ABSTRACT

A semi-continuous segmental probability model, which can be considered as a special form of continuous mixture segmental probability model with continuous output probability density functions sharing in a mixture Gaussian density codebook, is proposed in this paper. The amount of training data required, as well as the computational complexity of the semi-continuous segmental probability model (SCSPM), can be significantly reduced in comparison with the continuous segmental probability model (CSPM). Parameters of the vector quantization codebook and segmental probability model can be mutually optimized to achieve an optimal model/codebook combination, which leads to a unified modeling approach to vector quantization and segmental probability modeling of speech signals. The experimental results show that the recognition accuracy of the semi-continuous segmental probability model is higher than the semi-continuous hidden Markov model and continuous segmental probability model.

Keywords: hidden Markov model, segmental probability model, semi-continuous segmental probability model, speech recognition

1. Introduction

Currently, the acoustic-phonetic components of the most successful large-vocabulary automatic speech recognition (ASR) systems are almost exclusively based on hidden Markov models (HMMs) of some phonetically-defined subword units, including continuous HMM (CHMM) and Semi-continuous HMM (SCHMM) with a very similar specification for all the systems. However, these models are simply general statistical pattern matchers, whose inappropriate assumptions of HMM are linked with their frame-synchronous characteristic, whereby sates are associated with single acoustic feature vectors. Thus, some different modeling approaches are presented. Among those, segmental probability model (SPM)[1] incorporates the concept of modeling frame sequences rather than individual frames, with the aim of providing an accurate model of speech dynamics across an utterance in a way which takes into account predictable factors such as speaker continuity. Both CSPM and SCHMM have some apparent defects. For SCHMM, it still has a slightly higher computational complexity and a number of limited assumptions. As for CSPM, the first is the lack of the important description of duration information, and the second is that the effect of short duration segment is suppressed by that of the long one in the calculation of model likelihood.

Based on our modified SPM[3], SCSPM can overcome the limitations of the above two models. The semi-continuous output probability density function is represented by a combination of the discrete output probabilities of the model and the continuous Gaussian density functions of a mixture Gaussian density codebook.

The rest of the paper is organized as follows. MSPM is introduced in Section 2. Section 3 is dedicated to introduce SCSPM and the reestimation method. Some experimental results are given in section 4. Finally some conclusions are provided at the end of the paper, Section 5.

2. Modified Segmental Probability Model

The modified SPM is a Markov process. It can be depicted as that: the observation sequence of utterance, \( O = \{O_1, O_2, \ldots, O_T\} \) with \( T \) frames of feature vectors, is divided by segmentation algorithm into \( N \) segments, \( S = \{S_1, S_2, \ldots, S_N\} \). Assuming that, segment \( S_i \) and segment \( S_j \) are independent of each other in probability, where \( i \neq j \) and \( 1 \leq i, j \leq N \).

There are total \( M \) models, each model \( \lambda_j \ (j=1,2,\ldots,N) \) can produce its own probability \( p(\lambda_j) \), given the observation vector sequence \( O \) and the segmentation \( S \). The recognition procedure is virtually to find

\[
\lambda^* = \arg \max_{\lambda_j} \quad P(\lambda_j \mid O, S) = \arg \max_{\lambda_j} \frac{P(O, S \mid \lambda_j) P(\lambda_j)}{P(O, S)} \quad (1)
\]

Since \( P(O, S) \) is constant and when \( a \ priori \) knowledge is ambiguous \( P(\lambda_j) = P(\lambda_j), 1 \leq i, j \leq N \), the decision rule can be formulized as follows.

\[
\lambda^* = \arg \max_{\lambda_j} P(\lambda_j \mid O, S) = \arg \max_{\lambda_j} f(O, S \mid \lambda_j) \quad (2)
\]

*This project is supported by the Chinese National Natural Science Foundation (69982005) and “973” ( G199803050703).
Based on the theory of the conditional probability,

\[ f(O, S | \lambda) = f(O_i, S_i | \lambda) \]

\[ \prod_{i=2}^{N} f(O_i, S_i | \lambda, O_{i-1}, \lambda, O_{i-1}, S_{i-1}) \]  \hspace{1cm} (3)

where \( O_i \) is the observation vectors sequence of the \( i \)-th segment \( S_i \). By the independent assumption, there is

\[ f(O_i, S_i | \lambda, O_{i-1}, \lambda, O_{i-1}, S_{i-1}) = f(O_i, S_i | \lambda) \]

\[ = f(O_i | S_i, \lambda) \cdot P(S_i | \lambda) \]  \hspace{1cm} (4)

where \( f(O_i | S_i, \lambda) \) and \( P(S_i | \lambda) \) is the probability of the \( i \)-th segment observation sequence \( O_i \) and of \( S_i \) respectively. Thus, Equation (3) can be rewritten as

\[ f(O, S | \lambda) = \prod_{i=1}^{N} f(O_i | S_i, \lambda) \cdot P(S_i | \lambda) \]  \hspace{1cm} (5)

Since the independence between the segments, \( S_i \) relates not to other segments, only to its own segment length, \( \tau_i \). So \( P(S_i | \lambda) \) can be represented by the probable distribution of \( \tau_i \)

\[ P(S_i) = P(\tau_i) \]  \hspace{1cm} (6)

Now Equation (5) is transformed into

\[ f(O, S | \lambda) = \prod_{i=1}^{N} f(O_i | S_i, \lambda) \cdot \prod_{i=1}^{N} P(\tau_i | \lambda) \]  \hspace{1cm} (7)

We can find that the modified SPM not only retains the form of the segment observation probability, but also includes the product, \( \prod_{i=1}^{N} P(\tau_i | \lambda) \), which can be considered as the probability of the total states.

The normalization of duration and Gamma distribution of duration is used in the MSPM[3]. The normalized duration is

\[ \tau_i = \frac{\tau_i}{T} \]  \hspace{1cm} (8)

and Gamma distribution of the duration is

\[ P(\tau_i) = \frac{\gamma(\alpha_i, \beta_i)}{\Gamma(\alpha_i)} \cdot e^{-\beta_i \tau_i} \]  \hspace{1cm} (9)

The weighted geometric mean of the segment observation probability is defined as

\[ \bar{f}_i(O_i, S_i | \lambda) = k_i \left( \prod_{t=1}^{i} b_t(O_t) \right)^{1/i} \]  \hspace{1cm} (10)

where \( b_t(O_t) \) denotes the output probability of observation vector \( O_t \) and \( k_i \) stands for the weight coefficient, be set by experience or trained to stress some segments.

### 3. Semi-continuous Segmental Probability Model

It is well known that the difference between discrete, semi-continuous and continuous HMM lies in different form of output probability functions. The difference between semi-continuous and continuous SPM is the similar.

#### 3.1 Continuous Mixture Density Models

CSPM uses the multivariate Gaussian mixture density function as well as CHMM. In use of continuous output probability density functions, with \( M \) Gaussian mixture density functions, we have:

\[ b_{wk}(O_t) = \sum_{k=1}^{M} c_{wk} N(O_t | \mu_{wk}, \Sigma_{wk}) = \sum_{k=1}^{M} c_{wk} b_{wk}(O_t) \]  \hspace{1cm} (11)

where \( N(O_t | \mu_{wk}, \Sigma_{wk}) \) or \( b_{wk}(O_t) \) denotes \( w \)-th model and \( \mu_{wk} \) and \( \Sigma_{wk} \) are mean vector and covariance matrix for \( w \)-th segment’s single mixture Gaussian density function of mean vector \( \mu_{wk} \) and covariance matrix \( \Sigma_{wk} \) which includes the number of mixture-components; and \( c_{wk} \) is the weight for the \( k \)-th mixture component satisfying

\[ \sum_{k=1}^{M} c_{wk} = 1 \]  \hspace{1cm} (12)

#### 3.2 Semi-continuous Models

The semi-continuous model assumes the mixture density functions tied together across all the models to form some shared kernels. Thus, the VQ codebook typically used in the discrete model can be regarded as one of such shared kernels. Accordingly, equation (11) can be modified as:

\[ b_{wk}(O_t) = \sum_{k=1}^{M} c_{wk} N(O_t | \mu_k, \Sigma_k) \]  \hspace{1cm} (13)

\( N(O_t | \mu_k, \Sigma_k) \) is assumed to be independent of the segmented probability model and they are shared across all the models with a very large number of mixtures \( M \).
3.3 Re-estimation of SCSPM

Because of the binding of the continuous density functions, in the SCSPM, the number of free parameters as well as computational complexity is reduced in comparison to the CSPM.

We express the output probability \( f(O|S, \lambda) \) with respect to each single mixture component as:

\[
\begin{align*}
\hat{f}(O|\lambda, S) & = \sum_{w=1}^{W} P_w \sum_{u=1}^{U} U_w \sum_{s=1}^{S} f(O_u, \lambda_s, S_w) \\
& = \sum_{w=1}^{W} P_w \sum_{u=1}^{U} U_w \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) N(O_{u_t}, \mu_u, \Sigma_u) 
\end{align*}
\]  

where \( P_w \) denotes the prior probability of \( w \)th model and \( U_w \) is the number of \( w \)th model's train data.

First, derive the re-estimation formulation of weighting coefficients, \( c_{uw,m} \). Define an auxiliary function with the condition (12):

\[
Q(\hat{\lambda}, \hat{x}_{uw}) = \sum_{u=1}^{U} \sum_{w=1}^{W} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) \frac{f(O_{u_t}, \lambda_u, S_u)}{f(O_{u_t}, \lambda_{s_{t-1}}, S_w)} 
\]

where \( \hat{\lambda} \) and \( \hat{x}_{uw} \) denotes the parameters set of means and variations of \( m \)th code and \( \hat{x}_{uw} \) denotes the observation vector of \( u \)th sample of \( m \)th model.

By the algorithm of Lagrange’s factor, \( \Theta \):

\[
\hat{Q}(\lambda, \hat{x}_{uw}) = \hat{Q}(\lambda, \hat{x}_{uw}) - \Theta \left( 1 \sum_{w=1}^{W} \sum_{m=1}^{M} \hat{c}_{uw,m} - 1 \right) 
\]

It can be shown that the re-estimation equations for \( c_{uw,m} \) is:

\[
\hat{c}_{uw,m} = \frac{\sum_{u=1}^{U} \sum_{w=1}^{W} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)}{\sum_{m=1}^{M} \sum_{w=1}^{W} \sum_{u=1}^{U} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)} 
\]  

The re-estimation formulations can be more readily computed by defining two set of variables \( \alpha_{uw,m}(m) \) and \( \beta_{uw,m} \):

\[
\alpha_{uw,m}(m) = c_{uw,m} N(O_{uw_{m+1}}, \mu_{uw}, \Sigma_{uw}) 
\]

\[
\alpha_{uw,m}(m) = c_{uw,m} N(O_{uw}, \mu_u, \Sigma_u) \sum_{i=1}^{M} \alpha_{uw-1,i} 
\]

\[
\beta_{uw,m} = 1 
\]

So, the equation (17) can be rewritten as

\[
\hat{c}_{uw,m} = \frac{\sum_{u=1}^{U} \sum_{w=1}^{W} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)}{\sum_{m=1}^{M} \sum_{w=1}^{W} \sum_{u=1}^{U} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)} 
\]  

Second, derive the re-estimation formulation of \( \mu_u \). The auxiliary function is:

\[
\hat{Q}(\lambda, \hat{x}_{uw}) = \sum_{u=1}^{U} \sum_{w=1}^{W} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) \frac{f(O_{u_t}, \lambda_u, S_u)}{f(O_{u_t}, \lambda_{s_{t-1}}, S_w)} 
\]

\[
\hat{Q}(\lambda, \hat{x}_{uw}) = \sum_{u=1}^{U} \sum_{w=1}^{W} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) \frac{f(O_{u_t}, \lambda_u, S_u)}{f(O_{u_t}, \lambda_{s_{t-1}}, S_w)} 
\]

\[
\beta_{uw,m} = \sum_{k=1}^{M} c_{uw,k} N(O_{uw+1}, \mu_k, \Sigma_k) \beta_{uw+1} 
\]

Thus,

\[
\hat{\mu}_{uw,d} = \frac{\sum_{u=1}^{U} \sum_{w=1}^{W} \sum_{m=1}^{M} f(O_{uw}, \lambda_u, S_u) f(O_{uw}, \lambda_u, S_u) O_{uw_{m+1}}}{\sum_{m=1}^{M} \sum_{w=1}^{W} \sum_{u=1}^{U} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)} 
\]

Finally, the re-estimation formulation of \( \sigma_{uw,d} \) can be written as:

\[
\hat{\sigma}_{uw,d} = \frac{\sum_{u=1}^{U} \sum_{w=1}^{W} \sum_{m=1}^{M} f(O_{uw}, \lambda_u, S_u) f(O_{uw}, \lambda_u, S_u) O_{uw_{m+1}}^{2}}{\sum_{m=1}^{M} \sum_{w=1}^{W} \sum_{u=1}^{U} \prod_{t=1}^{t-1} f(O_{u_{t-1}}, \lambda_{s_{t-1}}, S_w) f(O_{u_t}, \lambda_u, S_u)} - \hat{\mu}_{uw,d}^{2} 
\]
Based on the re-estimation algorithm stated above, parameters of the vector quantization codebook and segmental probability model can be mutually optimized to achieve an optimal model/codebook combination, which leads to a unified modeling approach to vector quantization and segmental probability modeling of speech signals. It should be noted that the computational complexity for decoding with this new SCSPM is less than that of CSPM if the size of the VQ codebook is less than the number of output probability density functions in the continuous mixture SPM since \( f(\mathbf{O} | \mathbf{S}, \lambda) \) need only be calculated for each codebook index as opposed to each state with several mixture density functions. The segmentation can be performed by nonlinear algorithm. The Gamma distribution of duration and the weighted geometric means of observation probability in each segment are involved in the semi-continuous segmental probability model. These characters can overcome the limitations of CHMM.

4. Experimental Results
For both training and evaluation, the probability outputs a 39-dimensional vector. The first component of the vector consists of 12 MFCC coefficients and energy term. The other components of the feature vector are the first and second time derivatives of the first 13 components. Two experiments are test on the bi-word Chinese phrases. The first one train set is 30000 phrases, spoken by 30 persons and the test set is 10000 syllables spoken by 10 persons.

Table I: Comparison of CSPM, SCSPM and SCHMM on phrases with more train data

<table>
<thead>
<tr>
<th>Model</th>
<th>VQ level</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSPM</td>
<td>-</td>
<td>94.79%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>512</td>
<td>93.21%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>256</td>
<td>91.38%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>64</td>
<td>83.69%</td>
</tr>
<tr>
<td>SCHMM</td>
<td>256</td>
<td>87.31%</td>
</tr>
</tbody>
</table>

The second experiment, train test is 10000 phrases, spoken by 10 persons. The test set is 10000 phrases spoken by other 10 persons.

Table II: Comparison of CSPM, SCSPM and SCHMM on phrases with less train data

<table>
<thead>
<tr>
<th>Model</th>
<th>VQ level</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSPM</td>
<td>-</td>
<td>79.71%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>512</td>
<td>86.88%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>256</td>
<td>85.64%</td>
</tr>
<tr>
<td>SCSPM</td>
<td>64</td>
<td>76.41%</td>
</tr>
<tr>
<td>SCHMM</td>
<td>256</td>
<td>85.61%</td>
</tr>
</tbody>
</table>

From the two tables above, the VQ level varies from 64 to 512, the average phrase recognition accuracy of SCSPM will increase. With sufficient data, the recognition accuracy of SCSPM is comparable to SCSPM. In addition, when the data is less, the result of SCSPM (VQ level 256,512) is better than CSPM. Because the number of parameters need to be estimated in SCSPM is less than in SCSPM. In two experiments, the recognition accuracy of SCSPM is better than SCHMM with the same VQ level, 256.

5. Conclusion
The SCSPM take the advantage of SPM and has low computational complexity. It results in a powerful tool for modeling time-varying signals sources. The SCSPM technique described here can be considered as a method where a VQ codebook and model parameters are optimized together. Benefiting from the Gamma distribution of the duration and weighted geometric mean of the segment observation probability, SCSPM is a powerful new technique for modeling non-stationary stochastic processes with multi-model non-symmetric probabilistic functions of SPM.

Reference