



State-CodeBook based Quasi Continuous Density Hidden Markov Model with Applications to Recognition of Chinese Syllables

Ren-Hua Wang, Hui Jiang

Speech Communication Lab, University of Science and Technology of China,
P.O. Box 4, Hefei, Anhui, 230027, P.R.China

ABSTRACT

In this paper a new improved type of HMM, called State-CodeBook based quasi continuous density HMM (**SCBHMM**) is proposed and is tested in the recognition of Chinese syllables. The **SCBHMM** is composed of a set of model parameters, which can explicitly incorporate more acoustic characteristics of speech under limited training data. Here, the observation probability is associated with the static feature of speech within certain state and state transition probability is related to the temporal variations in speech spectra. **SCBHMM** suggests an effective method to integrate static and dynamic features in speech recognition. Preliminary experiments on the standard Chinese Speech Database CRDB¹, showed that the proposed **SCBHMM** not only achieved a great improvement over the original HMM, but also greatly reduced the computation consumption in the training process. An accuracy of more than 92% for the top one candidate and 99% for the top five candidates was achieved in the Chinese Syllables recognition.

1 Introduction

Since introduced to speech recognition, Hidden Markov Model(HMM) has been successfully applied to various speech recognition applications. However, a major limitation of HMM is that the normal HMM, consisted of a set of probability parameters and stochastic states, can not directly reflect the acoustic feature of a speech signal. HMM treats the speech feature vector of each frame as a stochastic vector and estimates its statistic distribution parameters based on a host of observations in the training data sets. Therefore, the state of HMM does not have any practical phonetic meanings about speech. The parameters of HMM, the observation probability distribution $b_j(k)$ or $b_j(X)$ and

¹The database CRDB was developed by Speech Communication Lab, Univ. of Scie. and Tech. of China (USTC). It is generally acknowledged as the standard database for Chinese syllable recognition in China.

state transition probability a_{ij} , only reflect the statistic characteristics of speech signal, are completely divorced from the acoustic characteristics of speech signal. Further, the simple one-order Markov hypothesis aggravates the inaccuracy of the model. In the literature, Poritz presented Linear Predictive Hidden Markov Model, which associates with each state s an all-pole filter A_s of degree N_s . M.Ostendorf suggested the Segment Statistic Model, that described the statistical dependence of all the frames of a speech segment. In this paper, we proposed a new improved type of HMM, called State-CodeBook based quasi continuous density HMM (**SCBHMM**), which is composed of a set of model parameters that can explicitly incorporate the more acoustic characteristics of speech while training data is insufficient. In this model, we make each HMM state directly correspond to a certain consecutive short-time stationary piece-wise speech segment. The observation probability distribution $b_j(X)$ can directly character the static characteristics of speech within a certain state and the state transition probability a_{ij} is related to the temporal changes of speech characteristics. It is a first try to associate the state transition a_{ij} with the temporal variations of the speech feature. In this way, **SCBHMM** suggested an effective method to integrate static and dynamic feature in speech recognition. In further, the experiments proved that the **SCBHMM** model was insensitive to the state partition, hence we can segment the speech in advance and avoid the iterative calculations in the Baum_welch algorithm. Therefore, a fast training algorithm for the **SCBHMM** was presented. We successfully apply the **SCBHMM** model to Chinese syllables recognition experiments. A recognition rate of 92% on top one candidate and nearly 99% on the top five candidates was achieved on an average. All our experimental results are reported on the Standard Chinese Speech Database CRDB, therefore they are trustworthy and comparable.

2 SCBHMM

Due to the defects of conventional HMM described above, we first assume that each state in the **SCBHMM** directly corresponds to a consecutive short-time relatively stationary piece-wise segment in the

speech stream. The topology of **SCBHMM** is simplified to left-to-right without state-skipping.

We consider the feature vectors belonged to the same state have more acoustic similarities than those belonged to different states. We suppose the observation probability distribution $b_j(X)$ is a function dependent on the current observation feature vector $X(t)$ and the state transition probability a_{ij} ² is dependent on the dynamic feature of current observation, such as the relative change of the current feature vector $\Delta X(t) = X(t) - X(t-\delta)$ (δ is a constant), i.e. a_{ij} is no longer a constant for a given state as that of normal HMM, it is related to the trend of current feature change in the **SCBHMM**. In this way, the capability of model to character the dynamic variations in the speech spectra is enhanced. Therefore, given the current observation feature vector $X(t)$ and the current state i in **SCBHMM**, we have

$$b_i(X(t)) = f_i(X(t)) \quad (1)$$

$$a_{ii}(X(t)) = g_i(\Delta X(t)) = g_i(X(t) - X(t - \delta)) \quad (2)$$

($i = 1, 2, \dots, N$)

$f_i()$ and $g_i()$ are the functions that describe the probability space distribution of speech feature belonged to the state i . As a result, the $b_i(X(t))$ directly character the static characteristics of speech with respect to the state i and the $a_{ii}(X)$ directly indicates the temporal changes in the spectra, which are believed to play an important role in the human perception. Therefore, the **SCBHMM** is believed to possess a greater capability to express acoustic property of speech than the conventional HMM.

The function $f_i()$ and $g_i()$ are distinct according to the different **SCBHMM** models or the different states in a **SCBHMM** model. How to accurately and conveniently express them and automatically estimate them from the training data is the key of **SCBHMM**.

Here we introduce two small codebooks SCB_i and $DSCB_i$ in each state for the function $f_i()$ and $g_i()$ respectively. Because of the similarity of the feature within a state, we suppose they are distributed in a relatively concentrated region of the feature space. It is reasonable for us to express feature distribution within a state with a very small codebook. Supposing the feature vectors corresponding to the state i are $X(1), X(2), \dots, X(L)$, we may cluster them into M classes with the LBG algorithm and get the codebook SCB_i at the same time. After we easily compute the differenced vectors with respect to state i , $\Delta X(1), \Delta X(2), \dots, \Delta X(L)$ ($\Delta X(t) = X(t) - X(t - \delta)$), we can also cluster them to obtain the codebook $DSCB_i$, which is close to the distribution of the temporal spectra changes in the feature space. Attention, the small codebook SCB_i and $DSCB_i$ are used in

²It becomes state-remaining probability a_{ii} in our model topology (from left-to-right and without state-skipping)

an utterly different way from the codebook in the normal Discrete Density HMM (DDHMM). If the current state is i and the observation vector is $X(t)$, instead of vector quantization as that in the DDHMM, we compute a **Similarity** $D_i(X)$ for $X(t)$ using the codebook SCB_i . $D_i(X)$ practically reflects how the input feature vector $X(t)$ accords with the feature space distribution described by codebook SCB_i . Different computation methods for $D_i(X)$ lead to various **SCBHMM** algorithms. Here we choose the simplest way,

$$D_i(X(t)) = \min_{Q_{ij} \in SCB_i} d(X(t), Q_{ij}) \quad (3)$$

$d(X(t), Q_{ij})$ is the distance measure between the $X(t)$ and a codeword Q_{ij} in the codebook SCB_i . Another alternative method for $D_i(X)$ is

$$D_i(X(t)) = \min_{Q_{ij} \in SCB_i} (X(t) - Q_{ij})^T [\sigma_{ij}]^{-1} (X(t) - Q_{ij}) \quad (4)$$

$[\sigma_{ij}]$ is covariance matrix of Voronoi cell corresponding to codeword Q_{ij} .

Further, we suppose that the relationship between current observation vector probability $b_i(X)$ and **Similarity** $D_i(X)$ can be expressed by the exponential function. i.e.

$$b_i(X) \propto e^{-D_i(X)} \quad (5)$$

so we can deduce the observation vector probability from the **Similarity** $D_i(X)$

$$b_i(X) = w_i e^{-D_i(X)} \quad (6)$$

let

$$j^* = \arg \min_{Q_{ij} \in SCB_i} d(X(t), Q_{ij}) \quad (7)$$

then

$$b_i(X) = w_i e^{-d(X(t), Q_{ij^*})} \quad (8)$$

w_i is a fixed weighting coefficient for the state i .

On the other hand, we can compute the current differenced feature vector $\Delta X(t) = X(t) - X(t - \delta)$. Similarly we can calculate another **Similarity** $DD_i(X)$ for $\Delta X(t)$ using the codebook $DSCB_i$,

$$DD_i(X(t)) = \min_{q_{ij} \in DSCB_i} d(\Delta X(t), q_{ij}) \quad (9)$$

Based on the same assumption, the state remaining probability

$$a_{ii}(X(t)) \propto e^{-DD_i(X(t))} \quad (10)$$

i.e.

$$a_{ii}(X(t)) = v_i e^{-DD_i(X(t))} \quad (11)$$

let

$$j^* = \arg \min_{q_{ij} \in DSCB_i} d(\Delta X(t), q_{ij}) \quad (12)$$

then

$$a_{ii}(X(t)) = v_i e^{-d(\Delta X(t), q_{ij^*})} \quad (13)$$

v_i is another weighting coefficient for state i . v_i is considered to be directly proportional to the magnitude of the constant state transition a_{ii} of the conventional HMM.

Unlike the normal HMM, the **SCBHMM** is composed of following parameters:

- N : state number in a **SCBHMM** model.
- M, M' : size of codebook SCB_i and $DSCB_i$.
- $SCB_i, DSCB_i$: static and dynamic feature vector codebook for state i . ($i=1,2,\dots,N$)
- $[\sigma_{ij}], [\sigma'_{ij}]$: covariance matrix for each codeword in SCB_i and $DSCB_i$ respectively. They are optional according to the methods to calculate the $D_i(X)$ and $DD_i(X)$.

Based on all of these parameters, we calculate the $b_j(X)$ and $a_{ij}(X)$ using the methods described above, then we can continue to use various algorithms in the conventional HMM for speech recognition, such as Baum-Welch and Viterbi algorithm.

3 Training and Recognition with SCBHMM

The experimental results showed that **SCBHMM** was insensitive to the state partition of speech signal. It is possible to presegment the speech signal and avoid the iterative calculations in the Baum-welch algorithm. Meanwhile, the size of codebook SCB_i and $DSCB_i$ is very small. Therefore, we can derive a fast training method for **SCBHMM**, which will greatly reduce the computation consumption over the Baum-welch algorithm. We segment the speech into N segments, each is relevant to a state in order. There are several methods to presegment the speech. We refer to a simpler segmenting algorithm, which segments it according to the relative distance between two adjacent frames.

Supposing the input speech feature vector sequence are $X(1), X(2), \dots, X(L)$. We denote the optimal segment point:

$$\{O_i | i = 1, 2, \dots, N, 0 \leq O_1 \leq O_2 \leq \dots \leq O_N = L\}$$

We defined the object distortion function:

$$D(O_1, O_2, \dots, O_N) = \sum_{i=1}^N \min_{Q_i \in V} \left[\sum_{s=O_{i-1}}^{O_i} d(X(s), Q_i) \right]$$

V indicates the whole feature space.

Obviously $\{O_i\}$ is the local extrema point of above function. i.e.

$$\{O_i\} = \arg \min_{0 \leq O_1 \leq O_2 \leq \dots \leq O_N = L} D(O_1, O_2, \dots, O_N)$$

However, we refer to a simpler segmenting algorithm, which segments it according to the relative distance between two adjacent frames.

The simple presegment algorithm can be written as follows:

1. First computing $\Delta_0 = \frac{\sum_{i=1}^{L-1} d(X(i), X(i-1))}{N}$
2. let $i=0, n=0, R=d(X(0), X(1))$
3. Let $X(i)$ belong to state n .
4. If $R \geq \Delta_0$ then $n = n + 1$, $R = R - \Delta_0$
else $R = R + d(X(i-1), X(i))$

5. $i=i+1$ and goto 3. until the end of speech.

After presegment, we use the normal K-means algorithm to cluster all feature vectors with respect to state i from various training samples to generate the codebook SCB_i and compute $[\sigma_{ij}]$ at the same time. Similarly, we can cluster all corresponding dynamic feature vectors with respect to state i into codebook $DSCB_i$ and get $[\sigma'_{ij}]$. So far, we obtain all parameters of **SCBHMM**. So we complete the whole training procedure of **SCBHMM**.

In the recognition process, we still presegment the input speech into N segments, each of which is corresponding to a state in order. For a given **SCBHMM** λ_k and segmented speech sequence

$$\bar{O} = X_1^1 X_2^1 \dots X_{L_1}^1 X_1^2 X_2^2 \dots X_{L_2}^2 \dots X_1^N X_2^N \dots X_{L_N}^N$$

symbol X_j^i denotes that the vector $X(t)$ is the j th vector within the state i , L_i is the number of vectors belonged to state i . The priori probability

$$\Pr(\bar{O} | \lambda_k) = \prod_{i=1}^N \left[\prod_{j=1}^{L_i} b_i(X_j^i(t)) \prod_{j=1}^{L_i-1} a_{ii}(X_j^i(t)) a_{i,i+1}(X_{L_i}^i(t)) \right] \quad (14)$$

Then the recognition output is k^*

$$k^* = \arg \max_k \Pr(\bar{O} | \lambda_k) \quad (15)$$

We know, the normal HMM must implement a state optimization iterative procedure based on the probability parameters. However, the **SCBHMM** can avoid this onerous task. It is the reason why we call it quasi HMM.

4 Chinese Syllables Recognition Experiments

In Chinese every character is pronounced as a syllable and there exist total of more than 1200 syllables. Furthermore, Chinese is a tonal language and there are in general 4 tones, every syllable is assigned to a tone. When the differences in tone are disregarded, the total number of different syllable is reduced to 406. Moreover, Chinese syllable has a unified phonetic structure, every syllable is conventionally decomposed into Initial part (ShenMu) and Final part (YunMu). Because of its unified phonetic structure, the presegment strategy is particularly suitable to the isolated chinese syllable recognition.

In the syllable recognition experiments, we build a **SCBHMM** model for each toneless syllable. (Recognition of tone is not considered here) Hence there are total 406 **SCBHMM** models. The used speech data is a subset of the Standard Chinese Recognition Speech Database CRDB, we select total 12 times of Chinese syllables utterances, each time contains every utterance

of 1264 Chinese tonal syllables. Among them, 6 times are uttered by a female speaker and another 6 times are from a male speaker at different time.

4.1 Signal Preprocessing

The speech in the CRDB database was digitized by 16Khz, 16bits. The digitized speech waveform was pre-emphasized with a filter whose transfer function is $1 - 0.97z^{-1}$, and bilinear transformed with function $\frac{(z^{-1}-0.6)}{(1-0.6z^{-1})}$. Then the waveform is blocked into frames and each frame spans 256 samples and consecutive frames overlap by 128 samples. From these smoothed speech samples we compute LPC coefficients using autocorrelation methods with order 16, then a 16 LPC-derived cepstral coefficients are computed as the feature vector for each frame.

4.2 Training and Test Procedure

we select the state number for each model is 5 ($N = 5$) and size of codebook SCB_i is 8 ($M = 8$) and size of $DSCB_i$ is 16 ($M' = 16$). To select the weighting coefficient $w_i, v_i, i = 1, 2, \dots, N$ conveniently, we tied all weighting coefficients with respect to the same state of different **SCBHMM** model to the same value. Hence we totally have only 10 weighting coefficients. In Chinese syllable recognition, the Initial part always are suppressed by Final part. In the model we can regulate the w_i, v_i to counteract the influence.

4.3 Experimental Results

In table 1, we enumerate six results from a host of recognition experiments:

- Experiment 1: Using normal DDHMM, trained by 3 times of male speech and tested in the rest.
- Experiment 2: Using the SCBHMM, trained by 3 times of male speech and tested in the rest.
- Experiment 3: Using the SCBHMM, trained by 5 times of male speech and tested in the rest.
- Experiment 4: Using normal DDHMM, trained by 3 times of female speech and tested in the rest.
- Experiment 5: Using the SCBHMM, trained by 3 times of female speech and tested in the rest.
- Experiment 6: Using the SCBHMM, trained by 5 times of female speech and tested in the rest.

With the conventional DDHMM, we get the recognition rate less than 70% on top one candidate. With the **SCBHMM** model we obtain the recognition rate about 92% on top one candidate and almost 99% on top five candidates on an average. Since the CRDB is generally

acknowledged as the standard syllable database for isolated syllable recognition in China, experimental results are relatively trustworthy and comparable.

Exp.	Top 1	Top 2	Top 3	Top 4	Top5
No.1	68.3%	83.9%	89.1%	91.6%	93.4%
No.2	92.9%	97.9%	99.1%	99.5%	99.7%
No.3	94.9%	99.3%	99.6%	99.8%	99.8%
No.4	65.6%	83.1%	88.5%	91.7%	93.1%
No.5	91.6%	97.4%	98.7%	99.0%	99.2%
No.6	95.0%	98.6%	99.0%	99.2%	99.3%

Table.1 Recognition rate using the SCBHMM and DDHMM

5 Summary

It is our philosophy that the correct recognition rate can be improved with the model that can capture the more acoustic, phonetic and linguistic knowledge of speech. Following the philosophy, we proposed the **SCBHMM** and achieved a good result in the Chinese syllables recognition experiments. With the limited training data, the **SCBHMM** can explicitly incorporate more acoustic characteristics of speech. In the **SCBHMM**, a set of parameters which are directly associated with the acoustic property of speech were introduced, the observation probability distribution is relevant to the static feature of speech within a state and state transition probability is related to the temporal changes in speech feature. The **SCBHMM** also provides an effective method to integrate the static and dynamic feature in recognition procedure. The experimental results demonstrated that the proposed **SCBHMM** not only achieved a great improvement over the original HMM, but also greatly reduced the computation consumption in the training process. Therefore, the **SCBHMM** is a good alternative to the conventional HMM in the speech recognition.

References

- [1] Ren-Hua Wang, et al "Development of A Chinese Voice Database for Machine Recognition", ACTA AUTOMATICA SINICA, Vol 18, No.3, 1992
- [2] Yumin Lee, Lin-shan Lee, "Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition", Computer Speech and Language(1993) 7, 247-263.
- [3] A.B.Poritz, "Linear Predictive hidden Markov Model and speech signals", ICASSP 1982, p1291-1294
- [4] M. Ostendorf and S.Roukous "A Stochastic Segment Model for Phneme-Based Continuous Speech Recognition", IEEE trans. on ASSP P1857-1869, Dec. 1989.