A Speech and Character Combined Recognition Engine for Mobile Devices

Min-Joung KIM[1], Soo-Young SUK[2], Ho-Youl JUNG[3], Hyun-Yeol CHUNG[3]

[1] School of EECS, Yeungnam University
214-1, Dae-Dong, Gyung-San, Gyungbuk, Republic of Korea
manjuk@paran.com
[2] Information Technology Research Institute, AIST
AIST Tsukuba Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
sy.suk@aist.go.jp
[3] School of EECS, Yeungnam University
214-1, Dae-Dong, Gyung-San, Gyungbuk, Republic of Korea
{hoyoul, hychung}@yu.ac.kr

Abstract. A Speech and Character Combined Recognition Engine (SCCRE) is developed for working on Personal Digital Assistants (PDA) or on mobile devices. In SCCRE, feature extraction from speech and 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 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. When we applied our developed engine which adopts SSMS to speech recognition for mobile devices, SSMS can reduce total number of Gaussian up to 40.0% compared with the fixed parameter models at the same recognition performance. This leads that SSMS can reduce the size of memory for models to 65% and that for processing to 82%. Moreover, recognition time decreases 17% with SSMS model but still maintains the recognition rate.

Keywords: Speech, Character, Recognition, SSS, Embedded

1 Introduction

There has been much interest in intelligent multimodal interfaces with the growth of mobile information devices. This is primarily motivated by a need for providing convenient user interface to small size of mobile devices such as PDA. In some customized PDAs, speech recognition and character recognition modalities have already offered, so as to maximize convenient user interfaces [1]. Such small mobile devices have employed two different engines for speech and character recognition so far. However, this recognition structure is not desirable for small size of mobile devices in

terms of memory management and cost. One solution is to use of a unified processor for both speech and character recognition modalities. In this paper, our interest is focused on the SCCRE.

Hidden Markov Model (HMM) is the most widely technique used in speech recognition and it has been successfully applied in the recognition of Korean on-line handwritings [2]. Therefore, our SCCRE employs HMM as a basic model structure for construction of both speech and character recognition units, so as to be effectively applied to memory limited low cost devices. Especially, context independent CHMM of phoneme or grapheme (Korean character phone) is used as a basic recognition unit in SCCRE. The following conditions should be satisfied for CHMM based SCCRE to be effectively applied to customize mobile devices; 1) Combined recognition engine has to maintain recognition accuracy as in each individual system. 2) Real time processing should be achieved. For these reason, the size of CHMM should be minimized for real time processing.

Usual CHMM has a fixed parameter model topology (i.e. a fixed number of states and a fixed number of mixtures). But this topology can not represent wide variety of distinctive features sufficiently in an individual recognition unit. In case of on-line character recognition, it is more effective to have a different number of states for the different units, for example "ㄱ(g)" and "뤄(rb)" have 2 and 6 states respectively[2]. For speech recognition, there have been similar trials. Several approaches such as parameter histogram, AIC (AKAIKE Information Criterion)[3], and BIC (Bayesian Information Criterion)[4] [10] have been proposed to reduce the number of parameters with the smallest error rate. These approaches have variable parameter model which consists of variable number of states and variable number of mixtures. However, these approaches determine the number of states and mixtures for a recognition unit (phoneme or grapheme) without considering those of other units. This can lead to decrease the recognition rate. 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 and this can cause real time processing difficult.

Therefore, our main interest is focused on developing a method that selects a suitable number of states and a suitable number of mixtures in each individual recognition unit. In this work, a splitting algorithm of GOPDD (Gaussian Output Probability Density Distribution) is employed to decide model topology automatically. This algorithm is similar to SSS (Successive State Splitting)[5], 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 not in context domain.

This paper is organized as follows. The following section gives a brief review of SCCRE architecture with the preprocessing of speech and on-line character recognition. Section 3 presents the previous variable parameters models. Section 4 describes the proposed splitting method of GOPDD. Recognition results for SCCRE are given in Section 5. Finally, some conclusions are presented in Section 6.

2 Speech and Character Combined Recognition Engine

2.1 System architecture

Fig. 1 shows combined recognition system architecture for working on PDA or on mobile devices. In this system, we assume that the character (cursive script) and speech inputs are taken through touch screen and microphone, respectively. The preprocessing and the feature extraction are carried out on each modality, but provide the parameter observations to CHMM based combined recognition engine. Total 115 M-mixture CHMM models are trained through labeling. The combined CHMM models consist of 48 phone-like units for speech and 67 graphemes for character. The recognition is performed by using OPDP (One Pass Dynamic Programming) algorithm [7].

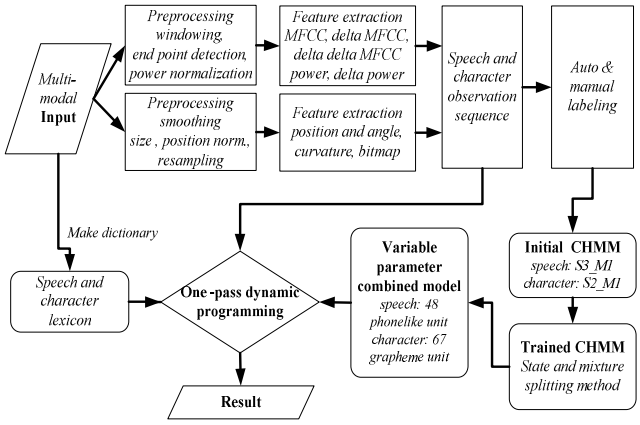

Fig. 1. System architecture

2.2 Preprocessing

39-order of MFCC (Mel Frequency Cepstral Coefficient) is extracted for speech recognition, where CMN (Cepstral Mean Normalization) is applied for taking account into environmental noise. The features for character recognition consist of 6-order of position parameters and 9-order of bitmap parameters. In the rest of this sub-section, we describe about preprocessing of Korean character briefly.

Fig.2 shows the preprocessing steps of on-line character recognition. Assume that the handwriting inputs are obtained through touch screen and the sampling rate is set to higher than 100 sample per second for different devices. In order to make users freely in writing style and writing position, the preprocessing step is given as follows:

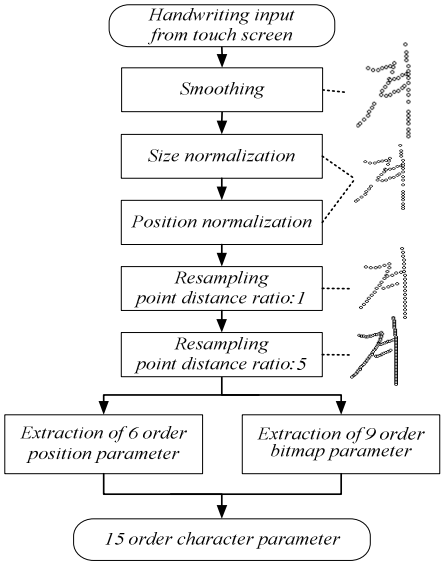

Fig. 2. Preprocessing of on-line character recognition

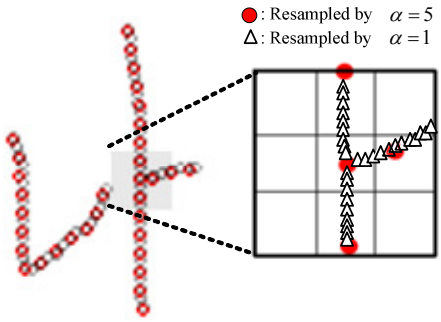

Fig. 3. Extraction of bitmap parameter

A) Smoothing

Smoothing is carried out on the pen trajectory of the input character to reduce noises from input device. A smoothed point $\hat{x}_i$ is obtained by convolution, as given by

$$\hat{x}_i = C_{-n} \cdot x_{i-n} + \ldots + C_0 \cdot x_i + \ldots + C_m \cdot x_{i+m} \quad (1)$$

Where C_j, $for\ j = -n, -(n-1), \Lambda\ , m$ denotes the impulse responses of the smoothing filter, and $n + m + 1$ denotes the size of smoothing filter.

B) Normalization

To rule out the variation of writing styles, width, height, and starting positions of input words should be normalized to the reference geometry. In word-based recognition system, the pre-determined height for all input words is used for the normalization of width and height. In other words, the height of input words are scaled to the reference height and width of the words are adjusted with the same scaling factor used in the height normalization. The starting position is also adjusted into the fixed position to remove the variations of each input words.

C) Re-sampling

Sampled data are re-sampled to yield a new sequence of data having equidistant in space to compensate different sampling rate and different writing speed. The following equidistant re-sampling procedure is applied. The coordinate of a new sample (px_j , py_i) is obtained by bi-linear interpolation as follows.

$$px_i = \alpha \frac{(px_j - px_{j-1})}{\sqrt{(px_{j-1} - x_j)^2 + (py_{j-1} - y_j)^2}} + px_{j-1}$$
$$py_i = \alpha \frac{(py_j - py_{j-1})}{\sqrt{(px_{j-1} - x_j)^2 + (py_{j-1} - y_j)^2}} + py_{j-1} \quad (2)$$

Where α denotes the desirable distance in spaces.

D) Feature extraction

For each data point of the re-sampled sequence, a 15 dimensional feature vector is calculated, which consists of 2 local information features for absolute x and y positions, 2 local angle parameters, 2 curvature parameters, and 9 bitmap based global information features [8]. Fig. 3 shows an example of bitmap parameter extraction. The circles and the triangles denote the re-sampled sequences in case of α=5 and α=1. Distribution of points (re-sampled with α=1) within 3×3 window is used for extracting bitmap parameter as show in this figure.

3. Conventional Variable Parameter Model

This chapter gives brief reviews of general variable parameter model topology selection methods with more details of ML and BIC algorithms.

3.1 Variable parameter model selection topology

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

$$\hat{T} = \arg\max_T P(T \mid X) = \arg\max_T p(T)P(X \mid T) \quad (3)$$

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

$$P(X \mid T) = \int p(X \mid T, \theta) p(\theta \mid T) d\theta \approx \log p(X \mid \theta_{ML}) - C(k, N) \quad (4)$$

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

3.2 ML topology selection method

ML topology selection method is to find the model, θ^*, that maximizes the log likelihood, so as to determine a suitable number of state and number of mixtures at each recognition unit.

$$\theta^* = \max_{\hat{\theta}_i} \{ \sum_{n=1}^{N} \log P(X_n \mid \hat{\theta}_i) \} \tag{5}$$

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

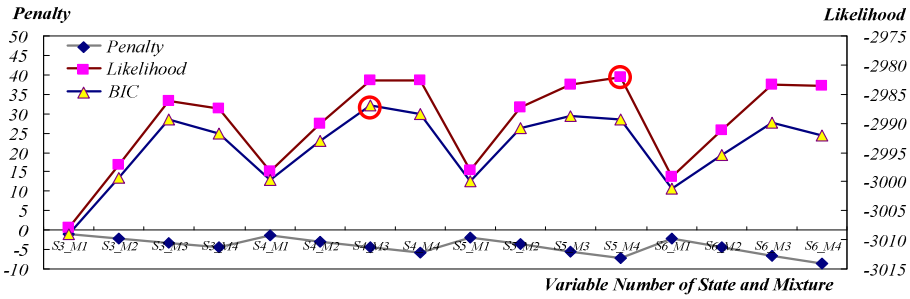

Fig. 4. Log likelihood, penalty and BIC of Korean phoneme "aa"

3.3 BIC topology selection method

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

$$\theta^{**} = \max_{\hat{\theta}_i} \{ \sum_{n=1}^{N} \log P(X_n \mid \hat{\theta}_i) - \frac{k_i}{2} \log N \} \tag{6}$$

Where k_i is the number of i-th model parameter and N is the size of the data set. Fig. 4 also shows that the S4_M3 model has the maximum BIC (sum of maximum likelihood and the penalty term).

4. Successive State and Mixture Splitting

The acoustical characteristics of phonemes are greatly influenced by phoneme context, speaker characteristics, and the speaking rate of utterance. Many algorithms such as SSS-FREE, ML-SSS[10], DT(Decision Tree)-SSS[6] have been proposed for constructing context dependent models. Generally, it is known that the context dependent models perform better than the context independent models, but require much more memory for processing. Taking account into low cost, memory limited mobile device, context independent model is applied to SCCRE in this paper.

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

Unlike the SSS algorithm generating context dependent model, the SSMS constructs context independent models 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 is illustrated in Fig. 5. The algorithm consists of three steps as follows.

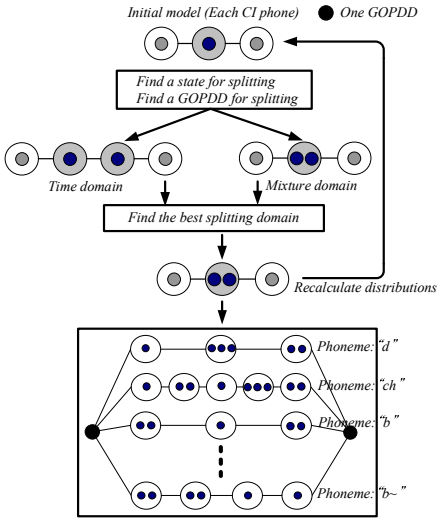

Fig. 5. Generation of a SSMS model

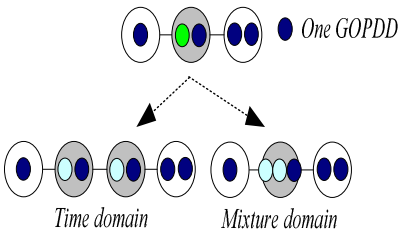

Fig. 6. The splitting examples in time and mixture domain

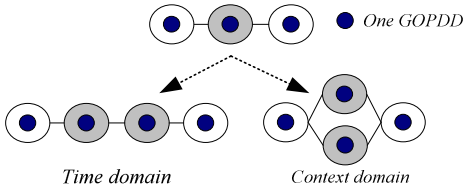

Fig. 7. The splitting examples in time and context domain

Step 1: Train initial models

Two different initial models should be constructed for speech and handwritten character, respectively. For speech, HMM with context-independent three states and one mixture is used as the initial model and context-independent two states and one mixture is used for character, so as to well represent up to the simple grapheme with the shortest length.

Step 2: Find GOPDD for splitting

For each state S(i) with M-mixture GOPDD, the normalized distribution size d_i is calculated. S(m) will be the state to split which gives the maximum d_i among all samples.

$$d_i = \sum_{k}^{K} \frac{\sigma_{ik}^{\ 2}}{\sigma_{Tk}^{\ 2}} \cdot \sqrt{n_i}$$
$$\sigma_{ik}^{\ 2} = \sum_{m}^{M} \lambda_{im} \sigma^2_{imk} + \sum_{m}^{M} \sum_{m'=m+1}^{M-1} \lambda_{im} \lambda_{im'} (\mu_{imk} - \mu_{im'k})^2 \quad (7)$$

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, $\sigma_{ik}^{\ 2}$ denotes the k-th variance of all samples.

Step 3: Split the GOPDD.
The selected state in step 2 is then split in time and mixture domain respectively. The Baum-Welch algorithm is applied to the split states in each domain in order to find maximum likelihood path. Fig. 6 shows a simple splitting example of SSMS. Where the large circle denotes one state and small circle denotes one GOPDD in corresponding state. 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. The original SSS algorithm split the states in both context and time domain as described in Fig. 7. Note that all split states have one mixture.

Step2 through Step 3 are repeated until M-mixture reaches the specified number. Because 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.

5. Experiments

5.1 Performance tests

452 KPBWs (Korean Phoneme Balanced Words) uttered by 38 male are used for constructing SI (speaker independent) model in speech recognition, and on-line cursive handwritten characters by 10 writers for character recognition. Table 1 shows the analysis conditions of SCCRE.

To show the effectiveness of variable parameter model using SSMS, we compare it with conventional fixed parameter model and DT-SSS HM-Nets (Hidden Markov Network)[6]. SI word recognition rate with fixed parameter model and variable pa-

rameter model using SSMS is shown in Fig. 8. As the figure indicates, 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 1. Analysis conditions for speech/character data

	Speech	Character
Preprocessing	8kHz sampling, 16bits 16ms hamming window 5ms frame shift	100 samples/sec smoothing size/position normalization distance re-sampling
Feature	12 MFCCs 12 delta MFCCs 12 delta delta MFCCs 1 power, 1 delta power	2 absolute X,Y positions 2 angles 2 curvatures 9 modified bitmaps
DB	KLE Korean words	KAIST Korean written characters
Model	M mixture variable parameter CHMM	

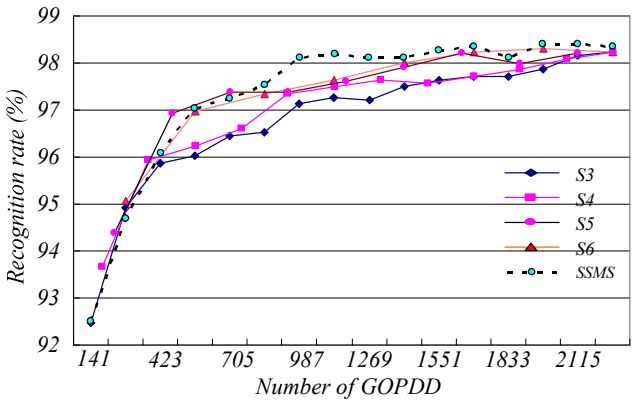

Fig. 8. Comparison of recognition rates between SI fixed parameter models from 3 states(*S3*) to 6 states(*S6*) and SI variable parameter model by SSMS(Where, no. of GOPDD= no. of phones *x* no. of states *x* no. of mixtures)

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

Model	S3	S4	S5	S6	SSMS
#GOPDD	2115	2256	1645	1692	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 models have 40% fewer parameters than the fixed model. For running in PDA, fixed parameter model has size of 630Kbyte and requires a memory of 3.42Mbyte; SSMS model has size of 410Kbyte and requires a memory of 2.81Mbyte, leading that SSMS can reduce the size of memory for models to 65% and that for processing to 82%. Moreover, recognition time decreases 17% with SSMS model but still maintains the recognition rate.. Moreover, recognition time decrease 17% with SSMS model but still maintain the recognition rate. 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 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. DT based SSS (#GOPDD) (#M: number of Mixture, #S: number of State)

#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 4 show the recognition rates of the context dependent model using DT based SSS. The results show that the context dependent model provides better recognition rate than the context independent models. Note that the DT based context dependent model requires, however, more than 1,000 of GOPDD, to achieve the recognition rate of 98.2%.

6. Conclusions

This paper describes an on-line SCCRE working on PDA or on mobile devices. In the SCCRE, feature extraction for speech and for character is carried out separately, but recognition is performed in an engine.

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 if variable parameter model is used to reduce the number of parameters while maintaining the recognition rate.

SSMS method was proposed for generating the variable parameter model automatically. The proposed allows reducing effectively the number of mixtures through splitting in mixture domain instead of in context domain. The experimental results indicate that the proposed model have the same recognition rate of the best fixed parameter model, with only 60% parameters of the fixed model. This leads that SSMS can reduce the size of memory for models to 65% and that for processing to 82%. Moreover, recognition time decreases 17% with SSMS model but still maintains the recognition rate. This means that the proposed SSMS is suitable for applying in compact mobile devices such as PDA.

Acknowledgement

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

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., (2002) 89-92
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., Zuchuba, Japan (1993) 147-150
3. Tong, H.: Determination of the order of a markov chain by Akaike`s information criterion, Journal of Applied Probability, 12, (1975) 488-497
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. (1992) 573-576
6. Takaki, H., Mashahru, K., Akinori, I. and Masaki, K.: A Study on HM-Nets using Decision Tree-based Successive Splitting, ICSP-97 Proc. (1997) 383-387
7. Nakagawa, S.: A connected spoken word recognition method by O(n) dynamic programming pattern matching algorithm, ICASSP Proc. (1983) 296-299
8. Ralph, G., Stefan, M. and Alex, W.: Run-on recognition in an on-line handwriting recognition system, Carnegie Mellon Univ. Press. (1997)
9. Biem, A., Ha, J.Y. and Subrahmonia, J.: A Bayesian model selection criterion For HMM Topology Optimization, ICASSP Proc. (2002) 989-992
10. Jitsuhiro, T., Nakamura, S.: Automatic generation of non-uniform HMM structures based on variational Bayesian approach, ICASSP Proc., Vol. 1. (2004) 805-808
11. Li Deng.: Distributed speech processing in MiPad's multimodal user interface, IEEE Trans. Speech and Audio., Vol. 10, No. 8. (2002) 605-619