# SMOOTHING AND TYING FOR KOREAN FLEXIBLE VOCABULARY ISOLATED WORD RECOGNITION

*Jae-Seung Choi°, Jong-Seok Lee, Hee-Youn Lee*
*Email : {seung111,ljs,hylee}@lgcit.com*

Information Technology Lab. MI group
LG Corporate Institute of Technology

## ABSTRACT

For large vocabulary recognition system, as well as for flexible vocabulary applications using hidden Markov model(HMM), parameter smoothing and tying have been used to increase the reliability of models. This paper describes bottom-up and top-down clustering techniques for state level tying. This paper also describes a method of applying parameter smoothing to the clustered states and covariance matrix of semicontinuous hidden Markov model(SCHMM). We present a new parameter smoothing method and apply it to the distribution of discrete hidden Markov model(DHMM) in the training procedure. A new model composition method for unseen triphone modeling in bottom-up clustering is also proposed and compared with traditional context-independent model backing-off method.

## 1. INTRODUCTION

It is necessary to use subword unit modeling for large vocabulary recognition system, as well as for flexible vocabulary applications. Context dependent phones, like triphones, are generally used for taking into account the co-articulation effect. As the number of units becomes larger by including more context dependencies, the amount of data available for each unit decreases and the model estimates become less reliable. Since we will never have sufficient training data to model the large amount of parameters, several techniques to increase the reliability of models were used.

Several solutions which have been proposed to create robust models can be summarized into two main classes, which are parameter smoothing and tying. Parameter smoothing methods include co-occurrence smoothing[1] based on joint probabilities of pairs of codebook symbols, interpolation of detailed context dependent models with less detailed but better trained models, and covariance matrix smoothing[2].

Parameter tying[3][4] by clustering similar units or similar distributions is also widely used for creating reliable units. Although any subsets of HMM parameters can be tied, parameter tying is generally applied in two levels which are state level and mixture level. Traditional method of dealing with parameter tying tend to be model based, but since model based approach cannot treat the left and right context independently, state based tying is preferred.

For state level tying, bottom-up and top-down clustering method were used. Although, bottom-up clustering approach is more flexible in considering all possible configurations for seen context, for unseen context modeling, top-down clustering method using decision tree is more appropriate. In bottom-up clustering method, context independent models are used to back off unseen context, which degrades system performance. In mixture level tying, the SCHMM which can be considered as a special form of continuous mixture hidden Markov model with the continuous output probability density functions sharing in a mixture Gaussian density codebook is generally used.

The organization of this paper is as follows. Section 2 describes parameter smoothing methods. Section 3 explains bottom-up and top-down state clustering. Finally, in Section 4, experiments and experimental results are described.

## 2. SMOOTHING

### 2.1 Training Procedure

Training procedure is composed of two stages. In the first stage, 52 initial context-independent DHMM phone models from hand-labeled data are created. These phone models are used to initialize context-independent DHMM and context-dependent models are estimated based on context-independent ones.

Bottom-up and top-down clustering methods are applied to the context dependent DHMM for reducing the number of states. The second stage is to train 4-codebook context-independent SCHMM using mapping table and context-independent DHMM in the first stage.

### 2.2 Senone Smoothing

The co-occurrence smoothing method(CSM)[1] is the method to smooth distributions by calculating co-occurrence probability which represents the similarity measure among all codewords.

When we apply the CSM to the senones[3], that is, shared-distributions, the co-occurrence probability of codeword *i* given codeword *j* is shown in Eq (1). This probability represent the similarity between codeword probability *i* and *j*.

In Eq(1), *p(k|p,d)* is the output probability of codeword *k* for distribution *d* in phoneme model *p* and $W_d$ represent the probability of each distribution, that is, the reliability of each distributions.

$$p(i|j) = \frac{\sum_{d=1}^{NS} p(i|d)p(j|d) \cdot W_d}{\sum_{k=1}^{NC}\sum_{d=1}^{NS} p(k|d)p(j|d) \cdot W_d} \quad (1)$$

where *NS* is the number of senone, and *NC* is the number of codeword.

We used the occupation counts during training as $W_d$. For each distribution, we calculate smoothed distribution by multiplying smoothing matrix. Finally, we interpolate original distribution and smoothed distribution.

## 2.3 Clustered Co-occurrence Smoothing Method

Traditionally, CSM has overall one smoothing matrix or smoothing matrix per phoneme. But this has a disadvantage of over-smoothing.

We propose clustered co-occurrence smoothing method and applied it to the distributions of DHMM. The proposed method has four steps. First we cluster similar states. Second, For each cluster of distributions, one smoothing matrix is calculated. The co-occurrence probability of *m*th cluster is shown in Eq (2).

$$p_m(i|j) = \frac{\sum_{d=1}^{ND(m)} p(i|d)p(j|d) \cdot W_d}{\sum_{k=1}^{NC}\sum_{d=1}^{ND(m)} p(k|d)p(j|d) \cdot W_d}, \quad 1 \le m \le N \quad (2)$$

where *N* is the number of cluster, and *ND(m)* is the number of distributions which belongs to the *m*th cluster.

Third, For each distribution, we determine to which cluster it belongs. And then calculate smoothed distribution by multiplying smoothing matrix. Finally, we interpolate original distribution and smoothed distribution.

## 2.4 Covariance Smoothing

We used covariance smoothing[2] to obtain more powerful acoustic models. A Gaussian density with a very sharp peak gives very low likelihood scores to feature points which are only slightly deviated from the mean and hence the model robustness is low.

Let there be *M* components in a mixture density with the covariance matrices $C_i$ , *i = 1,2,…,M*. The relative sharpness of the *i* th component $R_i$ is calculated as

$$R_i = \frac{1/|C_i|^{1/2}}{1/\left(\prod_{k=1}^{M}|C_k|^{1/2}\right)^{1/M}} \quad (3)$$

Let the covariance matrix of the unimodal density be *C*, then a smoothing is performed on $C_i$ if $R_i$ is above a threshold.

The smoothing is defined as a linear interpolation of the detailed estimate $C_i$ and the robust estimate *C*. The smoothed covariance matrix be $\widetilde{C}_j$

$$\widetilde{C}_i = \lambda C_i + (1-\lambda)C \quad (4)$$

where $0 \le \lambda \le 1$ is the interpolation parameter.

## 2.5 Incorporation Smoothing into Training Procedure

We applied distribution smoothing to the distributions of context-independent initial DHMM, context-dependent DHMM and context-dependent SCHMM in the training procedure. We also applied covariance smoothing to the covariances of SCHMM.

# 3. TYING

## 3.1 Bottom-up Clustering

We applied bottom-up[3] and top-down[4] clustering in state level tying. For both cases, we used the entropy reduction weighted by occurrence counts during training as the criterion for clustering. Eq (5) represents the criterion used.

$$\Delta\widetilde{H} = (P+Q)H(A+B) - PH(A) - QH(B) \quad (5)$$

where *P* is the summation of the count entries in probability distribution *A* and similarly, *Q* is the summation of the occurrence counts in *B*.

We apply four constraints in the bottom-up clustering for efficient and robust clustering. First, we prohibit HMM output distributions of different phones from being clustered. Second, we allow HMM output distributions to be merged only if they are associated with the same k-th Markov state in the model topology. Third, the central state of an allophone is tied to the central state of all the other allophones of the same phone. Fourth, we allow HMM output distributions to be merged only if occurrence counts during training is below the threshold.

In bottom-up clustering, the common method of the handling unseen triphone is backing-off to less specific models like monophones. For example, if we assume that the unseen triphone is g_a_n, by Backing-off method, it is replaced by monophone, $_a_$. Recently, unseen triphone modeling using diphone has been also proposed[5].

we present modified model composition method(MMCM) . In this method, each state of unseen triphone is tied with the similar state among trained states. Similarity is measured just simply comparing left and right context. For modeling the left state of unseen triphone, first, we estimate monophones, diphones, and triphones and then cluster monophones into the phonologically meaningful classes of phones. Among trained triphones, we select triphones which have the same left-context and belong to the same class in right-context with unseen triphone. Then we find the triphone which has the maximum training token among the selected triphones. The left-state of the unseen triphone is tied with the left-state of selected triphone. Right-state can be modeled similarly.

## 3.2 Top-down Clustering

By applying first three constraints of bottom-up clustering, we build a decision tree for each Markov states of each base phone except center state. Our system has 52 phonemes. When there are three states for each phonetic HMM, that makes 104 trees in total.

We classified Korean vowels and consonants into 30 classes according to the horizontal and vertical places of articulations. The linguistic simple questions querying about the left or right context of a triphone and composite questions to alleviate the data fragmentation problem are formed. First, we grow simple tree having about 5 to 10 leaves and then, combine the leaves into two groups and calculate the total entropy. Among all combination, the combinations which has minimal total entropy was chosen.

We used the information which was obtained from bottom-up clustering. The agglomerative approach is more flexible in considering all possible configurations for seen triphones. So we first performed bottom-up clustering to the distributions. Then in the pruning stage of top-down clustering, we used the number of clusters from bottom-up clustering as stopping criterion. In the growing stage, as was bottom-up clustering, we did not stop growing until a minimal entropy reduction failed. Then in the pruning stage, we pruned the merges that had the most delta entropy until the number of clusters from bottom-up clustering was left, which is expected to the proper number of cluster and the clusters are expected to be well trained by the given training corpus.

# 4. EXPERIMENTAL RESULTS

## 4.1 Experimental Conditions

Input speech for training and test is sampled at 16kHz. It is initially pre-emphasized (1-0.953$z^{-1}$) and grouped into frames of 320 samples with a shift of 160 samples. For each frame, a Hamming window was applied and then 12-dimensional LPC-derived cepstral vector was computed. The cepstrum vector and its first and second time derivatives are used as features. The first time derivative vector of cepstrum was computed with time difference of 20msec and 40msec. Thus each speech frame was represented by a vector of 48 features. A set of 52 phonemes in Korean language is used as a base phoneme set. Each subword unit is modeled by a three-state left-to-right HMM. Each state is characterized by a 4-mixture Gaussian state observation density. Training is done with Baum-Welch algorithm. Class-based bigram and A* N-best search algorithm are used as a language model and search algorithm, respectively.

We selected 6700 phonetically balanced words set which covers all phonological events in the noun words from general Korean dictionary to implement Korean flexible isolated word recognition system. That is, this set includes all the context-dependent phones in the noun words and has minimum words. We use only noun words for training, because Korean verbs and adjectives have the same ending sound for almost all words.

For training and test, 40000 utterances which are spoken by 240 males and 160 females were used. So, the system is gender-independent. For training purpose, we used 38000 utterances spoken by 228 males and 152 females. In order to test, as in-vocabulary test set, 2000 utterances spoken by 12 males and 8 females were used. As out-of vocabulary set, we used 768 words Korean company name set and 335 words Korean name set.

## 4.2 Experimental Results

Table 1 shows the test results of out-of-vocabulary test set without unseen triphone. We used 768 Korean company names as test set. In table 1, we clustered distributions into 1000 clusters. We used 1000 clusters to reduce memory requirement as much as possible for practical reason while maintaining system performance. When we used 1000 senones, the memory requirement is approximately 2 Mbyte.

| | Triphone | Bottom-up | Top-down |
|---|---|---|---|
| % Error | 6.13 | 6.00 | 6.88 |

**Table 1:** Error rate of out-of vocabulary set without unseen triphone

Since we used the cluster number of bottom-up clustering as the stopping criterion of pruning of top-down clustering, the number of senones of bottom-up and top-down clustering is almost same. Table 1 shows the results of bottom-up and top-down clustering when we used senone smoothing.

The effect of smoothing on the covariance matrices is investigated. The covariance matrices are smoothed by comparing the sharpness ratio $R_i$ with several threshold, such as 0.1,1,10,100. And for each threshold, those above the threshold are smoothed with the several interpolating parameters, λ, such as 0.2, 0.5, 0.7, 1.0. λ=1.0 means no covariance smoothing. Figure 1 shows the results of covariance smoothing when we apply it to the covariance matrix of SCHMM.

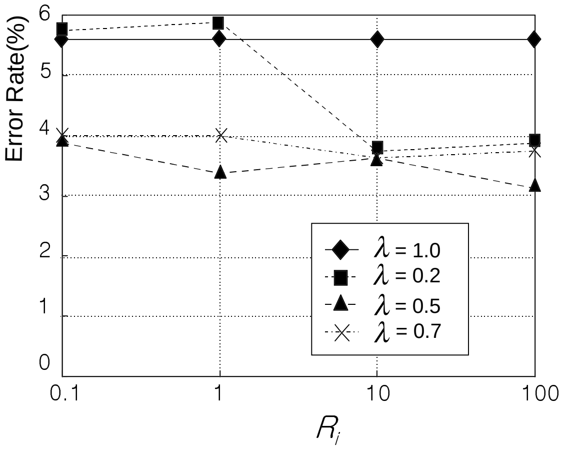

**Figure 1:** Results of covariance smoothing

Table 2 shows the test results of in-vocabulary test set. For in-vocabulary test, we used 20 speech files with a vocabulary of 1000 words.

| | Triphone | Bottom-up | Top-down |
|---|---|---|---|
| % Error | 12.60 | 12.15 | 13.10 |

**Table 2:** Error rate of in-vocabulary set

In table 2, we also clustered distributions into 1000 clusters. We compared the performance of floor method and CSM for the distributions of SCHMM. Table 3 shows the results.

| | Senone Smoothing | |
|---|---|---|
| | Floor Method | CSM |
| %Error | 13.15 | 12.15 |

**Table 3:** Results of senone smoothing

We used 335 Korean names as another out-of-vocabulary set with unseen triphone. The total number of unseen triphones in the test set is 244. So 0.728 unseen triphone per word was occurred. Table 4. shows the results.

| | Triphone | Bottom-up | | Top-down |
|---|---|---|---|---|
| | | Backoff | MMCM | |
| % error | 15.75 | 15.25 | 14.50 | 13.56 |

**Table 4:** Error rate of out-of vocabulary set with unseen triphone

Results of table 1,2 and 4 show that MMCM outperforms the traditional context-independent backing-off method and in case of out-of-vocabulary with unseen triphone, the top-down method outperforms the bottom-up method.

## 5. CONCLUSION

We applied state level tying by using bottom-up clustering and top-down clustering. We also applied co-occurrence smoothing method to the clustered states and variance smoothing to the covariance matrix of SCHMM. In senone smoothing, we obtained 7.6% error reduction rate by using CSM compared with uniform smoothing. And 48% error reduction rate was obtained by covariance smoothing .

We have proposed clustered co-occurrence smoothing method and by applying it to the distributions of DHMM, more accurate state clustering in the training procedure can be performed. And we also proposed a new model composition method for modeling of unseen triphone and obtained 5% error reduction rate compared with conventional context-independent phone baking-off method.

## 6. REFERENCES

1. R. Schwartz, O. Kimball, F. Kubala, M.W. Feng, Y.L. Chow, C. Barry, J. Makhoul, “Robust Smoothing Methods for Discrete Hidden Markov Models”, Proc. ICASSP 1989, pp. 548-551.
2. Y. Zhao, “A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density of Phoneme-Sized Units”, IEEE Trans. Speech and Audio Processing, vol. 1, no. 3, 1993, pp.345-361.
3. M.Y. Hwang, X. Huang, “Shared-Distribution Hidden Markov Models for Speech Recognition”, IEEE Trans. Speech and Audio Processing, vol. 1, no. 4, 1993, pp.414-420.
4. M.Y. Hwang, X. Huang, F.A. Alleva, “Predicting Unseen Triphones with Senones”, IEEE trans. Speech and Audio Processing, vol.4, no. 6, 1996, pp.412-419.
5. C.H. Lee, B.H. Juang, W. Chou, J.J.M. Perez, “ A Study on task-Independent Subword Selection and Modeling for Speech recognition”, Proc. ICSLP, 1996.