A Successive State and Mixture Splitting for Optimizing the Size of Models in Speech Recognition

Soo-Young SUK† , Seong-Jun HAHM‡ , Ho-Youl JUNG‡ and Hyun-Yeol CHUNG‡

† Information Technology Research Institute, AIST
AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
‡ School of EECS, Yeungnam University
214-1, Dae-Dong, Gyungsan, Gyungbuk, Republic of Korea
†sy.suk@aist.go.jp, ‡branden65@yumail.ac.kr;{hoyoul, hychung}@yu.ac.kr

Abstract

A Successive State and Mixture Splitting (SSMS) algorithm for optimizing the size of models used in speech recognition for small size of mobile devices is proposed in this paper. The proposed algorithm employs essentially Continuous Hidden Markov Model (CHMM) structure and this CHMM consists of variable parameter topology in order to minimize the number of model parameters and to reduce recognition time. SSMS splits the Gaussian Output Probability Density Distribution (GOPDD) for variable parameter context independent model. Unlike the Successive State Splitting generating context dependent model, the algorithm constructs context independent model with suitable number of states and mixtures for each recognition units by automatic splitting of GOPDD in time and mixture domain. The recognition results showed that the proposed SSMS could reduce the total number of Gaussian up to 40.0% compared with the fixed parameter models at the same performance in speech recognition.

Index Terms: speech recognition, SSMS, mixture splitting

1. Introduction

There has been much interest in intelligent multimodal interfaces with the growth of mobile information devices. This is primarily motivated by the need for providing convenient user interface to small size of mobile devices such as Personal Digital Assistants (PDA). In some customized PDAs, speech recognition and character recognition modalities have already offered, so as to maximize convenient user interfaces [1].

The Hidden Markov Model (HMM) is the most widely used technique in speech recognition and Phoneme based Continuous HMM (CHMM) is used as a basic recognition unit for various speech recognition system. The following conditions should be satisfied for CHMM to be effectively applied to customize mobile devices; 1) The recognition system has to maintain the recognition accuracy in conventional system. 2) Real time processing should be achieved. So, the size of CHMM should be minimized.

Usual CHMM has a fixed parameter model topology (i.e. a fixed number of states and a fixed number of mixtures). But this topology has a problem that could not represent wide variety of distinctive feature parameters sufficiently in an individual recognition unit. To solve this problem and to reduce the number of parameters with small error rate, Several approaches such as parameter histogram, AKAIKE Information Criterion (AIC) [2], and Bayesian Information Criterion (BIC) [3] have been reported. These approaches have variable parameter model, which consist of variable number of states and mixtures, but determine the number of states and mixtures for a phoneme without considering those of other phonemes. This can cause to decrease the recognition rate. As these approaches have the same number of mixtures for all phonemes, a phoneme that has a compact distribution must also have a complicated structure and this can cause real time processing difficult.

Therefore, our main interest is focused on developing a method that selects both suitable number of states and suitable number of mixtures in each individual phoneme automatically. In this paper, a splitting algorithm of Gaussian Output Probability Density Distribution (GOPDD) is employed to automatically decide model topology. This algorithm is similar to Successive State Splitting (SSS) [4], which is often used in tied states context dependent models. But, our method is different from the SSS, as it splits the GOPDD in mixture domain, instead of in context domain.

This paper is organized as follows. The following section presents conventional variable parameters models and section 3 describes the proposed splitting method of GOPDD. Section 4 gives a brief review of system architecture with the preprocessing of speech recognition. Recognition results of the system are reported in Section 5. Finally, conclusions are given.

2. Conventional Variable Parameter Model

Usual CHMM has a fixed parameter model topology (i.e. a fixed number of states and mixtures). However, such a topology could not represent distinctive features for individual recognition units.

Therefore, variable parameter model topology based methods such as Maximum Likelihood, parameter histogram, AIC, and BIC have been developed to reduce the number of parameters while maintaining the recognition rate. AIC [2] evaluates the likelihood with the penalty term to reduce the number of parameters. BIC [3] is very similar to the AIC, but uses the penalty term including the number of training data. The two methods are categorized to Information Criterion.

Fig. 1 shows an example of variable parameter model topology generated by these methods. In these approaches, the authors showed that variable parameter model topology has better performance than the fixed parameter models [3][8].

From the Fig.1 we can see that the model topology can give different number of states for different phoneme but has the same number of mixtures in a model.

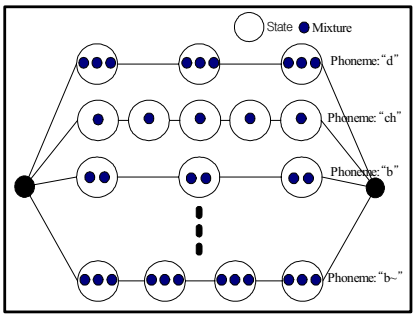

Figure 1 *An example of variable parameter model topology*

3. Successive State and Mixture Splitting

The acoustical characteristics of phonemes are greatly influenced by various factors, such as phoneme context, speaker characteristics, and the speaking rate of utterance. Many algorithms such as SSS-FREE, Maximum Likelihood SSS [9], Decision Tree SSS (DT-SSS) [5] have been proposed for constructing context dependent models. In general, it is known that the context dependent models perform better than the context independent models, but require much more memory. Taking account into low cost and memory limited mobile devices, context independent model is applied to the system in this paper.

Here, we propose a splitting algorithm, so called Successive State and Mixture Splitting (SSMS), which splits the GOPDD for variable parameter context independent model.

Unlike the SSS algorithm generating context dependent model, the SSMS algorithm constructs context independent model with suitable number of states and mixtures for each recognition units by splitting GOPDD. The SSS is done in time and context domains, while the SSMS splits the GOPDD in time and mixture domain. The outline of the SSMS algorithm is illustrated in Fig. 2. The algorithm consists of three steps as follows.

Step 1: Training of initial models

In the system, initial model should be constructed for recognition. 48 phonemes of context-independent HMM with three-state and one-mixture are used as the initial model.

Step 2: Find GOPDD for splitting

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

$$d_i = \sum_{k}^{K} \frac{\sigma_{ik}^2}{\sigma_{Tk}^2} \cdot \sqrt{n_i} \qquad (1)$$

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

Where, K denotes the dimension of the feature vector, $\lambda_{im}\lambda_{im'}$ represent weight coefficients, n_i denotes the number of training sample assigned to the state, σ_{ik}^2 denotes the k-th variance of all samples. A state is selected with maximum distribution value in time domain splitting, and with maximum weight value among M-mixture GOPDD in mixture domain splitting.

Step 3: Split of GOPDD.

The selected state in step 2 is split again in time and mixture domain respectively. The embedded training via forced alignment is applied to select maximum likelihood splitting domain because of changing the model's structure.

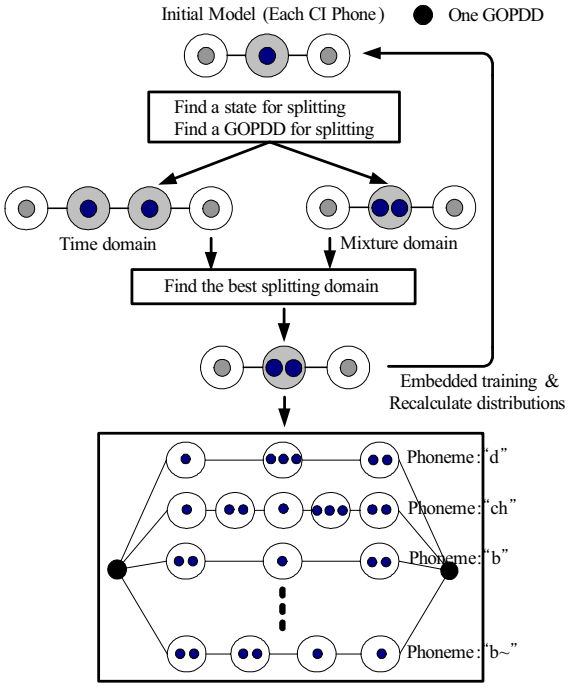

Figure 2 *Generation of a SSMS model*

Fig. 3 shows a simple splitting example of SSMS. Where the large circle denotes one state and small circle denotes one GOPDD in corresponding state.

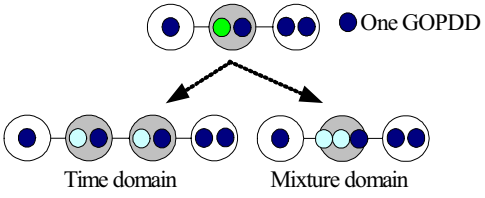

Figure 3 *The splitting examples in time and mixture domain*

In this example, the second state on 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. In the lower

right corner, two mixtures in the state can be split into three mixtures.

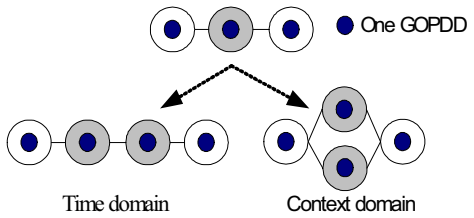

Figure 4 *The splitting examples in time and context domain*

The original SSS algorithm split the states in both context and time domain as illustrated in Fig. 4. Note that all split states have one mixture. Step2 through Step 3 are repeated until M reaches the specified number. As the model generated by the three steps of SSMS has suitable number of states and each state has appropriate number of mixtures, the proposed algorithm can be regarded as to be more general for generating variable context independent model. In addition, this algorithm allows more effective memory managements, in terms of the number of states and mixtures, than the fixed parameter model.

4. Speech Recognition System

Fig. 5 shows the recognition system architecture for working on PDA or on small size of mobile devices.

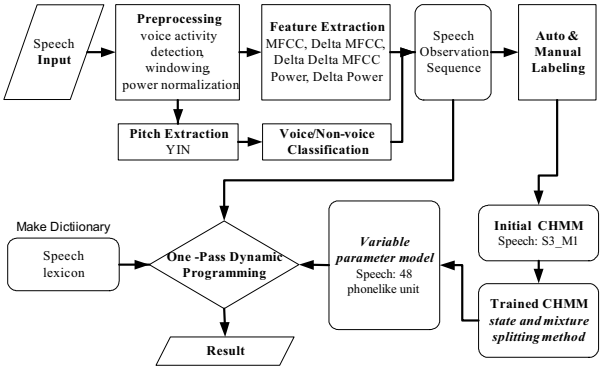

Figure 5 *System architecture*

In this system, speech data are taken through microphone on the devices and 39-order of Mel Frequency Cepstral Coefficient (MFCC) is extracted for speech recognition, where cepstral mean normalization is adopted for its use in the noisy environment. The voice activity detection, which rejects non-voiced input in pre-processing, is also adopted for realizing a highly reliable system. The detection is carried out by using the ratio of a reliable fundamental frequency contour of the whole input interval. Total of 48 CHMM models are trained through labeling. The recognition is performed by using one pass dynamic programming algorithm [6].

5. Experiments

The tasks are 452 Korean phoneme balanced words uttered by 38 male for Speaker Independent (SI) model for recognition. Table 1 shows the analysis conditions for the system. To show the effectiveness of variable parameter model using SSMS, we compare it with conventional fixed parameter model and DT-SSS [5]. Fig. 8 shows SI word recognition rate with fixed parameter model and variable parameter model using SSMS.

Table 1. *Analysis condition*

Preprocessing	8KHz sampling, 16bits 16ms hamming window 5ms frame shift
Feature	12 MFCCs,12 delta MFCCs, 12 delta delta MFCCs 1 power, 1 delta power, 1 delta delta power
DB	KLE Korean Words
Model	M mixture variable parameter CHMM

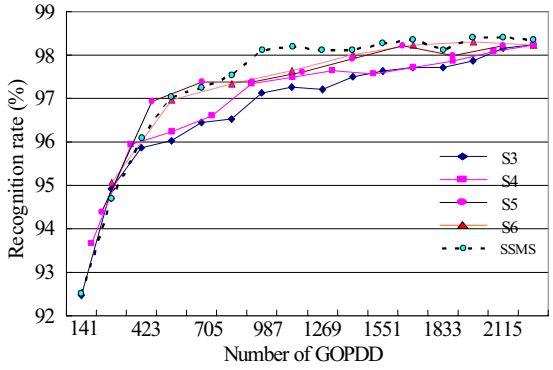

Figure 6 *Comparison of recognition rates between SI fixed parameter models from 3state (S3) to 6state (S6) and SI variable parameter model by SSMS (Where, no. of GOPDD= no. of phones x no. of states x no. of mixtures*

We can see from the result that the recognition accuracy increases as the number of GOPDD increases. The dotted line indicates the recognition performance by SSMS model, and straight-line indicates the recognition performance by fixed parameter models having number of states from 3 to 6. The recognition rate by SSMS model increases faster than other fixed models to the maximum recognition rate of 98.2%.

Table 2. *Number of GOPDD of each model reaching to the maximum recognition accuracy of 98.2%*

Model	S3	S4	S5	S6	BIC	SSMS
#GOPDD	2115	2256	1645	1692	2131	987

Table 2 shows the number of GOPDD of each model that reaches to the maximum recognition accuracy of 98.2%. In this table, we can find that the number of GOPDD is 1,692 for the fixed parameter model and 987 for SSMS model to reach 98.2% of recognition rate. Therefore, SSMS can reduce models up to 40% than the usual fixed parameter topology. Table 3 shows examples of model topology by SSMS that achieves maximum

recognition results. In case of phoneme "g," the number of state is 5 and the first state has 4 mixtures, the second four, the third 7…etc.

Table 4 show the recognition rates by context dependent model using DT-SSS. The results show that the context dependent model provides better recognition rate than the context independent models. In case of a model which have 300-state and 4-mixture, DT-SSS needs 24.7 MBytes of memory for running but SSMS 2.8Mbytes, meaning DT-SSS needs much processing time as compared with SSMS. So it can be said that SSMS will help reduce the size of models for recognition engine running on the small devices.

Note that the decision tree based context dependent model requires, however, more than 1,000 of GOPDD, to achieve the recognition rate of 98.2%.

Table 3. *Examples of model topology by SSMS model*

Phone	# Total state	1	2	3	4	5	6
g	5	4	4	7	5	6	-
gg	6	7	4	4	3	3	2
aa	3	3	7	8	-	-	-
ih	3	4	9	11	-	-	-

Table 4. *Recognition rates by decision tree based context dependent model (#GOPDD)*

		Number of mixture		
		1	2	4
Number of state	300	95.28 (300)	97.42 (600)	98.08 (1200)
	600	97.49 (600)	98.20 (1200)	98.71 (2400)
	1000	98.01 (1000)	98.67 (2000)	98.97 (4000)
	2000	98.75 (2000)	98.75 (4000)	99.19 (8000)

6. Conclusions

Usual CHMM has a fixed parameter model topology (i.e. a fixed number of states and a fixed number of mixture models), but can not represent wide variety of distinctive feature parameters sufficiently in an individual recognition unit. Therefore, it would be better for small mobile devices to have variable parameter models to reduce the number of parameters while maintaining the recognition rate.

SSMS method was proposed for generating the variable parameter model automatically. The proposed SSMS allows reducing effectively the number of mixtures through splitting in mixture domain instead of in context domain. The experimental results indicate that the proposed SSMS can save the number of models up to 40% while maintaining the best recognition accuracy of the fixed model. This means that the proposed SSMS enables to be applied to compact mobile devices such as PDA.

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

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

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

[5] Takaki, H., Mashahru, K., Akinori, I. and Masaki, K., "A Study on HM-Nets using Decision Tree-based Successive Splitting," *ICSP-97 Proc.,* 383-387, 1997.

[6] Nakagawa, S., "A connected spoken word recognition method by O(n) dynamic programming pattern matching algorithm," *ICASSP Proc*., 296-299, 1983.

[7] Ralph, G., Stefan, M. and Alex, W., *"Run-on recognition in an on-line handwriting recognition system,"* Carnegie Mellon Univ. Press., 1997.

[8] Biem, A., Ha, J.Y. and Subrahmonia, J., "A Bayesian model selection criterion For HMM Topology Optimization," *ICASSP Proc.,* 989-992, 2002.

[9] Singer, H., Ostendorf, M. ,"Maximum likelihood successive state splitting," *ICASSP Proc*., Vol 2, 601-604, 1996.