行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※

※ ※

※智慧型車輛之控制、感測與資訊處理技術研發（1/3）※

※ 子計畫二：智慧車用語音與聲響資訊處理技術研發 ※

※ ※

※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫 ■整合型計畫
計畫編號：NSC90－2213－E－009－096－
執行期間：90 年 08 月 01 日至 91 年 07 月 31 日

計畫主持人：林進燈 教授
共同主持人：

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學電機與控制工程研究所

中 華 民 國 91 年 7 月 31 日

行政院國家科學委員會專題研究計畫成果報告

子計畫二：智慧車用語音與聲響資訊處理技術研發

In-Car Speech and Audio Information Processing for Smart Car System

計畫編號：NSC90-2213-E-009-096

執行期限：90 年 08 月 01 日至 91 年 07 月 31 日

主持人：林進燈　教授 交通大學電控所

Abstract

A new speech recognition technique that is used in the environment inside the intelligent vehicles is proposed for continuous speech-independent recognition of spoken Mandarin digits. One popular tool for solving such a problem is the HMM-based one-state algorithm. However, two problems existing in this conventional method prevent it from practical use on our target problem. One is the lack of a proper selection mechanism for robust acoustic models for speaker-independent recognition. The other is the information of intersyllable co-articulatory effect in the acoustic model is contained or not. In this paper, we adopt the principle component analysis (PCA) technique to solve these two problems. At first, a generalized common-vector (GCV) approach is developed based on the eigenanalysis of covariance matrix to extract an invariant feature over different speakers as well as the acoustical environment effects and the phase or temporal difference. The GCV scheme is then integrated into the conventional HMM to form the new GCV-based HMM, called GCVHMM, which is good at speaker- independent recognition. For the second problem, context-dependent model is done in order to account for the co-articulatory effects of neighboring phones. It is important because the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances. However, there must be numerous context-dependent models generated because of modeling the variations of sounds and pronunciations. Furthermore, if the parameters in those models are all distinct, the total number of model parameters would be very huge. To solve the problems above, the decision tree state tying technique is used to reduce the number of parameter, hence reduce the computation complexity.

1. Introduction

Automatic speech recognition (ASR) is useful as a form of input. It is especially useful when someone's hands or eyes are busy. It also allows people with handicaps such as blindness or palsy to use computers. Especially for the environment inside the intelligent vehicles, automatic speech recognition is a helpful and friendly man-machine interface for the drivers of the intelligent vehicles. Because of the potential applications mentioned above, we attempt to develop a speaker-independent automatic speech recognition system for Mandarin digits.

In recent years, most automatic speech recognition technologies were based on the so-called Hidden Markov Models (HMM) and used the connected word pattern matching method to achieve continuous speech recognition. There exists many methods to solve the connected word pattern-matching problem. One well-known method is called one-state algorithm. There are two problems in continuous speech recognition based on the one-stage algorithm, one is how to build a reference model to characterize the acoustic feature of speech signal, the other is the information of intersyllable co-articulatory effect in the acoustic model is contained or not.

Due to the first problem mentioned above, the one-state algorithm is sensitive to the reference patterns, and thus the choice of reference patterns is important. One well-known and widely used statistical method of characterizing the spectral properties of the frames of a speech pattern is the HMM approach. The better the HMM models the acoustic signals, the better performance the one-state algorithm can achieve. One of the most important issues of speaker-independent (SI) speech recognition system is the estimation of robust speech model over different speakers. The statistical speech models for each phone unit of the recognition system should be estimated to cover the spectral variations in speech signal caused by intra-speaker differences. In this thesis, we propose a new framework of HMM called *generalized common-vector-based HMM* (GCVHMM) as a reference for speaker-independent automatic speech recognition.

For the second problem, context-dependent model is done in order to account for the co-articulatory effects of neighboring phones. It is important because the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances. However, there must be numerous context-dependent models generated because of modeling the variations of sounds and pronunciations. Furthermore, if the parameters in those models are all distinct, the total number of model parameters would be very huge. To solve the problems above, the decision tree state tying technique is used to reduce the number of parameter, hence reduce the computation complexity.

2. HMM

The HMM, which uses probabilistic functions of Markov chains to model random processes, is a model of stochastic process. The effectiveness of this model class lies in its ability to deal with non-stationarity that often appears in the observed data sequences. HMMs usually turn out to be a good model for non-stationary process, such as the sequence of the speech observation vectors.

2.1 Elements of an HMM

An HMM can be characterized by the set of parameters *A*, δ , and *B*. We list all these parameters as following to represent an HMM.

1. N, the number of states in the model. The states are hidden in HMM, which have some physical significance attached.
2. M, the number of mixtures per state for the output probability distribution of a continuous probability density function (pdf) of Gaussian mixtures.
3. The state transition probability distribution *A* = [$a_{i,j}$], where
$$a_{i,j} = P(\bar{\ }_t = j | \bar{\ }_{t-1} = i),\ 1 \le i, j \le N.$$
4. The observation probability distribution in state *j*, B = {$b_j(o_t)$}, where
$$b_j(o_t) = P(o_t | \bar{\ }_t = j),\ 1 \le j \le N.$$
5. The initial state distribution δ ={ δ_i} where
$$¢_i = P(\bar{\ }_0 = i),\ 1 \le i \le N.$$

It can be seen from the above discussion that a complete specification of a HMM requires specification of two model parameters (*N* and *M*), and the specification of the three probability measures *A*, *B*, and finally the initial state distribution δ. For convenience, we indicate the complete parameter set of the model by:

Ù=(A, ä, B)

2.2 Three Basic Issues for HMMs

In order to solve that the HMM can be used in real-world applications, there are three basic problems as follows:

Issue 1: Given the observation sequence *O*, and a model Ù, how do we efficiently compute *P(O|*Ù*)*, the probability of the observation sequence given the model?

Issue 2: Given the observation sequence *O*, and the model Ù, how do we choose a corresponding state sequence $\hat{\Theta}$, which is optimal in some meaningful sense?

Issue 3: How do we adjust the model parameters Ù = {*A*, ä, *B*} to maximize *P(O|Ù)*?

3. GCVHMM

The statistical speech models for each phone unit of the speaker-independent (SI) recognition system should be estimated to cover the spectral variations in speech signals caused by intra-speaker differences. Gülmezoðlu, et al. proposed a common vector approach (CVA) for SI isolated word recognition. In CVA, a common vector that represents common properties of one specific spoken word is obtained by estimating a common subspace. However, CVA needs the impractical assumption that the training data form a set of linearly independent vectors.

In this chapter, we generalize the CVA to relax its constrain and propose a new extension of HMM called *generalized common-vector-based HMM* (GCVHMM). There are two phases in the GCVHMM, extraction of robust features and estimation of HMM. In the first phase, a generalized CVA is developed based on the eigenanalysis of covariance matrix to extract an invariant feature, called generalized common vector (GCV). To relax the linearly independent assumption in the original CVA, we divide the eigenvalues of covariance matrix into two sets such that all the eigenvalues of the first set are greater than those of the second set. The common vector is obtained by projecting feature vectors on the subspace spanned by the eigenvectors whose corresponding eigenvalues are in the second set. In the second phase, the GCVs are used for the estimation of continuous observation density in HMM and form the so-called GCVHMM. In GCVHMM, in addition to the original elements of a traditional HMM, a new element, *GCV transformation matrix*, is added to extract GCV from speech feature vectors. Finally, a re-estimation algorithm based on Baum-Welch method to estimate all the parameters of GCVHMM is derived.

3.1 Structure of GCVHMM

In this thesis, a N state, left-to-right continuous observation density HMM, denoted as Ω, is considered. The initial probability for state i is denoted by $\delta_i = P(\theta_0 = i)$, 1 ≤ i ≤ N, and the transition probability from state i to state j by $a_{i,j} = P(\text{¯}_t=j \mid \text{¯}_{t-1}=i)$ for 1 ≤ i, j ≤ N. Denote $\text{¢} = \{\text{¢}_i\}_{i=1}^{N}$, and $A = \{a_{i,j}\}_{i,j=1}^{N}$. For the calculation of the observation density in state i, denoted as $b_i(o_t)$, for observation o_t, the generalized common vector of o_t given the matrix transformation of generalized common vector is first extracted. Then $b_i(o_t) = P(o_t \mid \text{¯}_t=i)$, $1 \le i \le N$ assumed to be a mixture of Gaussians is then given as

$$b_i(o_t) = \sum_{k=1}^{M} c_{i,k} b_{i,k}(o_t),\ 1 \le i \le N$$

where M is the mixture number, $c_{i,k}$ is the probability of mixture k in state i, and $b_{i,k}(o)$ is the gaussian distribution given by

$$b_{i,k}(o_t) = \frac{1}{\sqrt{(2\text{®})^{D_s} \left|\Lambda_{i,k}\right|}} e^{-\frac{1}{2}(y_{t,i,k} - \text{¦}_{i,k})\Lambda_{i,k}^{-1}(y_{t,i,k} - \text{¦}_{i,k})}$$

where D_s=D-D_g is the dimension of the extracted GCV $y_{t,i,k}$ from o_t, $y_{t,i,k}$ is the GCV of o_t for mixture k in state i, and $\Lambda_{i,k}$ and $\text{¦}_{i,k}$ are the covariance matrix and mean vector corresponding to mixture k in state i, respectively. $\Lambda_{i,k}$ is assumed to be diagonal, i.e.,

$$\Lambda_{i,k} = \begin{bmatrix} \text{±}_{i,k,1} & 0 & \cdots & 0 \\ 0 & \text{±}_{i,k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{±}_{i,k,D_s} \end{bmatrix}$$

so that $\left|\Lambda_{i,k}^{-1}\right| = \prod_{l=1}^{D_s} \text{±}_{i,k,l}^{-1}$. The GCV $y_{t,i,k}$ from o_t for mixture k in state i is defined as

$$y_{t,i,k} = V_{i,k,o_t}$$

where

$$V_{i,k} = \left[v_{i,k,1}, v_{i,k,2}, \ldots, v_{i,k,D_s} \right]^T$$

is matrix transformation of generalized common vector for mixture k in state i. For convenience in the following derivation, we also define

$$\text{¦}_{i,k} = V_{i,k}\text{«}_{i,k}$$

then we can write

$$z_{t,i,k} = y_{t,i,k} - \text{¦}_{i,k} = V_{i,k}\left(o_t - \text{«}_{i,k}\right)$$

Denote $B = \{b_i\}_{i=1}^{N}$ and Ω = {δ, A, B}.

3.2 Reestimation algorithm for the parameters of GCVHMM

For an observation sequence $O = (o_1, o_2, \ldots, o_T)$ unobserved state sequence $\Theta = (\text{¯}_0, \text{¯}_1, \text{¯}_2, \ldots, \text{¯}_T)$, and unobserved mixture component sequence $K = (k_1, k_2, \ldots, k_T)$, the joint probability density of $P(O, \Theta, K \mid \Omega)$ is defined as

$$P(O, \Theta, K \mid \Omega) = \text{¢}_{\text{¯}_0} \prod_{t=1}^{T} a_{\text{¯}_{t-1},\text{¯}_t} c_{\text{¯}_t,k_t} b_{\text{¯}_t,k_t}(o_t)$$

where T is the number of observation in O. It follows that the likelihood of O given Ω has the form

$$P(O \mid \Omega) = \sum_{\Theta} \sum_{K} P(O, \Theta, K \mid \Omega)$$

where the summations are over all possible state sequences and mixture component sequences.

Given an observation sequence O, the objective is to maximize $P(O|\Omega)$ over all parameters in Ω. It is, however, difficult to solve this problem by directly maximizing $P(O \mid \Omega)$ over Ω. In this following, we shall use the EM algorithm to estimate the parameters of HMM. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step (E step), we compute the auxiliary function for the equation

$$Q(\Omega,\Omega') = \sum_{all\,\Theta} \sum_{all\,K} P(O,\Theta,K \mid \Omega) \log P(O,\Theta,K \mid \Omega')$$

In the second step, called the maximization step (M step), we find the value of Ω' that maximizes $Q(\Omega,\Omega')$, i.e.,

$$\bar{\Omega} = \arg\max_{\Omega'} Q(\Omega, \Omega')$$

It has been shown that if $Q(\Omega,\Omega') \ge Q(\Omega,\Omega)$, then $P(O|\Omega') \ge P(O|\Omega)$. Therefore, iteratively applying the E and M steps of equations guarantees monotonic increase in the likelihood. The iterations are continued until the increase in the likelihood is less than some predetermined threshold.

From the following decomposition:

$$\log P(O, \Theta, K \mid \Omega') =$$

$$\log \text{¢}'_{\text{¯}_0} + \sum_{t=1}^{T} \log a'_{\text{¯}_{t-1},\text{¯}_t} + \sum_{t=1}^{T} \log c'_{\text{¯}_t,k_t} + \sum_{t=1}^{T} \log b'_{\text{¯}_t,k_t}(o_t)$$

it is straightforward to shown that $Q(\Omega,\Omega')$ can be decomposed into a sum of four auxiliary functions:

$$Q(\Omega, \Omega') = Q_{\text{¢}}(\Omega, \text{¢}') + \sum_{i=1}^{N} Q_{a_i}[\Omega, \{a'_{i,j}\}_{j=1}^{N}]$$

$$+ \sum_{j=1}^{N} Q_{c_j}[\Omega, \{c'_{j,k}\}_{k=1}^{M}] + \sum_{j=1}^{N} \sum_{k=1}^{M} Q_b(\Omega, b'_{j,k})$$

where

$$Q_{\text{¢}}(\Omega, \text{¢}') = \sum_{i=1}^{N} \sum_{K} P(O, \text{¯}_0 = i, K \mid \Omega) \log \text{¢}'_i$$

$$Q_{a_i}[\Omega, \{a'_{i,j}\}_{j=1}^{N}] = \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{K} P(O, \text{¯}_{t-1} = i, \text{¯}_t = j, K \mid \Omega) \log a'_{i,j}$$

$$Q_{c_j}[\Omega, \{c'_{j,k}\}_{k=1}^{M}] = \sum_{k=1}^{M} \sum_{t=1}^{T} P(O, \text{¯}_t = j, k_t = k \mid \Omega) \log c'_{j,k}$$

$$Q_b(\Omega, b'_{j,k}) = \sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega) \log b'_{j,k}(o_t)$$

This implies that the four sets of parameters can be independently maximized. The maximization results of first three auxiliary functions are

$$\text{¢}_i = \frac{P(O, \bar{\ }_0 = i \mid \Omega)}{P(O \mid \Omega)}$$

$$a_{i,j} = \frac{\sum_{t=1}^{T} P(O, \bar{\ }_{t-1} = i, \bar{\ }_t = j \mid \Omega)}{\sum_{t=1}^{T} P(O, \bar{\ }_{t-1} = i \mid \Omega)}$$

$$c_{j,k} = \frac{\sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega)}{\sum_{t=1}^{T} P(O, \bar{\ }_t = j \mid \Omega)}$$

Substituting the following decomposition

$$\log b'_{-_t,k_t}(o_t) = -\frac{D_s}{2}(2\text{®}) + \frac{1}{2}\log(\left|\Lambda'^{-1}_{-_t,k_t}\right|) - \frac{1}{2} z'^{T}_{t,-_t,k_t} \Lambda'^{-1}_{-_t,k_t} z'_{t,-_t,k_t}$$

where

$$z'_{t,-_t,k_t} = y'_{t,-_t,k_t} - \iota'_{-_t,k_t} = V'_{-_t,k_t}(o_t - \text{«}'_{-_t,k_t})$$

for $\log b_{-_t,k_t}(o_t)$ and differentiating it with respect to $\text{«}'_{j,k}$ and $\pm'^{-1}_{j,k,l}$,, we obtain

$$\overline{\text{«}}_{j,k} = \frac{\sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega) \cdot o_t}{\sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega)}$$

$$\overline{\pm}_{j,k,l} = \frac{\sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega) \cdot z'^{2}_{t,j,k,l}}{\sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega)}$$

where $z'_{t,j,k,l}$ is the lth element of $z'_{t,j,k}$.

To obtain the solution for $v'_{j,k,l}$, which is the lth element of $v'_{j,k}$:

$$\langle v'_{j,k,l}, v'_{j,k,l} \rangle = 1,\ 1 \le l \le D_s$$

the constrains

$$\frac{1}{2}\sum_{l=1}^{D_s} {}^{\circ}_{i,k,l}(\langle v'_{j,k,l}, v'_{j,k,l} \rangle - 1)$$

are added to $Q_b(\Omega, b'_{j,k})$. Then,

$$\frac{\partial Q_b(\Omega, b'_{j,k})}{\partial v'_{j,k,l}} = \vec{0}$$

we obtain

$$R_{j,k}\overline{v}_{j,k,l} = \text{£}_{j,k,l}\overline{v}_{j,k,l}$$

where

$$R_{j,k} = \sum_{t=1}^{T} P(O, \bar{\ }_t = j, k_t = k \mid \Omega)(o_t - \text{«}'_{j,k})(o_t - \text{«}'_{j,k})^T$$

It can be said that $R_{j,k}$ characterizes the variations for mixture k in state j so that it plays the same role as Φ_X in previous section. Thus, the eigenvectors of $R_{j,k}$ corresponding to the eigenvalues of smallest D_s are selected to constitute the GCV matrix transformation for mixture k in state j.

4. A Hybrid Decision Tree

To overcome this limitation, we have introduced the integrated generalized common vector approach into the conventional HMM in chapter 3, which is better at speaker-independent recognition because of its ability to extract common invariant features over different speakers. Besides modeling acoustic parameters, most of the variations are due to consistent contextual effects in practice. Therefore, we can focus our research on context-based information. Since the co-articulatory effect for continuous speech is significantly stronger than that for isolated utterances, it is important to study the modeling of context-dependent “subword” units. Here, “subword” means “Mandarin digits”, which indicate syllable equally.

The most important reason why we use the method of decision tree state tying is that the total number parameters in all models is prohibitively large. The computation complexity to train all these parameters would be intolerable. To reduce the total number of model parameters, one approach is to reduce the number of parameters in each model. The way of using continuous HMMs with tied parameters, parameter tying, reduces the parameter count while maintaining the model accuracy, and is popularly used in most ASR systems.

After incorporating common vector features mentioned in Chapter 3 with the structure of decision tree state tying, the decision tree algorithm should be modified as follows:

1. **Locate a (small) set of left context digit syllable questions manually.**
2. **For each center Mandarin digit syllable p:**
   - ■ Estimate all left context digit syllable GCVHMMs.
   - ■ For each Markov state k in the model topology, classify all the k-th output distribution in all left context digit syllables using a binary tree.
     - **a)Put all the training data in k-th state of all left context digit syllables into the root node.**
     - **b)Classify all the training data by each question in the question set. Using the clustered data to generate Gaussian distribution of common features by the method of GCVHMM introduced in Chapter 3. Then compute the likelihood of the parent node from Equation** (4.15)**:**

$$L = -\frac{1}{2}\left[D_S\left(1 + \ln\left(2\text{®}\right)\right) + \ln\left(\left|\Lambda_{i,k}\right|\right)\right] \cdot \sum_i \sum_t \Upsilon_i(t)$$

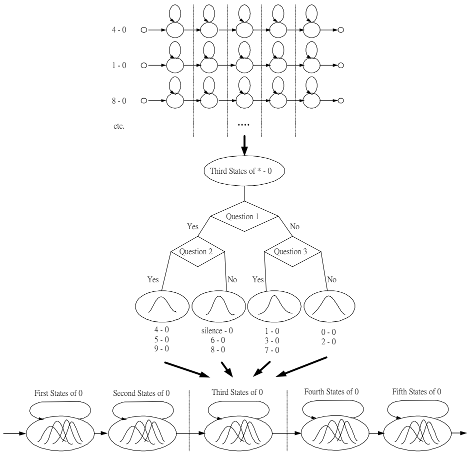
where D_S and $\Lambda_{i,k}$ represent the dimension and the covariance matrix

of common vector after the method of GCVHMM.

c) **Split the node by each question in the question set. By splitting, some training data that come from the left context digit syllables which answer yes to the question go to the yes-child node; those which answer no to the no-child node. Then calculate the every likelihood of two child nodes. Finally, compute the likelihood increment by each question in the question set.**

d) **Find the best question in the question set by computing the most likelihood increment for each of the newly created children.**

e) **Go to step b) unless some stop-growing criteria is met.**



GCVHMM of digit syllable "0"

5. Mandarin Digits Recognition Experiments

The speech data used in our experiments are the set of continuous Mandarin digits. We use a speech database from 20 persons including 10 males and 10 females. Each one speaks 10 times of each Mandarin digit. The recording sampling rate is 8kHz and stored as 16-bit integer.

5.1 Balanced Corpora

Digit Model	Decision Tree State Tying Based on GCVHMMs	GCVHMMs	HMM
0	91.667	83.333	58.333
1	61.111	27.778	0.000
2	61.538	53.846	76.923
3	100.000	89.474	100.000
4	83.333	50.000	50.000
5	100.000	94.118	100.000
6	60.000	30.000	0.000
7	100.000	46.154	7.692
8	86.667	40.000	60.000
9	100.000	69.231	23.077
Average	84.432	58.393	47.603

5.2 Unbalanced Corpora

Digit Model	Decision Tree State Tying Based on GCVHMMs	GCVHMMs	HMM
0	75.610	65.854	53.659
1	66.522	76.087	28.261
2	72.727	54.545	54.545
3	84.091	56.818	90.909
4	93.182	100.000	88.636
5	81.818	97.727	95.455
6	75.556	51.111	80.000
7	95.455	100.000	95.455
8	93.182	77.273	90.909
9	90.909	95.455	95.455
Average	82.9052	77.487	76.419

5.3 Balanced Tree

Digit Model	Balanced Decision Tree State Tying Based on GCVHMMs	Unbalanced Decision Tree State Tying Based on GCVHMMs
0	75.610	51.220
1	66.522	34.783
2	72.727	63.636
3	84.091	68.182
4	93.182	86.364
5	81.818	77.273
6	75.556	64.444
7	95.455	88.636
8	93.182	77.273
9	90.909	90.909
Average	82.9052	70.272

6. Conclusion

To consider the contextual effects of continuous speech that play an important role in Mandarin, we combine a method of the Decision Tree State Tying with GCVHMM. The balanced corpora mean that the count of females and males in the database are equivalent entirely. It shows 26.039% improvement when we replace GCVHMM with Decision Tree State Tying based on GCVHMM. Nevertheless, if the database is unbalanced, the performance comparison shows 5.4% improvement by employing the Decision Tree State Tying based on GCVHMM. To overcome leaving the major part of models behind in the unbalanced tree, we modify the tree as the balanced tree. We can find that the results show 12.6332% improvement by employing the balanced decision tree state tying based on GCVHMM. This technique is utilized as a helpful and friendly man-machine interface in the environment inside the intelligent vehicles.