A Novel Information Integration Algorithm for Speech Recognition System: Basing on Adaptive Clustering and Supervised State of Acoustic Feature

Chen Xu¹, * and Xi Xiao¹

¹Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

*E-mail address: c-xu16@mails.tsinghua.edu.cn

Abstract. When utilizing the most likely state sequence (MLSS) criterion in Gauss mixture model-hidden Markov model (GMM-HMM) to acquire the best state series of observations, only the maximum likelihood state of speech frame is considered. Therefore, the influence of other states is neglected, which leads to the losing of some important information, and further reduces the recognition rate of the system. In this paper, we propose two new features, which are called state likelihood cluster feature (SLCF) and supervised state feature (SSF), to both reflect acoustic features and fuse state information. Combining SLCF and SSF with Mel frequency cepstrum coefficient (MFCC), Mel frequency cepstrum & state likelihood cluster feature (MSLCF) and Mel frequency cepstrum & supervised state feature (MSSF) are formed, respectively. By the proposed MSLCF and MSSF in Chinese speech recognition experiment, the relative error rate of the isolated word recognition system declines 6.10% and 9.66%, respectively, and the relative error rate of the continuous speech recognition system declines 2.53% and 11.05%, respectively.

1. Introduction

Automatic speech recognition (ASR) is the key technology for realizing the active communications between people and machine [1]. From dynamic time warping (DTW) [2] to hidden Markov model (HMM) [3], and then to deep neural network (DNN) [4], different technologies are proposed, but they still are facing the problem of improving the accuracy of recognition. Chinese speech recognition system usually is based on the hidden Markov model (HMM) as an important part for model training and recognition, but in recent years, the accuracy improvement of HMM encounters the bottleneck. The improvements of the model mainly focus on reducing the inter-frame correlation or enhancing the model precision, which include providing linear prediction model [5], utilizing discriminative training to replace the method of most likelihood estimate (MLE) [6], and introducing neural network et.al. However, these improvements cannot fix the high redundancy of speech information, which influences the recognition performance.

At the moment, the researches on the features themselves with maximum likelihood (ML) criterion are lacking. One of the existing technologies is utilizing linear discriminant analysis to proceed linear transformation of eigenvector, which can reduce the dimension of eigenvector and remain the most discerning eigen component. Another technology is utilizing kernel function linear discriminant analysis to realize kernel-based ML linear transformation [8]. However, because these technologies use linear transformation matrices to decorrelate the parametric eigenvector, derived transformation matrices still rely on the original eigenspace, which cannot be utilized by models with different eigenvector spaces.
In addition, the computational complexity of these technologies is much higher.

In this paper, in section (3.1), coming from acoustic feature itself, we adaptively cluster all acoustic features of the speech training set, to get the probability model of the clustering model’s generative features. Then the scoring information of the training set acoustic features in state space is derived. Based on this, in section (3.2), we propose supervised state features to comprehensively reflect the distribution of acoustic features in state space.

Taking Gauss mixture model-HMM (GMM-HMM) system as an example, in the original model, GMM is used for the observed acoustic features distribution of the states. Then in HMM, based on the MLSS algorithm, the forced alignment and segmentation are implemented to get the best state sequence [9]. However, when utilizing the likelihood of this model, only the ML state of the current frame is considered. Therefore, the effect of other states to the current frame is neglected, and the information is lost, which reduces the recognition rate. However, the supervised state combines acoustic observation feature with state information, and this dual information can more accurately reflect the situation of acoustic feature in state space, which can improve the recognition accuracy of speech recognition system.

2. System overview

In this system, the Mel Frequency Cepstrum Coefficient (MFCC) is utilized as acoustic observation feature, the GMM-HMM is utilized as acoustic model. The expectation maximization (EM) algorithm is used to estimate the parameters of GMM, and the Viterbi decoding algorithm based on dynamic programming is utilized to search the optimal state sequence of the speech. In HMM, each of Markov chain states is modeling a basic phonetics unit. Among Chinese model, each tonal Chinese syllable in isolated speech is modeled by consonant-vowel (C-V) structure, while in continuous speech it’s model by V-C-V structure, i.e. triphone model[10], in order to reflect the relationship between syllables. By means of triphone modeling, the current syllable’s initial states is depend on the final states of the previous syllable, as in Figure 1, based on non-coarticulation states $I_0$, $I_1$, $F_0$, $F_1$, $F_2$, $F_3$, we add coarticulation states $I_{0c}$, $F_{2c}$, $F_{3c}$ to accurately represent the physical properties of continuous speech.

![Figure 1. Chinese three-tone word model](image)

There are 1254 Chinese tonal syllables, 856 non-coarticulation states and 49200 coarticulation states in this system. In order to control the scale of the model, we artificially define the dictionary to separately cluster consonants and vowels. More specifically, by the pronunciation rules of Chinese tonal syllables, 100 consonant states are further clustered into 27 classes, and 164 vowel states are clustered into 29 classes, which reduces the number of collaborative states to 2349. With the tail noise state model, there are total 3206 states in the tritone model.

In Chinese GMM-HMM recognition model, the MLSS algorithm is utilized in recognition network to find an optimal path with smallest accumulated distance. Then the states of the input series can be determined.

$$\max_{S^T_{1}} P(O^T_{1},S^T_{1} | \theta) = \max_{(t_1,t_2,..,t_N)} \Sigma_{t=1}^{N} \Sigma_{t=t-1}^{t} \ln b_t(O_t)$$ (1)

Given the training sample $O$, the model parameter $\theta$ is estimated to maximize equation (1), and this maximized state series $S^T_{1}$ is called the maximum likelihood series of this model. This state series $S^T_{1}$
also determines the segmentations of the model state as in equation (2). We also get state $S_t$ that each frame feature belongs to

\[ S_t \in \text{State}_i \ (t_i \leq t < t_{i+1}, \ i = 0, 1, ..., N - 1) \]  

(2)

Where $S_t$ represents state of feature $O_t$ at the time $t$, $t_i$ is the segmentation time that entering state $i$, and $\text{State}_i$ is the $i$-th state in the used set of states.

Through MLSS algorithm, the optimal path selection is based on the so-called optimal state with the maximum likelihood score. However, it ignores the influence of other sub-optimal state likelihood scores on the current frame, and therefore, abandons a lot of information. In this paper, we improve current Chinese GMM-HMM model from the single-frame’s perspective to enhance the utilization likelihood score information for all states. In addition, we fuse the state information with MFCC to improve speech features, and thus lower the recognition error rate.

3. Algorithm

3.1. State likelihood score model and likelihood distribution training

For each speech frame, we define the HMM state likelihood score vector which is obtained from MFCC under GMM-HMM model, as the state likelihood feature (SLF) in equation (3).

\[ \text{SLF} = [\text{SLF}_1, \text{SLF}_2, ..., \text{SLF}_{N-1}, \text{SLF}_N], \text{SLF}_i = b(o|\lambda, s_i) = g(o|\mu_{s_i}, \Sigma_{s_i}) \]  

(3)

And $g$ is the GMM, which can be extended to equation (4).

\[ \text{SLF}_i = b(o|\lambda, s_i) = \sum_{j=1}^{M} \alpha_j \cdot g(o|\mu_{s_i}^j, \Sigma_{s_i}^j). \sum_{j=1}^{M} \alpha_j = 1 \]  

(4)

Where $M$ denotes the mixture number of GMM, $\mu_{s_i}^j$ and $\Sigma_{s_i}^j$ denotes the mean vector and covariance matrix of the $j$-th Gaussian component of the GMM with state $s_i$, and $\alpha_j$ denotes the weight of the $j$-th Gaussian component.

Under a large number of samples, a GMM model is employed to train the model of generating the SLF that corresponding to each of HMM states. We call these GMM models as SLF codebooks. With SLF codebooks, likelihood scores of SLF can be obtained. And then by algorithm of MLSS, likelihood scores of SLF are used for searching the optimal state sequence. We define the above process as state likelihood score model. Based on the state likelihood score model, the implementation process of GMM-HMM speech recognition model is shown in Figure 2.

![Figure 2](image_url)  

**Figure 2.** The flow chart of GMM-HMM model based on state likelihood score model.

Taking the isolated word model as an example, the HMM model of isolated word system has 856 states, and then we use each 45-dimension MFCC acoustic feature to get the 856-dimension state
likelihood corresponding to 856 state codebooks. After that, the SLF codebook is trained utilizing diagonal matrix codebook because of the computational complexity and feasibility. Finally, we use SLF to test the description ability of state likelihood feature codebook.

The experiment finds that due to the large number of states, under the condition of limited labeling, the feature SLF was not suitable to merge into the original model. Because the model size and data volume are not matched, the recognition rate of the system cannot be improved. Therefore, clustering is utilized to reduce the dimension of SLF, and state likelihood clustering feature (SLCF) is obtained.

In our experiment, a GMM with covariance matrix is utilized to train SLCF model, and the new SLCF model is called SLCF codebook. Based on GMM-HMM system, we do the SLCF recognition experiment with MFCC. According to experiment, SLCF adds more comprehensive information of acoustic feature in the state space, and to some extent, it reflects the influence of sub-optimal state to current frame and efficiently improve the recognition rate of the model.

3.2. Supervised state and supervised state feature distribution
In HMM, acoustic feature observations reveal the likelihood of hidden states via ML criterion and Viterbi algorithm. In Chinese triphone model, states are evolved from C-V structure. These states are artificially classified for model training. However, there are two major problems in this process: one is whether the artificially defined state is reasonable, and the other one is whether the information in defined state space is fully utilized. Therefore, we propose the concept of supervised state. In this case, supervised state is the representative class of acoustic features in state space. In particular, \( SS = \{ SS_1, SS_2, ..., SS_i, ..., SS_N \} \) is the supervised state set of state space with \( N \) states in total, and acoustic features can be generated by these supervised states. Each supervised state \( SS_i \) has its own probability model \( G_i(O) \) for generating acoustic feature \( O \). These supervised state models can be selected manually based on expert knowledge or generated automatically based on machine learning. In this paper, we generate the supervised state models via machine learning, and we hope that it can provide richer information comparing with states in HMM. The output of vector \( O \) to each supervised state model \( G_i(O) \) is \( SSF_i \), then the score of vector \( O \) for all the supervised states constitutes the characteristic supervised state feature (SSF).

\[
SSF = [SSF_1, SSF_2 ... SSF_i ... SSF_{N-1}, SSF_N], SSF_i = P(O|SS_i) = G_i(O)
\]

The method of training adaptive supervising state distribution can be divided into two steps. Firstly, starting from Chinese pronunciation rules, the initial supervised state class center is defined artificially, and the Gaussian distribution of the initial supervised state is formulated. Secondly, the clustering center is updated automatically by computer via ML clustering.

\[
d(x, SS_i) = p(x|SS_i) = g(x|\mu_{SS_i}, \Sigma_{SS_i}) = \frac{1}{(2\pi)^{p/2}|\Sigma_{SS_i}|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_{SS_i})^T \Sigma_{SS_i}^{-1} (x - \mu_{SS_i}) \right\}
\]

In equation (6), where \( x \) represents the observed acoustic input feature, \( SS_i \) represents the \( i \)-th supervised state, and \( d(x, SS_i) \) represents the distance between \( x \) and \( SS_i \).

The specific steps of adaptive supervising state features are as follows:

**Step 1** Initialize the supervised state clustering center, and get the initial distribution of supervised state features.

**Step 2** Input phonetic feature \( x_j \), and traverse supervised state to get the label \( label_{x_j} \) of the feature \( x_j \), where \( label_{x_j} \in \{ 1, 2, ..., N \} \) and \( j \in \{ 1, 2, ..., M \} \). \( N \) denotes the total number of supervised states, \( M \) is the total number of training set phonetic features. The label \( label_{x_j} \) of the feature \( x_j \) can be denoted as follows:

\[
label_{x_j} = \arg \max_i d(x_j, SS_i)
\]

**Step 3** Count all phonetic features \( X = \{ x_1, x_2, ..., x_M \} \), and the label \( Label_X = \)
\{\text{label}_{x_1}, \text{label}_{x_2}, ..., \text{label}_{x_M}\}$, which corresponds to the cyclical features. Update the feature distribution of supervised state.

**Step 4** Repeat Step 2 and Step 3 until the distribution is stable, and get the feature distribution of supervised state.

### 3.3. The Chinese speech recognition model which combines the state likelihood score model or the supervised state model

The Chinese speech recognition model fusing two kinds of information, it utilizes MFCC to characterize the acoustic information of input speech features, and utilizes SSF or SLCF to reflect state information. Through feature fusion, we get Mel Frequency Cepstrum & State Likelihood Cluster Feature (MSLCF) and Mel Frequency Cepstrum & Supervised State Feature (MSSF). Thus, the novel feature codebook can be generated, trained and utilized for recognition.

![Figure 3. The GMM-HMM speech recognition model which fuses MSLCF or MSSF](image)

As in Figure 3, there are four steps to generate the MSLCF codebook. Step 1: Extracting MFCC acoustic feature of speech from database. Step 2: Get the optimal state series utilizing the ML estimation of MFCC, and then get the HMM state label of each frame. Step 3: Calculate the state likelihood value SLF of MFCC, get the SLCF by clustering, and obtain MSLCF through combining MFCC with SLCF. Step 4: MSLCF and corresponding labels of speech frames are counted to obtain the MSLCF distribution, i.e., the MSLCF codebook.

The SSF model can be trained in advance from the speech database. Based on the obtained supervised state feature distribution, the MSSF codebook is produced. The procedures are similar to generating the MSLCF codebook, but the Step 3 should be changed. In Step 3 of producing MSSF codebook, the observed feature vector MFCC is an input to the supervise state models get SSF, and then they are combined into MSSF.

When we train the MSLCF codebook or MSSF codebook, the Step 2 in generating codebook experiment is changed into utilizing the ML estimation of MSLCF or MSSF to obtain the optimal state series. Other steps are the same as those in generating experiment.

In these procedures, MFCC is a 45-dimension vector, which includes a 14-dimension Mel frequency cepstrum coefficient, a 1-dimension normalized energy coefficient, and their first order difference and second order difference coefficient. SLCF is a 57-dimension vector of score likelihood of clustered state, and SSF is a 57-dimension supervised state feature. MSLCF and MSSF are the 102-dimension features combined MFCC with SLCF and SSF, respectively.
4. Experiments

4.1. Isolated word recognition experiment with SLCF, MSLCF

The isolated-word experiment utilizes a 50 male isolated word speech database. 49 speech samples are selected as training set, and another one is the test set. The data sample rate is 16kHz, the frame length is 20ms, and the frame shift is 10ms. In this experiment, the acoustic model utilizes Chinese isolated-word monosyllable model with 1254 tonal syllables, which totally includes 857 isolated HMM state with mute state.

It can be shown from Figure 4 that combining with SLCF can assist acoustic feature recognition system. In GMM-HMM system, comparing with using single acoustic feature MFCC, utilizing fused feature MSLCF can decrease the absolute error rate by 0.6381%, and decrease the relative error rate by 6.10%.

Figure 4. The recognition effect of 50 isolated-word male acoustic samples with isolated-word model features SLCF, MSLCF and MFCC

4.2. Feature training and recognition experiment with supervised state SSF and MSSF in isolated word system and continuous speech system

In continuous speech experiment, the acoustic model is modelled basing on Chinese tritone model, which utilizes 1254 tonal syllables and 3206 HMM states. We utilize the recording data in “863” program of 83 people. We select the data of 70 people as the training set, and select the data of other 13 people as the test set.

In isolated-word GMM-HMM system, the feature observation probability utilizes 45-dimension single Gaussian full covariance matrix GMM. The result of the experiment is shown in Table 1, and we can see that comparing with the baseline system utilizing MFCC feature, utilizing MSLCF and MSSF can decrease the absolute error rate by 0.64% and 1.01%, respectively, which can also decrease the relative error rate by 6.10% and 9.66%, respectively. It can be seen that utilizing supervised states and state likelihood clustering features not only reduces the recognition error rate, but also improves the training and recognition efficiency.

Table 1. The experiment results of different features in isolated-word speech recognition system.

| Isolated-word experiment | Error rate of the system (Relative error rate decrease) |
|-------------------------|--------------------------------------------------------|
| MFCC                    | 10.46%                                                 |
| MSLCF                   | 9.82% (6.10%)                                          |
| MSSF                    | 9.45% (9.66%)                                          |
In continuous speech GMM-HMM system, the observation probability of feature uses two 45-dimension mixture full covariance matrix GMM. The result of the experiment is shown in Table 2. Comparing with the baseline system, utilizing MSLCF and MSSF can decrease the absolute error rate by 0.58% and 3.12, respectively, which can also decrease the relative error rate by 2.53% and 11.05%, respectively.

Table 2. The experiment results of different features in continuous speech recognition system.

| Continuous speech experiment | Error rate of the system (Relative error rate decrease) |
|------------------------------|-------------------------------------------------------|
| MFCC                         | 22.89%                                                |
| MSLCF                        | 22.31% (2.53%)                                        |
| MSSF                         | 20.36% (11.05%)                                       |

5. Conclusion
In this paper, based on the speech recognition system of GMM-HMM and Chinese triphone model, we successfully propose the SLCF and SSF features to improve the speech recognition performance over the original system. Under the circumstance that isolated-word speech database is limited, the supervised state distribution can also be obtained through continuous speech database training, which can assist the isolated-word recognition system. In this paper, we combine the acoustic feature MFCC with SSF and SLCF to get two novel features, i.e. MSSF and MSLCF, which contain both the acoustic feature and state information. The experiment shows that the MSSF and MSLCF features can significantly reduce the system error rate in Chinese continuous speech recognition system and isolated-word speech recognition system.

Acknowledgments
We would like to thank Zhijian Ou and Jiasong Sun at Speech Processing and Machine Intelligence Lab Tsinghua University for valuable discussions.

References
[1] Baker J M, Deng L, Glass J, Khudanpur S, and O'Shaughnessy D 2009 Developments and directions in speech recognition and understanding, part 1 [dsp education] IEEE Signal Processing Magazine 26 75-80
[2] Rabiner L R 1990 A tutorial on hidden Markov models and selected applications in speech recognition [J] Readings in Speech Recognition 77 267-96
[3] Sadaoki F 2000 Digital Speech Processing, Synthesis, and Recognition (New York: Marcel Dekker)
[4] Yu D, Deng L and Seide F 2013 The deep tensor neural network with applications to large vocabulary speech recognition IEEE Transactions on Audio Speech and Language Processing 21 388-96
[5] Zhijian O, Zuoying W 2002 A hybrid model from linear prediction HMM to a new speech recognition[J] Chinese Journal of Electronics 30 1313-16
[6] Pang Z, Shikui T U, Dan S U, Xihong W U and Lei X U 2011 Discriminative training of GMM-HMMacoustic model by RPCL learning. Frontiers of Electrical and Electronic Engineering 6 283-90
[7] Reynolds D A and Rose R C 1995 Robust text-independent speaker identification using Gaussian mixture speaker models IEEE Trans Speach & Audio Processing 3 72-83
[8] Haebumbach R, and Ney H 1992 Linear discriminant analysis for improved large vocabulary continuous speech recognition IEEE International Conference on Acoustics
[9] Chun L, Zuoying W 2001 Triphone recognition unit based on phonetics category[C] The 6th National Conference of Human Computer Speech Communication. Shenzhen, China: CIPSC 257-62
[10] You Z, Xiao X and Wang Z 2009 Continuous speech recognition based on the triphone DDBHMM Journal of Tsinghua University 49 506-09