UNKNOWN-MULTIPLE SIGNAL SOURCE CLUSTERING PROBLEM USING ERGODIC HMM AND APPLIED TO SPEAKER CLASSIFICATION

*J. Murakami †, M. Sugiyama ‡, H.Watanabe **

†NTT Information and Communication Systems Laboratories E-mail: murakami@isl.ntt.jp
‡The University of Aizu, School of Computer Science and Engineering
* ATR Interpreting Telecommunication Research Laboratories

ABSTRACT

In this paper, we consider signals originated from a sequence of sources. More specifically, the problems of segmenting such signals and relating the segments to their sources are addressed. This issue has wide applications in many fields. This report describes a resolution method that is based on an Ergodic Hidden Markov Model (HMM), in which each HMM state corresponds to a signal source. The signal source sequence can be determined by using a decoding procedure (Viterbi algorithm or Forward algorithm) over the observed sequence. Baum-Welch training is used to estimate HMM parameters from the training material. As an example of the multiple signal source classification problem, an experiment is performed on unknown speaker classification. The results show a classification rate of 79% for 4 male speakers. The results also indicate that the model is sensitive to the initial values of the Ergodic HMM and that employing the long-distance LPC cepstrum is effective for signal preprocessing.

1. INTRODUCTION

In this paper, we consider signals originating from a sequence of sources. More specifically, the problems of segmenting such signals and relating the segments to their sources are addressed. This issue has wide applications in many fields. For example, the automatic determination of the acoustic unit, speaker discrimination, and the discrimination of the utterance mode are candidates

The unknown-multiple signal source clustering problem divides into the following four sub- problems.

1. Extract the feature parameter that characterizes each category.
2. Determine the segmentation point to best represent category transition.
3. Classify segmented blocks for N-categories.
4. Estimate the number of categories (N).

This paper assumes that the number of categories N is known. Therefore, we consider the problem of automatically segmenting the observed signal sequence and classifying the category for each segmented interval.

On the other hand, in the areas of speaker identification [3] and language modeling [5], the Ergodic HMM is often used, where all states are connected to each other. When such an Ergodic HMM is applied to the unknown-multiple signal source clustering problem, it is expected that the category corresponds to the state, and the signal sequence corresponds to the symbol sequence output from the state. Therefore, the signal source sequence can be determined by using the Viterbi algorithm over the observed sequence. Baum-Welch training is also used to estimate HMM parameters from observed sequences.

As an example of the unknown-multiple signal source cluster problem, an experiment is performed on speaker classification. Each speaker speaks randomly and this speech data is recorded using only one channel. The purpose is to segment the speech data according to the speaker. The following results are obtained. Using LPC cepstrum with a long-term window for 4 male speakers, the average classification rate is 67.5%. Selecting the Ergodic HMM with high likelihood yields the average classification rate of 78.8%.

Relevant studies are as follows. A method using Kullback information based on the code book is presented in[1]. In this study, segmentation boundaries and the number of categories are already known. Papers [4] discuss the same classification problem for multiple speaker utterances. They assumed that the acoustic parameter follows a Gaussian distribution for each speaker, and try to solve the problem using VQ clustering. We note that for the normal speaker identification problem [3], each speaker model is constructed beforehand. These models are applied to the input speech to identify the speaker. This paper considers capturing the speech sequential speech of several speakers, and the problem is to segment and classify the speech of each speaker.

2. UNKNOWN-MULTIPLE SIGNAL SOURCE CLUSTERING PROBLEM

2.1. Formulation of problem

Let the signal sequence be $X = (x_t)(t = 1, 2, ..., T)$, which is generated by n signal sources, as is shown in Fig.1. It is assumed that the sequence consists of K blocks X_k ($k = 1, 2, ..., K$) , and each block is generated from one of $N(\leq K)$ categories C_j ($j = 1, 2, ..., N$) . Block X_k is written as $(x_{t_{k-1}+1}...x_{t_k})$ (where $t_0 = 0, t_K = T$), and the time sequence is set as M_k $(= t_k - t_{k-1})$. The unknown-multiple signal source clustering problem is that, where the signal sources are given, to find the block boundaries $t_k(k = 1, 2, ..., K - 1)$ based on signal characteristics. The K blocks are to be classified into $N(\leq K)$ categories. Also, the number of categories N is to be estimated.

3. SOLUTION METHOD USING ERGODIC HMM

3.1. Ergodic HMM

When the number of categories N is known and segmentation boundaries are unknown, it is possible to apply the Ergodic HMM. In this case, one can consider that 'category' corresponds to 'state' and the signal sequence corresponds to the symbol generated from the state. The problem divides into the following two problems.

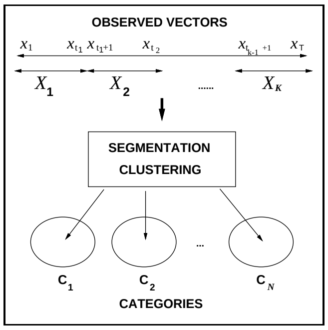

Figure 1. Definition of N-signal Problem

1. The problem of estimating the HMM parameter X that maximizes the likelihood of the signal sequence. Parameter X consists of the initial state probability $\pi = (\pi_i)$, the state transition probability $A = (a_{ij})$ and the symbol output probability $B = (b_j(l))$.
2. The problem of estimating the state transition sequence that generate the highest probability of outputting the signal sequence X for the HMM parameter M (estimate of optimal state sequence).

Fig.2 outlines our solution.

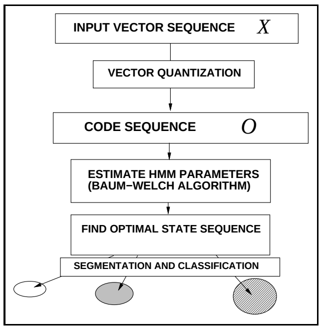

Figure 2. Flow chart that includes an Ergodic HMM for the N-signal source problem

3.2. HMM parameter estimation

It is known that the HMM parameter M can be estimated by the Baum-Welch algorithm. This is a kind of EM algorithm [2] whose HMM parameters are iteratively calculated. This training algorithm runs to local maxima, so the initial parameters are important.

3.3. Estimation of optimal state sequence

In the following, the problem considered is that of estimating the optimal state transition sequence for generating the signal sequence.

For such a purpose, Viterbi algorithm is well known. Also we propose the forward algorithm which is shown below. Using the optimal state transition sequence $S^* = \{s_1^*, ..., s_T^*\}$ obtained by such algorithms, the segmentation boundaries and the category identification are directly derived.

3.3.1. Viterbi algorithm

It is well known that the Viterbi algorithm can be used to estimate the optimal state sequence that generates the observed sequence. Detail is described in literature[2].

3.3.2. Forward algorithm

To estimate the optimal state sequence, we propose the new algorithm called the forward algorithm.

This algorithm summarizes the likelihood for all states. The optimal state sequence is then chosen as the maximum likelihood state at each time (frame).

This forward algorithm is described as follows.

1. For all$i \in \{1, ..., N\}$, let
$$\delta_1(i) = \pi_i \times b_i(o_1)$$
$$s_1^* = arg\max_i \delta_1(i)$$
2. Along the time axis $t = 2, ..., T$, for all$j \in \{1, ..., N\}$, let
$$\delta_t(j) = \sum_i [\delta_{t-1}(i) \times a_{ij} \times b_j(o_t)]$$
$$s_t^* = arg\max_j \delta_t(j)$$
3. The likelihood for all possible state transition sequences and the optimal state at time T are given as follows.
$$\sum_S P(O, S \mid M) = \sum_j \delta_T(j)$$
$$s_T^* = arg\max_j \delta_T(j)$$

4. SPEAKER CLASSIFICATION PROBLEM

In this paper, the speaker classification problem is addressed as an example of unknown-multiple signal source clustering. In this problem, the signal sequence corresponds to the LPC cepstrum, each category corresponds to a speaker, and the segmentation boundaries correspond to speaker transitions.

4.1. Speech Data

For the experiment, we made a pseudo speech using word utterance speech data (ATR 5240 word utterance speech data). The number of speakers was 4. Each speaker uttered 8 blocks, each of which consisted of 8 different words. Silences before and after each word were deleted. The words were randomly concatenated so each speech set consisted of 32 blocks and 31 speaker transitions. Fig.3 shows an example of the speech data. We use 8 sets of speech data in the experiment. Average set length was about 150 seconds.

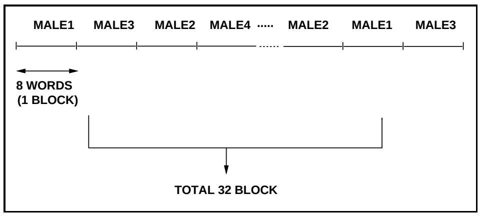

Figure 3. Fig.3 Example of speech data (Speaker: 4 males)

4.2. Analysis condition of Acoustic Parameter

In this experiment, the 16th order LPC cepstrum was used as the acoustic parameters representing speaker features. Table 1 shows the condition of the speech analysis. The universal code book is constructed using the Euclid distance from the speech data.

4.3. Initial HMM parameters

In this study, we used the Moore type discrete Ergodic HMM. The number of states (categories) of the Ergodic HMM was set at 4, the same as the number of speakers.

It is well known that the Baum-Welch algorithm runs toward local maxima, Therefore, the speaker classification rate seems greatly depends on the initial parameters. Accordingly, the initial parameters were calculated in the following three ways.

Table 1. analysis condition

Acoustic parameters	LPC cepstrum
sampling period	12 kHz
LPC	16 dimension
LPC analysis	14 order
window length	21.3 ms (256 point)
window interval	10.7 ms (128 point)
high frequency weighting	$(1 - 0.97z^{-1})$
universal code book size	256

1. Experiment 1 : (true values are given to all parameters)
   As the initial HMM parameters M, true values were given to the initial state probability $\pi^{(0)}$, the state transition probability $A^{(0)}$ and the symbol output probability $B^{(0)}$.
2. Experiment 2 : (true value is given only to the symbol output probability)
   Symbol output probability $B^{(0)}$ was given its true value. Initial state probability $\pi^{(0)}$ and the state transition probability $A^{(0)}$ were given uniform probabilities.
3. Experiment 3 : (random values)
   The initial state probability $\pi^{(0)}$ and the state transition probability $A^{(0)}$ were given uniform probabilities. The symbol output probability $B^{(0)}$ was assigned random probabilities. Note that the Baum-Welch learning algorithm does not work if a uniform probability is given to $B^{(0)}$.

4.4. Evaluation Method of Speaker Classification Rate

Even though the optimal state sequence is obtained using forward decoding or Viterbi decoding, the relation between the classification and the speaker is still unknown. Therefore, we calculated the classification rate by the following expression.

$$R = \frac{1}{T} \max_{\sigma} \sum_{t=1}^{T} d(\tau(x_t), \sigma(S_t)) \qquad (1)$$

In the above, τ is the optimal state sequence. σ is an arbitrary permutation of $(1, 2, \ldots, N)$, and S_t is the correct speaker of the utterance. d is the variable that takes value 1 if the values agree, and 0 if otherwise.

In this study, the classification numbers are related to σ, the correct classification rate is calculated for each of $N!$ permutations, and the maximum is defined as the classification rate. Consequently, in the case of 4 speakers, $24(= 4!)$ combinations are examined.

5. EXPERIMENTAL RESULTS

Fig.4 shows the result of experiment, where the LPC cepstrum is calculated with the window length is 21.3ms. The average classification rate is given by the average for 8 sets of speech data for experiments 1 and 2. In experiment 3, 16 different initial models were tried for each of the 8 sets of speech samples. Therefore, the rate is the average of 128 trials.

The following observations can be obtained from this figure.

1. The average speaker classification rates of the Viterbi and the forward algorithm differ little. However, in experiment1 , (true values are given to all parameters), forward decoding gives 94.0 %, while Viterbi decoding gives only 48.3 %.

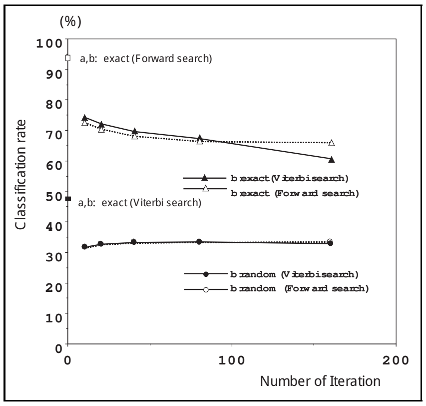

Figure 4. Relationship between number of iterations and classification rate

2. In experiment 2 (the true value is given only to the symbol output probability $B^{(0)}$), the average classification rate is about 75%. However, the average classification rate decreases with training.
3. In experiment 3 ($B^{(0)}$ is set at random), the average classification rate is low, being 30% to 35%. This value is not improved even if the amount of training is increased.
   The reason seems to be as follows. In this experiment, the LPC analysis condition is the same as used for speech recognition [2]. Thus the parameters have a phoneme feature, but not a speaker feature.

5.1. Speaker Feature and Long Window Analysis

It is known from a study of speaker identification that the long-term average spectrum is useful. Therefore, it is expected that the speaker classification rate will be improved by increasing the LPC analysis window length. From such a viewpoint, we examined the classification rate with various LPC analysis window lengths. The universal code book size was set at 64 and 256. In both cases, the frame length (window interval)was set at half the LPC analysis window length. For each speech data, 16 kinds of random initial models were constructed using random variables. The average for 8 sets of speech data, i.e., the average for 128 trials in total, was taken as the average speaker classification rate. The end condition of HMM training was set to 160 iterations. Other experimental conditions were the same as in experiment 3. Fig.5 shows the results of this experiment.

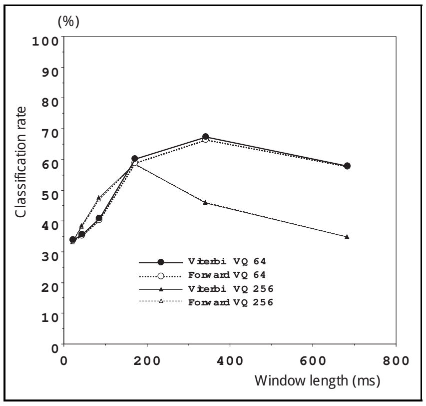

Figure 5. Relationship between window length and speaker classification rate

The following observations can be obtained from this figure.

1. When the LPC analysis window length is long, the average classification rate is increases, but decreases beyond a certain point.
2. The classification rate is the highest when the code book size is 64 and the analysis window length is 341ms.

5.2. Dependency of classification rate and likelihood

The Baum-Welch training algorithm runs toward local maxima so the likelihood of Ergodic HMM depends on the initial parameters. From such a viewpoint, the relation between the classification rate and the likelihood was examined. In the experiment, the LPC analysis window length was 341.3ms, the code book size was 64 and HMM training was iterated 160 times. Other experimental conditions were the same as in experiment 3. We use the 8 speech data sets and 16 initial models were used for each set. Therefore, 128 experiments were carried out. The results are shown in Fig.6. This figure show that there is a relationship between HMM likelihood and the classification rate.

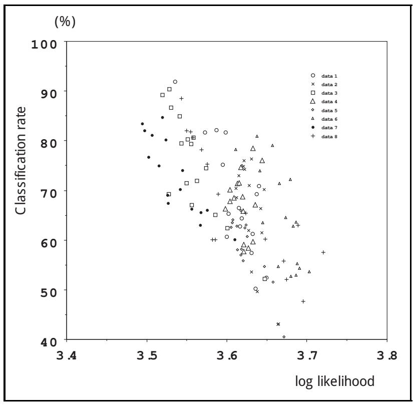

Figure 6. Relationship between HMM likelihood and classification rate

5.3. Selection of initial model

In the previous experiment, the model with high likelihood gave the highest speaker classification rate. We tested this relationship in the following experiment. Sixteen different initial models were constructed at random. After Baum-Welch learning, the model with the highest likelihood was selected. Fig.7 shows the average classification rates. The LPC analysis window length is set at 341.3 ms. The universal code book size was varied from 32 and 256. Other experimental conditions were the same as in experiment 3. As can be seen from these results, the average classification rate is 78.8% which means that the performance is improved by about 10%.

6. DISCUSSIONS

This paper described the problem of identifying multiple speaker utterances, as an example of the unknown-multiple signal source clustering problem. However many unsolved problems remain as described below.

1. Evaluation of classification rate
   It is necessary to develop a method for evaluating the speaker classification rate when the number of speakers is large. This paper examined all possibilities, and the highest value was taken as the classification rate. However, the number of possible combinations is the factorial of the number of speakers. Thus, it is necessary to speed up the evaluation.

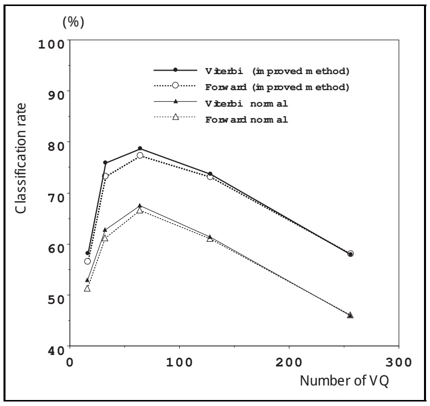

Figure 7. Comparison of selection method and base-line method

2. Time resolution
   Speaker transition occurred in some frames because of the LPC analysis. In such a frame, the speaker can not be determined uniquely. In other words, the time resolution of the speaker classification depends on the LPC frame window length. This problem must be studied and resolved.
3. Estimation of the number of categories N
   In this experiment, the number of speakers (the number of categories) was set as 4. This means that the number of speakers is a priori knowledge. A technique is needed that can estimate the number of speakers.

7. CONCLUSION

This paper considered the problem of decomposing a signal sequence into multiple signal sources, and proposed a method based on the Ergodic HMM. As an example of this problem, the speaker classification problem was considered, and speaker classification experiments were carried out. The following results were obtained.

1. The initial parameters of the Ergodic HMM are important in determining the segmentation boundaries and the category simultaneously.
2. Among the initial HMM parameters, the symbol output probability is the most important in obtaining good performance.
3. In the speakers classification problem, an excellent classification rate is obtained by using the LPC long cepstrum (the LPC analysis window length was 341ms in these experiments).
4. The average speaker classification rate is improved by selecting the Ergodic HMM that has high likelihood.

REFERENCES

[1] M.Sugiyama, J.Murakami H.Watanabe, "Speech Segmentation and Clustering Based on Speaker Feature", ICASSP93, 2-398, pp.2-395.2-398

[2] X.D. Huang, Y. Ariki and M.A. Jack, "Hidden Marcov Models for Speech Recognition", Edinburgh University Press, Edinburgh (1990).

[3] T.Matsui, S.Furui, *Speaker Recognition Based on Ergodic HMMs*, 3-6-14 (Oct. 1991) (in Japanese).

[4] G.Yu, M.Siu, H.Gish, "An Unsupervised, Sequential, Learning Algorithm for the segmentaion of Speech Waveform with Multiple Speakers", Proc. of ICASSP92, 2-189 (Apr.1992).

[5] K.Lari, S.J.Young, "The estimation of stochastic context-free grammars using the Inside-Outside algorithm" , speech recoginition, Computer Speech and Language, vol. 4, pp.35-56 (1990).