Automatic Generation of Context-Independent Variable Parameter Models using Successive State and Mixture Splitting

Soo-Young SUK† , Ho-Youl JUNG‡ , Hyun-Yeol CHUNG‡

School of EECS, Yeungnam University
214-1, Dae-Dong, Gyung-San, Gyungbuk, Republic of Korea
† lover@yumail.ac.kr, ‡ {hoyoul, hychung}@yu.ac.kr

Abstract

A Speech and Character Combined Recognition System (SCCRS) is developed for working on PDA (Personal Digital Assistants) or on mobile devices. In SCCRS, feature extraction for speech and for character is carried out separately, but recognition is performed in an engine. The recognition engine employs essentially CHMM (Continuous Hidden Markov Model) structure and this CHMM consists of variable parameter topology in order to minimize the number of model parameters and to reduce recognition time. This model also adopts our proposed SSMS (Successive State and Mixture Splitting) for generating context independent model. SSMS optimizes the number of mixtures through splitting in mixture domain and the number of states through splitting in time domain. The recognition results show that the proposed SSMS method can reduce the total number of Gaussian up to 40.0% compared with the fixed parameter models at the same recognition performance in speech recognition system.

1. Introduction

With the growth of mobile information devices, interest in intelligent multimodal interfaces has significantly increased, particularly to provide a convenient user interface for small sized mobile devices, such as Personal Digital Assistants (PDAs). In fact, some customized PDAs already offer speech recognition and character recognition modalities [1].

So far, the inclusion of speech and character recognition in small mobile devices has employed two different engines. However, this recognition structure is inappropriate for small sized mobile devices in terms of memory management and cost. One solution is the use of a unified processor for both speech and character recognition modalities. Accordingly, the current paper focuses on a unified Speech and Character Combined Recognition System (SCCRS).

The Hidden Markov Model (HMM) is the most widely used technique in speech recognition and has been successfully applied to Korean on-line handwriting recognition [2]. Therefore, the proposed SCCRS employs HMM as the basic model structure for constructing both the speech and character recognition units, thereby facilitating application to memory-limited low-cost devices. In particular, a context-independent CHMM of a phoneme or grapheme (Korean character phone) is used as the basic recognition unit in the proposed SCCRS. The following conditions need to be satisfied for a CHMM-based SCCRS to be effectively applied to customized mobile devices; 1) The combined recognition system must maintain recognition accuracy in each individual system. 2) Real-time processing must be achieved. As such, the size of CHMM must be minimized for real-time processing.

Normally, CHMM has a fixed parameter model topology (i.e. fixed number of states and fixed number of mixtures). Yet this topology is unable to sufficiently represent a wide variety of distinctive feature parameters from an individual recognition unit. In the case of on-line character recognition, it is more effective to have a different number of states for different units - "ㄱ(g)" and "ㄹㅂ(rb)" have 2 and 6 states respectively[2]. Similar trials have also been conducted for speech recognition, and various approaches, such as a parameter histogram, AIC (AKAIKE Information Criterion)[3], and BIC (Bayesian Information Criterion)[4] have been reported to reduce the number of parameters with the smallest error rate. These approaches have a variable parameter model, which consists of a variable number of states and variable number of mixtures. However, since these approaches determine the number of states and mixtures for a recognition unit (phoneme or grapheme) without considering those of other units, this can decrease the recognition rate. Furthermore, as these approaches have the same number of mixtures for all recognition units, a recognition unit that has a compact distribution must also have a complicated structure, resulting in real-time processing difficulties.

Consequently, the current interest is focused on developing a method that selects a suitable number of states and suitable number of mixtures from each individual recognition unit. Thus, a splitting algorithm, Gaussian Output Probability Density Distribution (GOPDD), is employed to automatically decide the model topology. This algorithm is similar to Successive State Splitting (SSS)[5], which is often used in tied-states context-dependent models. Yet, the proposed method is different from SSS, as the GOPDD is split in the mixture domain, instead of in the context domain.

The remainder of this paper is organized as follows: The next section presents some conventional variable parameter models. Section 3 describes the proposed GOPDD splitting method. The recognition results for the proposed SCCR system are reported in Section 4, and some final conclusions are given in Section 5.

2. Conventional variable parameter models

CHMM has been widely used in speech and character recognition. Usual CHMM has a fixed parameter model topology (i.e. fixed number of states and fixed number of mixture models). However, this fixed parameter topology is unable to represent the distinctive features of individual recognition units.

Therefore, variable parameter model topology-based methods, such as ML (Maximum Likelihood), a parameter histogram, AIC, and BIC have been developed to reduce the number of parameters, while maintaining the recognition rate. AIC [3] evaluates the likelihood with a penalty term to reduce the number of parameters. BIC [4] is very similar to AIC, yet uses a penalty term including the number of training data. These two methods, categorized as an Information Criterion(IC), can identify a suitable number of states and mixtures for each recognition unit (phoneme/grapheme) and demonstrate that a variable parameter model topology can produce a better performance than a fixed parameter model topology. Fig. 1 shows an example of the variable parameter model topology generated by these methods.

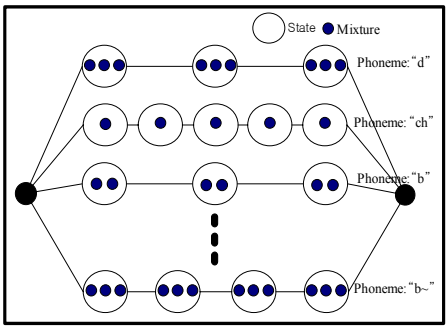

Figure 1. An example of variable parameter model topology

The rest of this section gives a brief review of other variable parameter topology selection methods with further details on the ML and BIC algorithms. These reviews will be helpful in understanding the proposed GOPDD splitting method.

2.1. Variable parameter model selection topology

In general, model selection is performed by choosing the topology $\hat{T}$ such that

$$\hat{T} = \arg\max_{T} P(T \mid X) = \arg\max_{T} p(T)P(X \mid T) \qquad (1)$$

A common practice in Bayesian model selection is to ignore the prior over the structure $P(T)$ (that is, assuming an equal prior across all topologies) and using the evidence $P(X \mid T)$ as the sole criterion for model selection such that

$$P(X \mid T) = \int p(X \mid T, \theta)p(\theta \mid T)d\theta \approx \log p(X \mid \theta_{ML}) - C(k, N) \qquad (2)$$

Where θ_{ML} is the estimated model using the MLE (maximum likelihood estimate) of θ. Note that (2) is written as the likelihood term and penalty term $C(k,N)$ [4], which depend on the number of training data N and number of parameter k, respectively.

2.2. ML topology selection method

The ML topology selection method identifies the model, θ^*, that maximizes the log likelihood, so as to determine a suitable number of states and number of mixtures for each recognition unit.

$$\theta^* = \max_{\hat{\theta}_i}\{\sum_{n=1}^{N} \log P(X_n \mid \hat{\theta}_i)\} \qquad (3)$$

Where $\hat{\theta}_i$ is the i-th model trained by the maximum likelihood estimate, and X_n is the n-th data, N is the size of the data set. Fig. 2 shows an example of a log likelihood. In the case of the Korean phoneme "aa", a 5 states and 4 mixtures model, denoted as S5_M4, exhibits the maximum likelihood over the interval of 3 ~ 6 states and 1 ~ 4 mixtures.

2.3. BIC topology selection method

The BIC is defined as the sum of the log likelihood and a penalty term. The penalty term depends on the number of model parameters and size of the data set. The BIC topology selection method identifies the model, θ^{**}, that maximizes the BIC, so as to determine a suitable number of states and number of mixtures for each recognition unit.

$$\theta^{**} = \max_{\hat{\theta}_i}\{\sum_{n=1}^{N} \log P(X_n \mid \hat{\theta}_i) - \frac{k_i}{2}\log N\} \qquad (4)$$

Where k_i is the number of the i-th model parameter and N is the size of the data set. Fig. 2 also shows that the S4_M3 model produces the maximum BIC (sum of the maximum likelihood and penalty term).

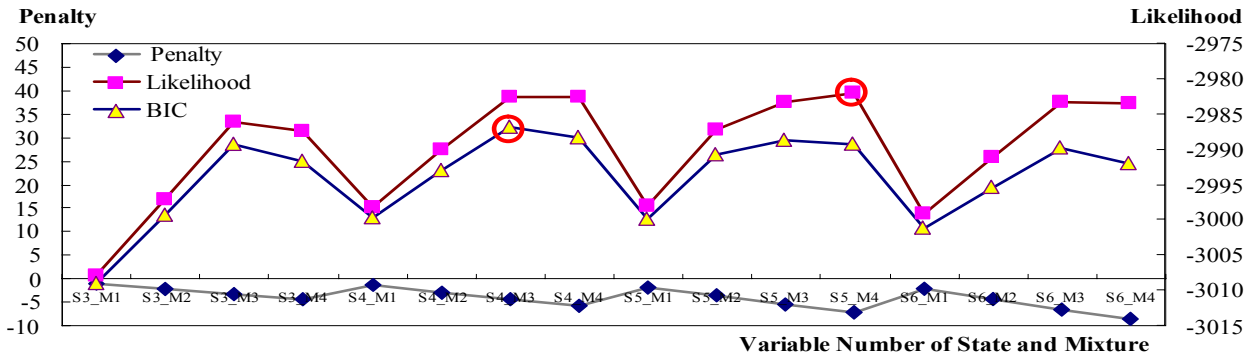

Figure 2. Log likelihood, penalty, and BIC of Korean phoneme "aa"

3. Successive State and Mixture Splitting

The acoustical characteristics of phonemes are significantly influenced by various factors, such as the phoneme context, speaker characteristics, and speaking rate of utterance. Many algorithms, such as SSS-FREE, ML (Maximum Likelihood)-SSS, DT (Decision Tree)-SSS[6] have already been proposed for constructing context-dependent models. In general, it is known that context-dependent models perform better than context-independent models, yet require much more memory. As such, the current paper uses a context-independent model for the proposed SCCRS, while also taking account of the limitations of low-cost memory-limited mobile devices.

A splitting algorithm, called Successive State and Mixture Splitting (SSMS), is proposed, which splits the GOPDD for a variable parameter context-independent model.

Unlike the SSS algorithm generating a context-dependent model, the SSMS algorithm constructs a context-independent model with a suitable number of states and mixtures for each recognition unit based on splitting the GOPDD. The SSS is performed within the time and context domains, while the SSMS splits the GOPDD within the time and mixture domains. An outline of the SSMS algorithm is illustrated in Fig. 3. The algorithm consists of three steps as follows:

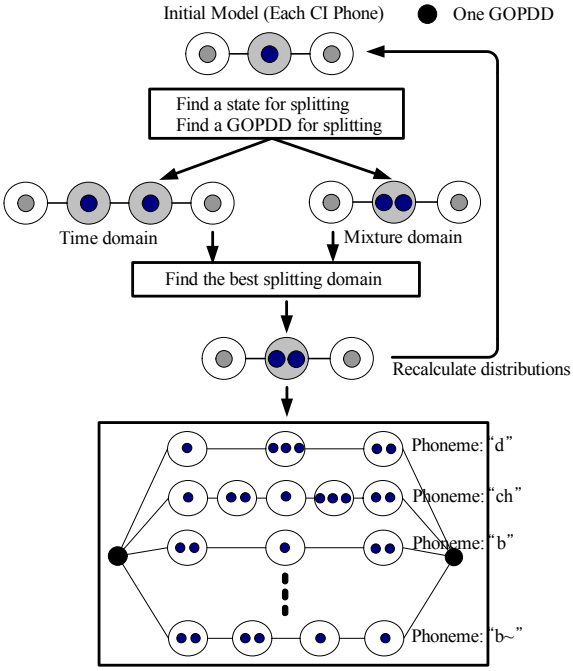

Figure 3. Generation of SSMS model

Step 1: Train initial models

In the proposed SCCRS, two different initial models are constructed for speech recognition and handwritten character recognition, respectively. For speech, an HMM with three context-independent states and one mixture is used as the initial model. For characters, two context-independent states and one mixture-based model is used, thereby representing a simple grapheme with the shortest length, for example "ㄱ(g)" or "ㄴ(n)".

Step 2: Identify GOPDD for splitting

For each state S(i) with an M-mixture GOPDD, calculate the normalized distribution size d_i. Let S(m) be the state to split that gives the maximum d_i:

$$d_i = \sum_{k}^{K} \frac{\sigma_{ik}^{\ 2}}{\sigma_{Tk}^{\ 2}} \cdot \sqrt{n_i} \qquad (5)$$

$$\sigma_{ik}^{\ 2} = \sum_{m}^{M} \lambda_{im} \sigma^2_{imk} + \sum_{m}^{M} \sum_{m'=m+1}^{M-1} \lambda_{im} \lambda_{im'} (\mu_{imk} - \mu_{im'k})^2$$

Where, K denotes the dimension of the feature vector, $\lambda_{im}\lambda_{im'}$ represent the weight coefficients, n_i denotes the number of training samples assigned to the state, and $\sigma_{ik}^{\ 2}$ denotes the k-th variance of all samples.

Step 3: Split the GOPDD.

The selected state in step 2 is split within the time and mixture domain, respectively. The Baum-Welch algorithm is applied to the split states in each domain to identify the maximum likelihood path. Fig. 4 shows a simple example of SSMS splitting. Where the large circle denotes one state and the small circle denotes the GOPDD in the corresponding state.

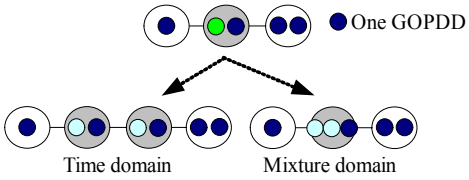

Figure 4. An example of splitting in time and mixture domains

In this example, the second state on the upper line is split by SSMS. The lower left corner shows that the state can be split into two states with the same number of mixtures. Meanwhile, in the lower right corner, the two mixtures in the state are split into three mixtures.

The original SSS algorithm splits the states in both the context and time domain, as described in Fig. 5. Note that all split states have one mixture.

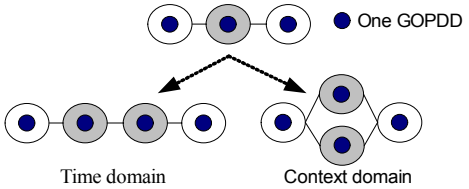

Figure 5. An example of splitting in time and context domains

Step2 through Step 3 are repeated until M reaches a specified number. As the model generated by the three steps of SSMS has a suitable number of states and each state has an appropriate number of mixtures, the proposed algorithm can generate a variable context-independent model. In addition, the proposed algorithm allows for more effective memory management in terms of the number of states and mixtures, compared with a fixed parameter model.

4. Experiments

The tasks were 452 Korean Phoneme Balanced Words (KPBWs) uttered by 38 males for a speaker-independent (SI) model for speech recognition, and on-line cursive characters handwritten by 10 writers for character recognition. Table 1 shows the analysis conditions for SCCRS.

Table 1. Analysis conditions for speech/character data

	Speech	On-line Character
Prepro cessing	8KHz Sampling, 16bits 16ms Hamming Window 5ms frame shift	100 samples/sec size/position norm. smoothing distance resampling
Feature	12 MFCCs 12 Delta MFCCs 12 Delta Delta MFCCs 1 Power, 1 Delta Power	2 Absolute X,Y position, 2 Directions 2 Curvatures 9 Modified bitmaps
DB	KLE Korean Words	KAIST Korean Written Characters
Model	M Mixture Variable Parameter CHMM	

To show the effectiveness of a variable parameter model using SSMS, it was compared with a conventional fixed parameter model and DT-SSS HM-Net[6]. Fig. 6 shows the SI word recognition rate with the fixed parameter model and variable parameter model using SSMS. The recognition accuracy increased as the GOPDD number increased. The dotted line indicates the recognition performance of the SSMS model, while the straight-line indicates the recognition performance of the fixed parameter model. The recognition of the SSMS model increased faster than that of the other models. The maximum recognition rate of the fixed parameter model was nearly 98.2%.

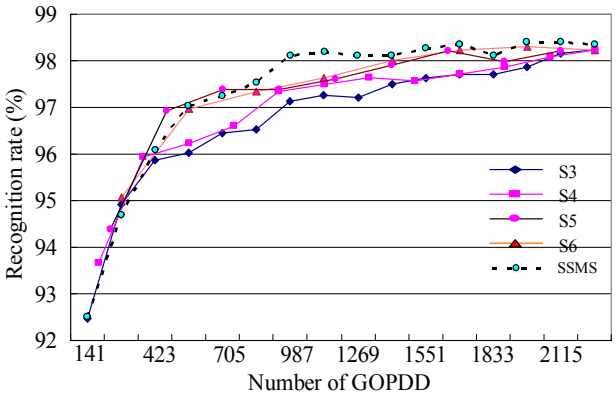

Figure 6. Comparison of recognition rate of SI fixed parameter model and SI variable parameter model using SSMS.(# GOPDD = # phone: 47 × # state: 3~6 × # mixture: 1~ 16)

Table 2. GOPDD number for each model with maximum recognition accuracy

Model	S3	S4	S5	S6	SSMS
#GOPDD	2115	2256	1645	1692	987

Table 2 shows the GOPDD number for each model with the maximum recognition accuracy. The recognition rates for all the models achieved the same maximum recognition rate (about 98.2%). However, the minimum GOPDD number was 1,692 for the fixed parameter model and 987 for the SSMS model. Therefore, the SSMS models had 40% fewer parameters than the fixed models.

Table 3. Decision-Tree-based SSS (#GOPDD)
(#M: number of Mixtures, #S: number of States)

#S \ #M	1	2	4
300	95.28 (300)	97.42 (600)	98.08 (1200)
1000	98.01 (1000)	98.67 (2000)	98.97 (4000)
2000	98.75 (2000)	98.75 (4000)	99.19 (8000)

Table 3 shows the recognition rates of the context-dependent model using a Decision Tree (DT)-based SSS. The results show that the context-dependent model provided a better recognition rate than the context-independent models. However, it should be noted that the DT-based context-dependent model required a GOPDD of more than 1,000 to achieve a recognition rate of 98.2%.

5. Conclusions

CHMM normally has a fixed parameter model topology (i.e. fixed number of states and fixed number of mixture models), yet is unable to sufficiently represent a wide variety of distinctive feature parameters within an individual recognition unit. Therefore, a variable parameter model is more effective at reducing the number of parameters, while maintaining the recognition rate.

Accordingly, the SSMS method was proposed to automatically generate a variable parameter model. The proposed method effectively reduces the number of mixtures through splitting in the mixture domain, instead of in the context domain. Experimental results indicated that the proposed model can achieve the same recognition rate as the best fixed parameter model, with only 60% of the parameters of the fixed model. Consequently, the proposed SSMS can be applied to compact mobile devices, such as PDAs.

6. References

[1] Suk, S.Y., Kim, M.J. and Chung, H.Y. "An on-line speech and character combined recognition system for multimodal interfaces", *EALPIIT Proc.*, 89-92, 2002.

[2] Sin, B. K. and Kim, J. "A Statistical Approach with HMMs for On-line Cursive Hangul(Korean Script) Recognition", *Second International Conference on Document Analysis and Recognition Proc.*, 147-150, Zuchuba, Japan, 1993.

[3] Tong, H. "Determination of hte order of a markov 0chain by Akaike`s information criterion", *Journal of Applied Probability*, 12:488-497, 1975.

[4] Li, D., Biem, A. and Subrahmonia, J. "Hmm topology optimization for handwriting recognition," *ICASSP Proc.,* 2001.

[5] Takami, J. and Sagayama, S., "A successive state splitting algorithm for efficient allophone modeling", *ICASSP-92 Proc.*, *Vol* 1, 573-576, 1992.

[6] Takaki, H., Mashahru, K., Akinori, I. and Masaki, K., "A Study on HM-Nets using Decision Tree-based Successive Splitting," Proc. ICSP-97, pp383-387, 1997.