\mu}^{(n)}_{m^{(n)}}$ can be easily estimated by setting the derivative to zero. The update of covariance matrix is difficult to solve analytically, accordingly the Newton method is applied for each covariance matrix. To reduce the influence of the update order, all mean vectors of all trees are iteratively updated until a convergence and then the covariance matrices are updated in the same manner. These update process for means and covariances are also iterated until a convergence.

2.2. Context Clustering for Multiple Decision Trees

To automatically extract the additive structure from training data, we propose a context clustering algorithm for multiple decision trees. As the EM algorithm for the additive structure, it is easy to construct a decision tree if the tree structures and the parameters of the other components are fixed. However, tree structures of the additive components interact each other to compose of the output probabilities, the multiple decision tress for additive components should be constructed simultaneously.

The procedure of the the proposed clustering algorithm is as follows:

Step 0. Set the number of trees N to one, and create the root node of the first tree and compute its likelihood.

Step 1. Evaluate questions at all leaf nodes of all trees and a root node of a new tree. The likelihood after splitting the node is calculated by estimating the ML parameters of all leaf nodes in all trees.

Step 2. Select the pair of a node and question which gives the maximum likelihood, and split the node into two by applying the question. The model parameters of all leaf nodes are updated by the ML parameters.

Step 3. If the change of the likelihood after splitting is below a predefined threshold, stop the procedure. Otherwise, go to Step 2.

In the procedure, there are some differences from the conventional clustering algorithm: first, in Step 2, the ML estimates of all parameters of all trees are required to evaluate questions at a candidate node. In the conventional clustering, the ML parameters of the splitted two nodes can be obtained independently of the other nodes. However, in the proposed model, the change of likelihood before and after splitting a node is calculated not only by the parameters of splitted nodes but also the parameters of the other trees. From the same reason, the likelihood of a candidate node is affected by splitting other nodes in the additive structure models. Therefore, all questions should be re-evaluated at all leaf nodes after splitting a node.

It can be seen that the proposed model which is restricted to have single tree is equivalent to the conventional decision tree based context clustering. If all trees have the only two node (applied one question), the proposed model is equivalent to a linear regression model. Thus, the proposed model can be regarded as an intermediate model between decision tree based context clustering and a linear regression model, and includes them as special cases. Furthermore, the derived algorithm can extract additive components which independently affect to acoustic features and determine an appropriate number of additive components automatically.

3. EXPERIMENTS

To evaluate the proposed technique, a continuous speech recognition experiment was conducted. We used phonetically balanced 503 sentences uttered by a single male speaker MHT from the ATR Japanese speech database b-set. The 250 sentences were used for training HMMs and the remaining 253 sentences were used for testing. The speech data was downsampled from 20kHz to 16kHz, windowed at a 10-ms frame rate using a 25-ms Blackman window, and parameterized into 19 mel-cepstral coefficients with a mel-cepstral analysis technique [6]. Static coefficients including the zero-th coefficients and their deltas and delta-deltas were used as feature parameters.

Three-state left-to-right HMMs were used to model 43 Japanese phonemes, and 118 questions about left and right phonetic contexts were prepared in decision tree construction. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix. The maximum number of decision trees were varied from one to five. The MDL criterion was applied as the stopping criterion for the proposed algorithm. For a fair comparison, the conventional triphone HMMs which have the same number of free parameters with the proposed models are constructed for each number of trees.

Figure 2 shows the relationship between the number of trees and the number of context dependent models which are constructed by the convolution of the additive component distributions. According to the figure, there is a little change in the number of additive component distributions obtained by the MDL criterion as increasing the number of trees. However, the number of composed distributions significantly increase with proportion to the number of trees. This means that the proposed algorithm captures the additive structure of training data, and context dependencies of acoustic features are efficiently represented by the multiple decision trees.

Figure 3 compares the phoneme accuracy of the proposed models and the conventional triphone HMMs. It can be seen that almost the same results were obtained for the proposed models and the conventional triphone HMMs when the number of trees is one, because these two approaches are equivalent except for the training procedures. As increasing the number of trees, the proposed method improved the recognition accuracy and achieved 5.3–10.5% phoneme relative error reduction over the conventional triphone HMMs. It is considered that the increase of the number of distributions provides a greater complexity still having sufficient robustness because it has the same number of free parameters.

4. CONCLUSIONS

In this paper, we have presented a new modeling approach considering contextual additive structure. The proposed technique can extract independent additive components composing acoustic feature vectors and their optimum context dependencies automatically. In the speaker-dependent continuous phoneme recognition experiments, the proposed model successfully reduced the phoneme relative error rates by 10.5% over traditional models. Future works include applications to speaker-independent large vocabulary continuous speech recognition and HMM-based speech synthesis.

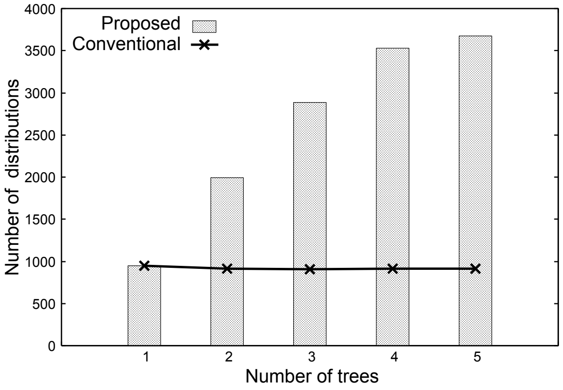

Fig. 2. Number of composed distributions.

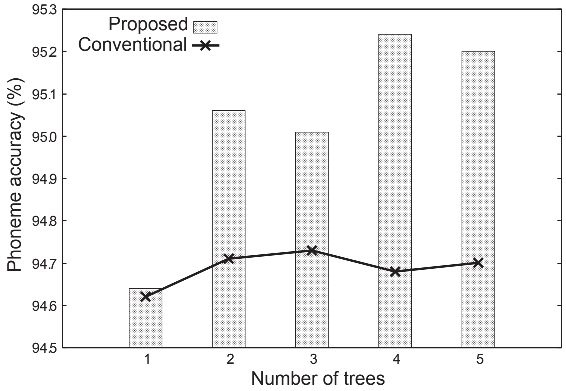

Fig. 3. Phoneme accuracy.

5. ACKNOWLEDGMENT

The authors thank Dr. Takashi Masuko (currently with Corporate Research & Development Center, Toshiba Corporation) for his comments and suggestions.

6. REFERENCES

[1] J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, University of Cambridge, 1995.

[2] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, 2000.

[3] Y. Abe and K. Nakajima "Speech Recognition Using Dynamic Transformation of Phoneme Templates Depending of Acoustic/Phonetic Environments," Proc. ICASSP'89, pp.326–329, 1992.

[5] S. Matsoukas and G. Zavaliagkos, "Convolutional Density Estimation In Hidden Markov Models For Speech Recognition," Proc. ICASSP'99, pp.113-116, 1999.

[6] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech,"Proc. ICASSP'92, vol.1,pp.137-140, 1992.