STATE-DEPENDENT MIXTURE TYING WITH VARIABLE CODEBOOK SIZE FOR ACCENTED SPEECH RECOGNITION

LIU Yi[1], *ZHENG Fang*[1], *HE Lei*[2], *XIA Yunqing*[1]

[1]Center for Speech and Language Technologies, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology, Beijing, China
{eeyliu,fzheng, yqxia}@tsinghua.edu.cn

[2]Toshiba (China) Research and Development Center, Beijing, China
helei@rdc.toshiba.com.cn

ABSTRACT

In this paper, we propose a state-dependent tied mixture (SDTM) models with variable codebook size to improve the model robustness for accented phonetic variations while maintaining model discriminative ability. State tying and mixture tying are combined to generate SDTM models. Compared to a pure mixture tying system, the SDTM model uses state tying to reserve the state identity; compared to the sole state tying system, such model uses a small set of parameters to discard the overlapping mixture distributions for robust model estimation. The codebook size of SDTM model is varied according to the confusion probability of states. The more confusable a state is, the larger its codebook size gets for a higher degree of model resolution. The codebook size is governed by state level variation probability of accented phonetic confusions which can be automatically extracted by frame-to-state alignment based on the local model mismatch. The effectiveness of this approach is evaluated on Mandarin accented speech. Our method yields a significant 2.1%, 9.5% and 3.5% absolute word error rate reduction compared with state tying, mixture tying and state-based phonetic tied mixtures, respectively.

Index Terms— State-dependent tied mixture models, variable codebook size

1. INTRODUCTION

Continuous-density hidden Markov model (CDHMM) triphones have been widely used in most state-of-the-art automatic speech recognition systems for robust model generation [1, 2]. With sufficient parameters, triphone models have strong ability to cover many substitution and insertion errors [3]. However, triphone models are independent from each other and have their own mixture distributions, many of which have overlapping distributions and lead to high model complexity with thousands of states and thousands upon thousands Gaussian distributions. Estimation of such a large amount of parameters requires a lot more training data for a reliable estimation [2, 4]. In addition, with the increased acoustic and phonetic variability as well as coarticulations in accented speech, mixture distributions tend to be inadequate for covering the high degree of acoustic variations within phonetic units.

In this situation, state tying and mixture tying approaches can be adopted at different levels to reduce the redundant parameters and provides a good balance between recognition accuracy and training data availability [1, 2, 4]. State tying addresses the conflict between model accuracy and available training data [4]. Different triphone models are tied according to their acoustic similarity. Similar triphones are grouped when they are poorly trained or when they are acoustically close to each other, and share a common set of parameters. Data-driven method and decision tree based state tying approach [2, 5] are commonly used to generate generalized-triphone models and tied-state triphone models. Another method of complexity reduction is mixture tying in which a universal codebook of Gaussian distributions is shared by all HMM states while each state has different mixture weights. It aims at keeping the model accuracy by using a large number of Gaussian components with appropriate model complexity. Tied-mixture (TM) models and phonetic tied-mixture (PTM) model share parameters in the acoustic space, and are commonly used in mixture tying system [6]. In the PTM model, the whole set of Gaussian distributions in the acoustic space is classed into different independent sets consisted of Gaussian components for phone-dependent HMM states.

State tying and mixture tying address the fundamental conflict between the model complexity and robustness with a limited amount of training data, challenges still remain in these approaches, especially for accented speech recognition. State tying forces certain states to be identical which causes the acoustic parameters corresponding to tied states to be invariably overlapped, and seriously degrades the model discriminative ability to represent the high degree of phonetic confusions in accented speech. In conventional TM and PTM triphone models, all Gaussian distributions in different states or phone-dependent states are covered by a single codebook with large size, under-training can be a severe problem since each state has a large number of mixture weight parameters to estimate, and a large number of mixture weights are small in magnitude [7]. Therefore, the efficiency of the mixture distributions is low and lacks enough discriminative ability to model the rich phonetic variations in accented speech. Moreover, for Gaussian sharing approach, only the shared Gaussians can be estimated efficiently, whereas the way the parameters are tied must be fixed beforehand [8].

In this paper, we propose to combine state tying with mixture tying to generate state-dependent tied-mixture models for an

efficient complexity reduction of triphone models. Compared to a pure mixture tying system, the SDTM model uses state tying to reserve the state identity; compared to the sole state tying system, such model uses a small set of parameters to discard the overlapping mixture distributions for robust model estimation. Unlike conventional TM or PTM models where all HMM states or phone-dependent HMM states share a single codebook, our SDTM models use an independent codebook for each separate HMM tied state. Unlike commonly used decision tree based state tying method where each individual tied state has its own mixture distribution, our SDTM models allow different tied states belonging to a same decision tree share a same codebook. Furthermore, we generate variable codebook size based on the state level variation probability (SLVP) learned from data to improve the discriminative power of our model to deal with the high degree of phonetic confusions in accented speech. The variable-size codebook includes mixture distributions of its own model as well as those borrowed from the relevant triphone models governed by SLVP. The more confusable a model is, the larger codebook size it gets. The SDTM model with variable codebook size can be regarded as an intermediate level description between the CDHMM tied-state triphone models and the TM and PTM models. It achieves a better tradeoff between complexity and accuracy.

The paper is organized as follows. In Section 2, we describe the approach of combining state tying with mixture tying. Section 3 outlines our method of generating variable codebook size based on SLVP. In Section 4, experimental results on accented Mandarin telephony speech are presented. We conclude in Section 5.

2. STATE TYING WITH MIXTURE TYING

The generation of SDTM model is a combination of state tying and mixture tying. In the SDTM model, state tying is used to reserve the state identity; mixture tying is used to improve the efficiency of Gaussians and discard the overlapping mixture distributions for robust model estimation.

2.1. State Tying

State tying is usually achieved by decision tree based state clustering [2]. Decision trees are phonetic binary trees in which a yes/no phonetic question is attached to each node. Initially, all states in a given item list, typically a specific phone state position, are placed at the root node of a tree. Depending on each answer, the pool of states is successively split and this continues until the states have trickled down to leaf nodes. All states in the same leaf node are then tied. The set of questions designed is based on the phonetic knowledge and is regarded as clustering rules. Generally, the phonetic questions are symmetric. The question at each node is chosen to maximize the likelihood of the training data given the final set of tied states. In this tree structure, the root of each decision tree is a basic phonetic unit with a certain state topological location, triphone variants with the same central phone but different contextual phones are clustered to different leaf nodes according to the clustering rules.

2.2. Mixture Tying

We use Gaussian clustering and merging based on the minimum local distance distortion to perform mixture tying. We establish a set of Gaussian components to form a codebook that is shared by tied states belonging to a same decision tree. Bhattachyaryya distance measure is used to determine the similarity between two Gaussian components. Given two Gaussian components, $G_1(\mu_1,\sigma_1)$ and $G_2(\mu_2,\sigma_2)$, the Bhattachyaryya distance measure is represented as

$$D(G_1,G_2)=\frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{\sigma_1+\sigma_2}{2}\right)^{-1}\times(\mu_1-\mu_2)+\frac{1}{2}\ln\frac{\left|(\sigma_1+\sigma_2)/2\right|}{|\sigma_1|^{\frac{1}{2}}\cdot|\sigma_2|^{\frac{1}{2}}} \quad (1)$$

where μ and σ are the mean and variance of a Gaussian component. Suppose we have M decision trees, i.e., the whole acoustic space is divided into M sub-acoustic spaces. Each sub-acoustic space contains a set of Gaussian components derived from the accumulated mixture distributions of tied states within a same decision tree. Based on the Bhattachyaryya distance measure, a set of Gaussian components in a sub-acoustic space are clustered and merged by successive binary splitting. These components in a sub-acoustic space are divided into different clusters, components in a same cluster are merged to form a single Gaussian component.

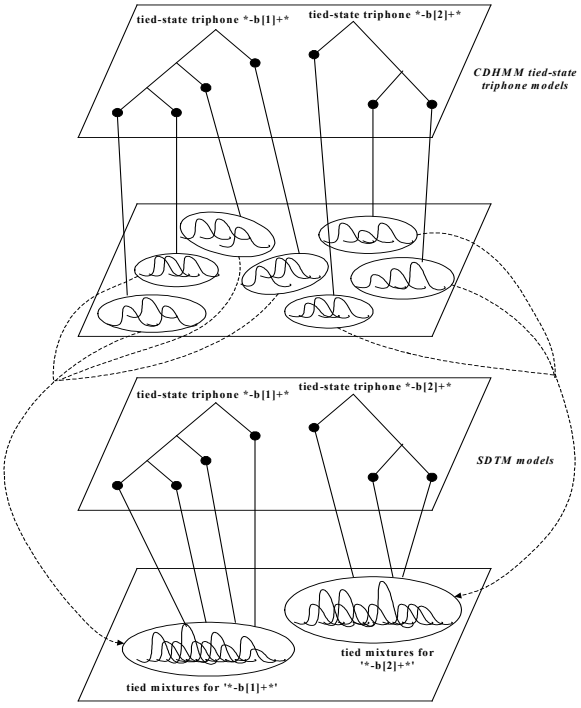

Fig.1: The structure of SDTM model versus tied-state triphones.

The merge algorithm based on likelihood loss computation [9] is illustrated in the following equations. Given two Gaussian components, $G_1(\mu_1,\sigma_1)$ and $G_2(\mu_2,\sigma_2)$ with their relevant occurrence counts in the training set c_1 and c_2. When G_1 and G_2 are merged to form a signal Gaussian component, the new component has the following coefficients:

$$
\begin{array}{cccccccccccccc}
 & \overbrace{\quad\quad\quad\quad}^{Class\ 1} & & & & \overbrace{\quad\quad\quad\quad}^{Class\ 2} & & & & & \overbrace{\quad\quad\quad\quad}^{ClassM} & & & \\
 & u_{1,1} & u_{1,2} & \cdots & u_{1,N_1} & u_{2,1} & u_{2,2} & \cdots & u_{2,N_1} & & u_{M,1} & u_{M,2} & \cdots & u_{M,N_M} \\
v_{1,1} & p_{1|1,1} & p_{1|1,2} & \cdots & p_{1|1,N_1} & v_{2,1} & p_{1|2,1} & p_{1|2,2} & \cdots & p_{1|2,N_2} \cdots & v_{M,1} & p_{1|M,1} & p_{1|M,2} & \cdots & p_{1|M,N_M} \\
v_{1,2} & p_{2|1,1} & p_{2|1,2} & \cdots & p_{2|1,N_1} & v_{2,2} & p_{2|2,1} & p_{2|2,2} & \cdots & p_{2|2,N_2} \cdots & v_{M,2} & p_{2|M,1} & p_{2|M,2} & \cdots & p_{2|M,N_M} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
v_{1,K} & p_{K|1,1} & p_{K|1,2} & \cdots & p_{K|1,N_1} & v_{2,K} & p_{K|2,1} & p_{K|2,2} & \cdots & p_{K|2,N_2} \cdots & v_{M,K} & p_{K|M,1} & p_{K|M,2} & \cdots & p_{K|M,N_M}
\end{array}
$$

Fig.2: a weighted coefficient matrix of SDTM models.

$$
\begin{aligned}
& c = c_1 + c_2;\ \mu = \frac{c_1\mu_1 + c_2\mu_2}{c} \\
& \sigma = \frac{c_1\left[\sigma_1 + (\mu_1 - \mu)(\mu_1 - \mu)^T\right]}{c} \\
& \quad + \frac{c_2\left[\sigma_2 + (\mu_2 - \mu)(\mu_2 - \mu)^T\right]}{c}
\end{aligned} \tag{2}
$$

The likelihood loss due to the merge of G_1 and G_2 is

$$
\Delta_{1+2} = \frac{c\log|\sigma| - c_1\log|\sigma_1| - c_2\log|\sigma_2|}{2} \tag{3}
$$

The most similar pair of Gaussian components is thus defined to be the pair that, when merged, gives the least likelihood loss. The clustering and merging procedure can be performed iteratively. This obtained codebook of Gaussian components is shared across all of the tied states within a same decision tree. The difference between the state tying triphone models and SDTM models is illustrated in Fig. 1.

2.3. Structure of SDTM Models

In SDTM models, the emission probability distributions of tied states within a decision tree share the same codebook with different mixture weights. Suppose we have M decision trees, the codebook for each class is composed of K codewords, i.e., Gaussian components. The structure of SDTM can be represented as a weighted coefficient matrix $\{p_{k|m,n}\cdot v_{m,k}\}_{K\times M\times N_m}$ as shown in Fig. 2, where $v_{m,k}$ is the k th codeword belonging to class m, and $p_{k|m,n}$ is a coefficient which has the probabilistic interpolation and can be regarded as mixture weight. u denotes the tied-state, N_m the total number of tied-states of the m th class. Given a tied state n in a certain decision tree of class m, the state output distribution is

$$
\begin{aligned}
& P(x \mid u_{m,n}) = \sum_{k=1}^{K} p_{k|m,n} f(x \mid v_{m,k}) \\
& 1 \le m \le M,\ 1 \le n \le N_m
\end{aligned} \tag{4}
$$

where the mixture weights $p_{k|m,n}$ satisfy $\sum_{k=1}^{K} p_{k|m,n} = 1$.

The SDTM model can be regarded as a bridge between the CDHMM tied-state triphone models and the conventional TM and PTM models. Compared to CDHMM, the SDTM model uses state tying to reserve state identity and can be considered as a special form of CDHMM in which tied states are separate from each other, and mixture distributions of those tied states belonging to a same decision tree are tied. From Fig. 2, if the codebook $\{v_{m,k}, k = 1,2,\cdots,K\}$ of class m is different from state to state in a decision tree, the SDTM model becomes a CDHMM tied-state model. Compared to TM or PTM, the SDTM model separates the whole acoustic space into M classes according to the total number of decision trees instead of using a universal codebook of Gaussian components for all HMM states or phone-dependent HMM states. Different codebooks are assigned to different decision trees. If all the codebooks $\{v_{m,k}, k = 1,2,\cdots,K\}$ belonging to different classes are accumulated to form a universal codebook v_k, where $v_k = \sum_{m=1}^{M} v_{m,k}$, the SDTM models back off to conventional TM models. On the other hand, if the codebooks of different state location topology with a same phonetic unit are clustered, the SDTM models back off to conventional PTM models.

3. VARIABLE CODEBOOK SIZE GENERATION

3.1. State Level Variation Probability

In order to capture the diversity of variations in accented speech at state levels, we estimate the SLVP from frame-to-state alignment that considers the frame level error caused by local model mismatch. The alignment is between the frame sequence and the HMM state sequence, and it does not require the HMMs to have the identical state number that allows the topology of HMMs to be flexible. Taking the frame level errors into consideration enables the estimation to be reliable, and the spurious mappings at the state level caused by noise can be discarded.

Let b and s denote the baseform (canonical) and surface form (alternative) sequence. Let $U^b = u_1^b, u_2^b, \cdots, u_T^b$ and $U^s = u_1^s, u_2^s, \cdots, u_T^s$ are the time sequence of baseform and surface form states, respectively. In addition, define $X = x_1, x_2, \cdots x_T$ the input acoustic vectors. T is the total frame number of each input utterance. The procedure of estimating SLVP based on frame-to-state alignment is:

1. Generate the baseform time state sequence U^b using the forced alignment, and keep the track of full state sequence at the same time. In addition, at each frame t, the output acoustic likelihood score $L_t^b = P(x_t \mid u_t = u^b)$ is saved.

2. Generate the surface form time state sequence U^s using phone recognition, and also keep the track of full state sequence. Similarly, the output acoustic likelihood score $L_t^s = P(x_t \mid u_t = u^s)$ at each frame is saved.
3. At frame t, if $u^b \neq u^s$, frame level error occurs. The frame level error is denoted as $E = \left| L_t^b - L_t^s \right|$. The smaller the E, the more similar the two mapping states. If the frame level error of this type of state pair occurs frequently in the training set, it means the baseform phone state and the surface form state is always confused. Table 1 indicates an example of frame level errors.
4. Set *threshold1* to filter the state pairs which have the high frame level error score. These errors may be caused by the noise or the accidental frame level state mappings.
5. Set *threshold2* to discard those state pairs which have the rare occurrence numbers. Calculate the SLVP according to Eq. (5), where i, j denote the state number and $Occ(\cdot)$ is the total occurrence number.

$$P(u^{s_i} \mid u^{b_j}) = \frac{Occ(u^{b_j}, u^{s_i})}{\sum_{u^{si}} Occ(u^{b_j}, u^{s_i})} \qquad (5)$$

When the SLVP for all the baseform and surface form state pairs in the training set has been calculated, stop.

Therefore, a direct relation between the baseform state and surface form state can be established through the frame-to-state alignment. Obviously, it is not a one-to-one match during the alignment. A baseform state may be aligned to one or more surface form states and vice versa. On the other hand, conventional methods of state-to-state alignment relying on phoneme-to-phone alignment can only produce one-to-one match. Hence, SLVP can be estimated more flexibly and covers more combination of variations.

Frame index	Baseform state sequence	Surface form state sequence	Frame level errors
128	i1[2]	ze[4]	0.828867
129	i1[3]	ze[4]	0.394684
130	i1[3]	u[2]	2.395267
131	i1[4]	u[3]	4.855088
132	i1[4]	u[3]	3.583070
133	i1[4]	u[3]	6.597553
134	i1[4]	u[3]	2.272062
135	i1[4]	u[4]	2.363436
136	za[2]	u[4]	5.697863
137	za[2]	da[2]	1.097567
138	za[2]	da[2]	5.531465
139	za[2]	da[2]	6.324440
140	za[3]	da[2]	6.409951
141	za[3]	da[3]	8.266272
142	za[4]	da[3]	1.695197
143	za[4]	da[4]	0.909124
144	a[2]	da[4]	**20.130319**
145	a[2]	da[4]	**8.988168**

Table 1: An example of frame-to-state mapping and local frame level errors.

3.2. Adjust Codebook Size

The SLVP can be incorporated into acoustic model at state level to adjust the codebook size of SDTM models. Based on the structure of SDTM shown in Fig.2, the state output distribution is as follows:

$$P(x \mid u^b) = \sum_{k=1}^{K} p_{k|C_i,n_i} f(x \mid v_{C_i,k})$$
$$1 \le C_i \le M \ , \ 1 \le n_i \le N_{C_i} \qquad (6)$$

where u^b is a tied state of a decision tree corresponding to the class C_i, N_{c_i} is the total number of tied states of such decision tree, $f(x \mid v_{C_i,k})$ is a Gaussian density whose mean and variance are specified in the codeword $v_{C_i,k}$. The codebook can be simply represented by a set of codewords $\{v_{C_i,k}, k = 1,2,\cdots,K\}$. Similar description is for surface form state u^s in class C_j ($1 \le C_j \le M$ and $C_i \neq C_j$). In the following equations, $f(x \mid v_{C_i,k})$ is represented as $f_{C_i,k}(\cdot)$ for simplification.

Let $P'(x \mid u^b)$ be the new output distribution of the baseform state after taking SLVP and the effects of other surface form states into consideration, we have

$$P'(x \mid u^b) = \sum_{u^s} P(x \mid u^s) P(u^s \mid u^b)$$
$$= \lambda P(x \mid u^b) + (1-\lambda) \sum_{l=1}^{L} P(x \mid u^s) P(u^s \mid u^b) \quad (7)$$

where λ can be regarded as a linear interpolation coefficient for combining the different acoustic models. It is determined by the probability of the baseform state being recognized as itself. For instance, if 'p[2]' has 70% probability to be recognized as 'p[2]' and 30% for other variations, then $\lambda = 0.7$.

Accented variations are unidirectional [10], and are one-to-many mapping representing different types of phonetic changes, e.g., $b \to d$ and $b \to p$, Substituting Eq. (6) into Eq. (7) giving us

$$P'(x \mid u^b) = \sum_{k=1}^{K} p' f_{C_i,k}(\cdot) + \sum_{l=1}^{L} \sum_{k=1}^{K} p'' f_{C_j^l,k}(\cdot) \qquad (8)$$

where L is the total number of "borrowed" models related to the baseform state u^b. In addition, the baseform state u^b and surface form state u^s correspond to classes C_i and C_j, tied states in a decision tree share one codebook mixture distributions. Therefore, the class confusion probability $P(C_j^l \mid C_i)$ is equivalent to SLVP $P(u^s \mid u^b)$. p' and p'' are the weights of the mixture distribution, they are

$$p' = \lambda \cdot p_{k|C_i,n_i}$$
$$p'' = (1-\lambda) \cdot P(C_j^l \mid C_i) \cdot p_{k|C_j^l,n_j} \qquad (9)$$

For simplification, the output distribution can be expressed as

$$P'(x \mid u_b) = \sum_{k=1}^{K+LK} p''' f'(\cdot) \qquad (10)$$

where p''' is the relevant mixture weight. It includes both p' and p'', satisfies $\sum_{k=1}^{K+LK} p''' = 1$, and is proportional to the class confusion probability $P(u^s \mid u^b)$. From Eq. (10) we can see that the codebook size is augmented, and varies according to the value of L. The new codebook includes its original mixture distribution as well as those borrowed from the surface form states. If a tied state has a higher confusability, it gets more mixture distributions from the codebook of the relevant surface form states for higher model robustness. The codebook size of the reconstructed baseform acoustic model varies according to the confusability of states. If the matrix of SLVP is diagonal, the codebook size of SDTM model is unchanged. Otherwise, it is adjusted based on the SLVP. Using borrowed codebook of mixture distributions from alternative surface form states adjusts the original mixture distribution structure of baseform state, which enables the borrowed mixture distributions to cover the boundaries of the original mixture distribution. More Gaussians in this region makes it possible to model in detailed distributions that may be ignored by simply increasing the codebook size through mixture splitting method.

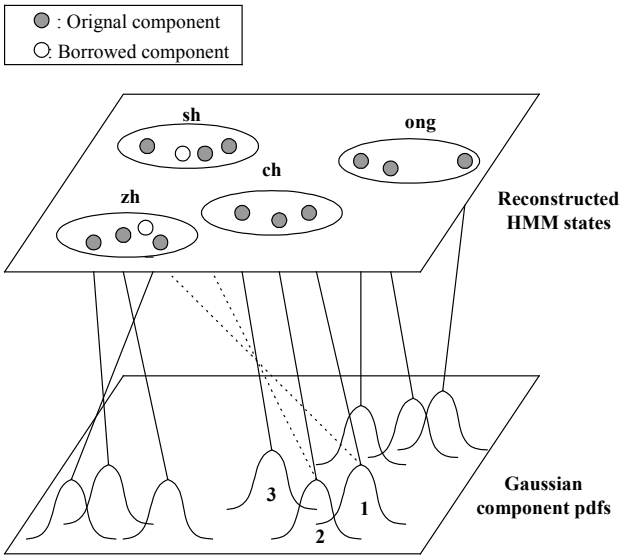

Fig.3: Use of the dominant Gaussian for variable codebook size.

Moreover, each Gaussian in the codebook of surface form states does not have the identical contribution to accommodate accented changes. Only those Gaussians representing the confusing acoustic samples are required to considered and mixed in the baseform states as shown in Fig.3. In order to consider the unidirectional and asymmetric property of acoustic confusions in accented speech, an asymmetric acoustic distance measure described in [10] was used to select dominant Gaussians for mixture borrowing to form variable codebook size.

4. RECOGNITION EXPERIMENTS

We evaluated our approach in a Chinese telephony short phrase speech recognition task. There is no word n-gram in these short phrases so that we can isolate the effect of our approach without the influence from high-level information. All speech data were sampled at 8 kHz and 8 bit rate. The baseline acoustic model was trained using 100 speakers' utterances with around 50 hours of Putonghua speech. The HMM topology is three-states, left-to-right without skips. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. Standard Chinese 21 initials and 38 finals were used to generate HMMs. Dev_set contained 2000 utterances from 10 Cantonese-accented speakers and 10 Wu-accented speakers, each speaker has 100 utterances. The Dev_set was used to estimate SLVP. The testing data consisted of two parts: Test_set1 includes 1800 Cantonese-accented and Wu-accented utterances from 18 speakers (8 females and 10 males), excluded from Dev_set. Test_set2 is used for performance comparison and consists of 900 Putonghua utterances selected from 9 native speakers. Speakers were instructed to speak the same phrases in these two test sets.

We started from the context-independent initial and final CDHMM models, and used the HTK flat-start to generate state tying, mixture tying and SDTM models with variable codebook size, respectively. We established 177 decision trees. Through the use of these trees, the overall 25000 triphones were tied to 12 Gaussian-component models with 5500 tied-states (M1), there were 66000 Gaussians in total. We generated a conventional PTM model in which 15140 (59*256) Gaussian components were clustered into 59 different codebooks shared by all phone-dependent states (M4). The codebook size of M4 is 256. We also generated a state-based PTM model discussed in [7] where the codebook size was fixed and cloned from monophone models (M5). The total number of Gaussians of M5 is 15222 (59*3*86). Using mixture tying method shown in Section 2.2 as well as SLVP with variable codebook size adjustment, 66000 Gaussians of M1 were tied to 12480, on average, the codebook size was 71 (12480/(59*3)), i.e., each decision tree has 71 Gaussians and varied according to accented phonetic confusions (M6). For a fair comparison, we also generated state tying triphone models with 3 Gaussian-component models, 16500 Gaussians in total (M2). Moreover, we mixed Dev_set2 data with training set, and retrained a tied-state triphone models with 16500 Gaussians (M3) for a better comparison.

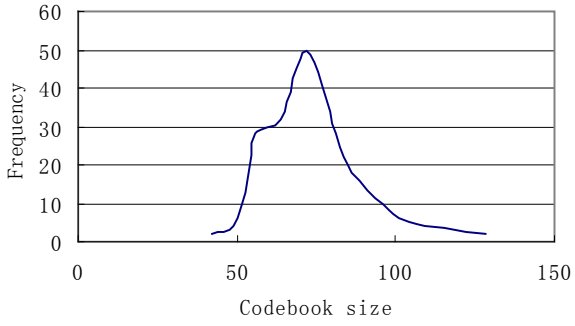

Fig. 4: Distribution of codebook size in SDTM models.

Fig.4 shows the codebook size distribution of the reconstructed models by taking into account of SLVP. It is found that the codebook size of decision tree distributes from 41 to 132 according to its degree of accented phonetic variations. We found that the high degree of accented confusions relates to retroflexed affricatives such as 'zh, ch, sh, r'. As a result, the codebook size of those is the largest among the 177 decision trees. Meanwhile, voiced stops, unvoiced stops and voiced fricatives also obtain a larger codebook size compared to other initial units due to the reason that they are short and flexible in pronunciations, especially in Wu-accented and Cantonese-accented speech.

Besides 12 Gaussian-component models with 5500 tied-states models, these five models were compared at the same state complexity and the similar number of Gaussian components. The results are shown in Table 2. It shows that the use of SDTM models with variable codebook size achieves the lowest WER at a comparable model complexity. It gives significant 2.1% (M6 to M2), 9.5% (M6 to M4) and 3.5% (M6 to M5) WER absolute reductions compared with state tying, mixture tying and state-based phonetic tied mixtures, respectively. Even in testing on Putonghua speech with the high identity between training set and testing set with fewer accented phonetic confusions, using our SDTM models with variable codebook size does not lead to performance degradation. Moreover, using the SDTM models yields additional 1.2% WER reduction with respect to the use of mixed data training model, which means the SLVP with variable codebook size improves the robustness of models to cover accented variations. In addition, the results suggest that reserving the state identity information together with a variable small size codebook for mixture tying is beneficial for keeping the model resolution. Compared to M1 model, our SDTM model achieves a comparable WER only with 1/5 Gaussian mixture numbers. We believe the improvement comes from the higher model robustness with variable codebook size and a more efficient mixture distribution of each codebook in our proposed model.

System	Total Gaussians	Word Error Rate (WER) %	
		(Test_set1) Accented speech	(Test_set2) Putonghua speech
State tying models (**M1**)	66000 (5500*12)	15.8%	7.9%
State tying with less Gau. (**M2**)	16500 (5500*3)	18.3%	8.1%
Mixed training models (**M3**)	16500 (5500*3)	17.4%	8.8%
PTM Models (**M4**)	15140 (59*256)	25.7%	11.3%
State-based PTM (**M5**)	15222 (59*3*86)	19.7%	10.5%
Our SDTM models (**M6**)	**12480**	**16.2%**	**7.9%**

Table 2: The SDTM model outperforms the state tying and the mixture tying models at a comparable model complexity.

5. CONCULSIONS

We presented a method to combine state tying with mixture tying to generate state-dependent tied-mixture (SDTM) models with variable codebook size for an efficient complexity reduction of triphone models. The codebook size is governed by state level variation probability of accented phonetic confusions which can be automatically extracted by frame-to-state alignment based on local model mismatch. The more confusable a state is, the larger its codebook size gets for a higher degree of model resolution. Compared to a pure mixture tying system, the SDTM model uses state tying to reserve the state identity; compared to the sole state tying system, such model uses a small set of parameters to discard the overlapping mixture distributions for robust model estimation. We have shown that our proposed model leads to significant WER absolute reduction of 2.1%, 9.5% and 3.5% compared with state tying, mixture tying and state-based phonetic tied mixtures, respectively. In addition, our method provides an additional 1.2% WER reduction with respect to the use of mixed data training model. It proves that the variable codebook size with state tying and mixture tying together improves the robustness of the model while keep good model resolution. We achieved a good balance between model complexity and robustness.

6. REFERENCES

1. X. Huang, A. Acero, H. Hon and R. Reddy, Spoken language processing: A guide to theory, algorithm, and system development. Upper Saddle River, NJ: Prentice Hall PTR, 2001.
2. S. Young, et al., The HTK book (for HTK Version 3.2). Entropic Cambridge Research laboratory, 2002.
3. Jurafsky, D., Ward, W., Zhang, J.P., Herold, K., Yu, X.Y. and Zhang, S., "What kind of pronunciation variation is hard for triphones to model?" In Proc. ICASSP, pp.577-580, 2001.
4. L. Gu and K. Rose, "Sub-state tying in tied mixture Hidden Markov Models," in Proc. ICASSP, pp.1013-1016, 2000.
5. X. Luo and F. Jelineck, "Probabilistic classification of HMM states for large vocabulary continuous speech recognition," in Proc. ICASSP, pp.353-356, 1999.
6. J. Park, H. Ko, "A new state-dependent phonetic tied-mixture model with head-body-tail structured HMM for real-time continuous phoneme recognition system", in Proc. ICSLP, pp.1583-1586, 2006.
7. A. Lee, T. Kawahara, K. Takeda and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," in Proc. ICASSP, pp.1269-1272, 2001.
8. Y. Liu and P. Fung, "State-Dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition" IEEE Trans. Speech and Audio Processing, Vol.12, No.4, pp.351-364, Jul. 2004.
9. L. Deng, A. Acero, L. Jiang, J. Droppo and X.D. Huang, "High-performance robust speech recognition using stereo training data," in Proc. ICASSP, pp.301-304, 2001.
10. P. Fung and Y. Liu, "Effects and Modeling of Phonetic and Acoustic Confusions in Accented Speech Recognition". Journal of the Acoustical Society of America, Vol.118, Issue 5, pp.3279 – 3293, Nov. 2005.