SEMI-TIED COVARIANCE MATRICES FOR ACOUSTIC MODELS BASED ON RANDOM FORESTS OF PHONETIC DECISION TREES

Jian Xue, Lili Che and Yunxin Zhao

IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA
Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
{jxue, lilyche}@us.ibm.com, zhaoy@missouri.edu

ABSTRACT

In this paper, we investigate combining semi-tied covariance matrices and Random Forests (RFs) based phonetic decision trees (PDTs) for acoustic modeling in conversational speech recognition. We first use the RF method to train multiple PDTs for each phone state unit, and generate multiple sets of acoustic models accordingly. We then apply semi-tied covariance matrices to each set of acoustic models to improve their fit to data. In decoding search we combine the likelihood scores from the multiple acoustic models for each speech frame. The viability of semi-tied covariance matrices with different tying classes are studied from their effects on the diversity of RF-based acoustic models as well as on the word accuracy of our task of telehealth automatic captioning. Experimental results indicate that semi-tied covariance matrices help enhance the diversity of the RFs-PDTs based acoustic models as well as increase word accuracy.

Index Terms— semi-tied covariance matrices, Random Forests, phonetic decision trees, acoustic modeling

1. INTRODUCTION

In HMM-based speech recognition systems, diagonal covariance matrices are usually used for ease of training and computation. Although full or block-diagonal covariance matrices allow better modeling of the correlations among speech feature components, they cause a dramatic increase in the number of parameters per Gaussian density, and therefore limit the number of densities which may be robustly estimated. Semi-tied covariance matrices [1] allow a few full covariance matrices to be shared by multiple Gaussian distributions, while each distribution maintains its own diagonal covariance matrix. This new form of covariance matrices decreases the number of parameters significantly, and it also effectively models the correlations of speech feature components.

Phonetic decision tree (PDT) based state tying is commonly used in acoustic modeling for large vocabulary continuous speech recognition. PDT incorporates phonetic knowledge into context-dependent phone-state clustering and model allophones that are not seen in training data [2]. Many efforts have been reported on improving PDT state tying in acoustic modeling [3]-[5], where the focuses were on finding optimal node splits in the tree growing procedure or finding an optimal model structure according to some information criteria. We recently proposed to use RFs to train a set of PDTs for each phone-state, and accordingly obtained multiple acoustic models [6]. During decoding search the multiple acoustic scores from the multiple acoustic models were combined to generate more accurate acoustic scores for each speech feature vector. On our telehealth automatic captioning task, this technique improved recognition performance significantly due to complementary contributions from multiple models.

Although the multiple acoustic models generated from the RF PDTs constitute a strong acoustic model for each RF tied state (see Section 2), only diagonal covariance matrices have been used for each acoustic model [6]. Given the success of semi-tied covariance matrices in conventional HMMs, it is of interest to examine whether this benefit would carry over to multiple acoustic models. This issue arises from the fact that the effectiveness of multiple models is jointly determined by the strengths of individual models and the diversity among the models. While using semi-tied covariance matrices would improve the strength of individual models, its effect on the diversity of multiple models is unknown. Additionally, it is of interest to study the effects of different semi-tied classes on the multiple models in comparison with the conventional models.

This paper is organized as follows. We introduce semi-tied covariance matrices in section 2, and RFs based PDTs in section 3. In section 4 we describe our method of combining RFs based PDTs and semi-tied covariance matrices. In section 5 we provide a detailed account of experimental results. We conclude our work in section 6.

---

1 This work was conducted when the first two authors were with the Department of Computer Science, University of Missouri – Columbia.
2. SEMI-TIED COVARIANCE MATRICES

Semi-tied covariance matrices as originally proposed by Gales [1] are an extension to the standard diagonal, block-diagonal, or full covariance matrices used with HMM's. Instead of having a distinct covariance matrix for every Gaussian component in the acoustic model, each covariance matrix consists of two elements, a component specific diagonal covariance element, $\Sigma_{diag}^{(m)}$, and a semi-tied class-dependent, non-diagonal matrix, $H^{(r)}$ (referred to as the semi-tied transform). The form of the covariance matrix becomes:

$$
\Sigma^{(m)} = H^{(r)} \Sigma_{diag}^{(m)} H^{(r)T}
$$

(1)

$H^{(r)}$ may be tied over a set of Gaussian components, for example all those associated with the tied states of a particular context-independent phone, or a set of context-independent phones. Each Gaussian component, $m$, has the parameters of a component weight, a component mean $\mu^{(m)}$, a component diagonal element of the semi-tied covariance matrix $\Sigma_{diag}^{(m)}$, and additionally it is associated with a semi-tied class $r$, which has an associated semi-tied transform $H^{(r)}$. Please refer to [1] for details of the semi-tied covariance matrices.

3. RANDOM FORESTS OF PHONETIC DECISION TREES

We use RFs to train a set of PDTs for each state of each phone unit, where the phone units are described by hidden Markov models (HMM) with 3 emitting states per HMM as defined in HTK [7]. We randomly (with uniform probability) choose a subset of $m$ phonetic questions without replacement out of a total of $M$ questions to train one set of PDTs for all phone states, with one PDT for one phone state, where the procedure of constructing a PDT is the same as the commonly adopted deterministic greedy method. We then randomly choose another subset of $m$ questions to generate another set of PDTs and so on.

Multiple sets of PDTs are thus generated for all phone-state units, and from which we obtain multiple sets of acoustic models $MS_{11}, \ldots, MS_{K}$. Each set $k$ has $N_k$ models $M_{1k}, \ldots, M_{N_k}$, corresponding to the $N_k$ tied states in the $k$th set of PDTs. Each triphone state $S_i$ therefore has $K$ models $M_{1i}, \ldots, M_{Ki}$, where $r_k$ is a mapping from $S_i$ to a tied state in the model set $k$. If for two triphone states $S_i$ and $S_j$, $i \neq j$, for $k = 1, \ldots, K$, that is, $S_i$ and $S_j$ belong to the same tied state in each of the $K$ PDTs, then we say that $S_i$ and $S_j$ belong to the same RF tied state. Fig. 1 illustrates the construction of RF tied states from two PDTs.

The triphone states of a RF tied state $ES_i$ share the same multiple models $M_{1i}, \ldots, M_{Ki}$. In single PDT-based method, a triphone state corresponds to a tied state which is modeled by a Gaussian mixture density (GMD). In the RF-based method, a triphone state corresponds to a RF tied state which is modeled by multiple GMDs.

![Fig. 1 An illustration of RF tied states.](image)

The performance of the RFs-PDTs based multiple acoustic models depends on the strength of individual PDTs, as well as the correlations among different PDT sets. Here we measure the correlations among acoustic model sets resulting from different PDT sets. Assume we have $U$ phone units, $K$ sets of acoustic models, and $N$ triphone states. Given a speech feature vector $x$, we compute the posterior probabilities of the triphone state $n$ for each model set $k$, i.e.,

$$
P_t(n|x) = P_t(x|n) / \sum_{n=1}^{N} P_t(x|n')
$$

(2)

where the prior probabilities are assumed uniform.

The speech feature vectors $x$ are grouped into phone units sets $\Omega_u$, $u = 1, \ldots, U$, based on Viterbi alignment [7] with baseline acoustic models. For model sets $i$ and $j$ and $x \in \Omega_u$, we have the posterior probability vector pairs $P_{ij} = [P_i(1|x), \ldots, P_i(N|x)]'$ and $P_{ij} = [P_j(1|x), \ldots, P_j(N|x)]'$. For each triphone state $n$, we compute the correlation coefficient between the posterior probabilities $P_t(n|x)$ and $P_t(n|x)$ by

$$
Corr_{uv}(i,j) = \frac{\sum_{t=1}^{T} [P_t(n|x_t) - \bar{P}_i(n|x)] [P_t(n|x_t) - \bar{P}_j(n|x)]}{\sqrt{\sum_{t=1}^{T} [P_t(n|x_t) - \bar{P}_i(n|x)]^2} \sqrt{\sum_{t=1}^{T} [P_t(n|x_t) - \bar{P}_j(n|x)]^2}}
$$

(3)

The final correlation among the $K$ sets of acoustic models is obtained by averaging $Corr_{uv}(i,j)$ over the $N$ triphone states, the $U$ phone units, and all $(i,j)$ pairs of model sets.

4. COMBINATION OF RFS-PDTS AND SEMI-TIED COVARIANCE MATRICES

The following procedure is used to combine semi-tied covariance matrices with RFs-PDTS. Firstly, we use the RFs method to train multiple PDTs for each phone state, and generate multiple sets of acoustic models accordingly, where diagonal covariance matrices are used in these steps. Next, based on the tying relationships and model structures already obtained in each set of acoustic models, we retrain the acoustic models by using the semi-tied covariance.
matrices, where separate semi-tied covariance matrices are generated for different model sets.

In [1], all states of all context-dependent phones associated with the same monophone were assigned to the same semi-tied class. In the current work we consider 5 different semi-tied classes as defined below.

1) Global: all the Gaussian components use the same semi-tied transform.
2) Two: Gaussians associated with the states of silence “sil” or short pause “sp” belong to one class, and the other components belong to another class.
3) Three: the first class is defined for the states of “sil” and “sp”, the second class is for the states of vowels, and the third one is for the states of consonants.
4) Fifteen: fifteen classes are defined in Table I below.
5) Phone: the class definition is similar to the one used in [1], and additionally, the filled-pause units are assigned to one class, and “sil” and “sp” are assigned to another class, with a total of 44 classes.

<table>
<thead>
<tr>
<th>Types of phonemes</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front vowel</td>
<td>iy, ih, eh, ae</td>
</tr>
<tr>
<td>Mid vowel</td>
<td>aa, er, ah, ax, ao</td>
</tr>
<tr>
<td>Back vowel</td>
<td>uw, uh, ow</td>
</tr>
<tr>
<td>Diphthong</td>
<td>ay, oy, aw, ey</td>
</tr>
<tr>
<td>Liquid semivowel</td>
<td>w, l</td>
</tr>
<tr>
<td>Glide semivowel</td>
<td>r, y</td>
</tr>
<tr>
<td>Nasal</td>
<td>m, n, nx, en</td>
</tr>
<tr>
<td>Voiced stop</td>
<td>b, d, g</td>
</tr>
<tr>
<td>Unvoiced stop</td>
<td>k, p, t</td>
</tr>
<tr>
<td>Voiced fricatives</td>
<td>dh, v, z, zh</td>
</tr>
<tr>
<td>Unvoiced fricatives</td>
<td>f, s, sh, th</td>
</tr>
<tr>
<td>Whisper</td>
<td>lh, wh</td>
</tr>
<tr>
<td>Articulate</td>
<td>ch, jh</td>
</tr>
<tr>
<td>Filled pause</td>
<td>aah, huh, oh, om, uhh, umm</td>
</tr>
<tr>
<td>Silence</td>
<td>sil, sp</td>
</tr>
</tbody>
</table>

5. EXPERIMENTAL RESULTS

5.1. Experimental setup

The proposed methods were evaluated on the Telemedicine automatic captioning system developed at the University of Missouri-Columbia [9]. The training and test datasets were speech data extracted from healthcare providers’ conversation with clients in mock telemedicine interviews. Speech features consisted of 39 components including 13 MFCCs and their first and second order time derivatives. Gaussian mixture density based hidden Markov models were used for within-word triphone modeling. Speaker dependent acoustic models were trained for five speakers Dr1 - Dr5. The task vocabulary size was 46,480. Baseline acoustic model used single PDT-based state tying, with the average number of tied states being 1429. Please refer to [9] for a detailed description of the experimental setup.

5.2. Experimental results

a. Word accuracy performance

We trained 50 sets of PDTs by using the RF technique in [6] and obtained 50 sets of acoustic models with diagonal covariance matrices. For each set of acoustic models, the number of randomly sampled phonetic questions \( m \) was set to be 150 out of the \( M = 216 \) questions, which has been verified in [6] as a good choice for balancing the strength of individual models and the correlations among the model sets. The number of Gaussian components per GMD was set to be 16 and 24. We then retrained the acoustic models using semi-tied covariance matrices, with the different semi-tied classes as discussed in Section 4. Table II gives the performances in word recognition accuracy averaged over the 5 speakers’ test sets, where the baseline used single PDTs with diagonal covariance matrices, and the decoding parameters of language model scale and word penalty were set to 14 and 6 for all the cases.

<table>
<thead>
<tr>
<th>GMD Mixture size</th>
<th>16</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78.96</td>
<td>78.15</td>
</tr>
<tr>
<td>RFs</td>
<td>80.98</td>
<td>81.97</td>
</tr>
</tbody>
</table>

Table II Word accuracies (%) averaged over five speakers

We observed that for the mixture sizes of 16 and 24, the RFs-PDTs based acoustic models improved word accuracy over the baselines by 2.02% and 3.82% absolute, respectively, and the semi-tied transforms improved baseline word accuracy by 0.31%–0.64% and 0.3%–0.55% absolute, respectively. Using semi-tied covariance matrices on top of the RFs-PDTs improved RFs-PDTs’ word accuracy by 0.57%–0.94% and 0.24%–0.47% absolute, respectively for the mixture sizes of 16 and 24. The effects of semi-tied covariance matrices were dependent on the definition of semi-tied classes as well as the types of acoustic models. For the baseline with semi-tied transforms, using the “three” semi-tied classes gave the best performance, while for the RFs-PDTs with semi-tied transforms, using the phone-level semi-tied classes gave the best performance. As the number of semi-tied classes increases, the number of parameters to be estimated increases also. With our limited training data in the telehealth task, the number of semi-tied classes cannot be too large for single PDTs. Since RF-based state tying is more robust to overfitting, we could use a larger number of semi-tied classes than the single PDT–based state tying. The robustness of the RF-based state tying to overfitting is also shown in its better performance for the larger mixture size of 24, where as a contrast, the baseline model obtained the best performances for the smaller mixture size of 16.

b. Performance versus model correlation
By using the method discussed in Section 2, we measured the correlations among acoustic model sets with the 5 different semi-tied classes. Table III provides the measured correlation data and the word accuracy on the dataset of Dr. 1. The correlation for RFs based PDTs alone was 0.790 and the word accuracy was 79.53%.

From Table III we can see that the use of semi-tied covariance matrices reduced the correlations among acoustic model sets, which contributed to the improvements in word recognition accuracy. However, among the five semi-tied classes, there does not exist a simple relation between correlation and word accuracy, since the difference among the correlation values are not large, and the performance of the RFs-PDTs also depends on the strengths of the individual PDTs, i.e., each set of acoustic models.

Table III Correlation and word accuracy versus semi-tied classes for Dr.1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>80.17</td>
<td>0.762</td>
</tr>
<tr>
<td>Two</td>
<td>80.42</td>
<td>0.765</td>
</tr>
<tr>
<td>Three</td>
<td>80.11</td>
<td>0.753</td>
</tr>
<tr>
<td>Fifteen</td>
<td>80.02</td>
<td>0.753</td>
</tr>
<tr>
<td>Phone</td>
<td>80.54</td>
<td>0.760</td>
</tr>
</tbody>
</table>

c. Rebalancing acoustic and language model scores

We have found that with RFs-PDTs and semi-tied transforms, the range of acoustic scores were increased relative to the baseline, indicating a strengthening of the acoustic models. Therefore, in decoding search, the weights for score contributions from the acoustic and the language models should be rebalanced to put more emphasis on the acoustic component. Fig. 2 shows the variation of word accuracy performances of the baseline, the RFs-PDTs, and the RFs-PDTs with semi-tied transforms with different decoding parameter settings for language model scale and word penalty, where the word accuracies were averaged over the 5 test sets as well as over the 5 definitions of semi-tied classes for RFs with semi-tied transforms.

Fig. 2 Word accuracies vs. decoding parameters of language model scale and word penalty

From Fig. 2 we observe that while the decoding parameter of (14, 6) was a good choice for the baseline, the RFs-PDTs with semi-tied transforms benefitted largely from using smaller language model parameters. For example, using the decoding parameters of (12, 4) gave the best results, where for the global, two, three, fifteen, and the phone-level semi-tied classes, the word accuracies were 82.75%, 82.68%, 82.59%, 82.61%, and 82.80%, respectively, all are much larger than the accuracy figures given in Table II with the decoding parameter of (14, 6).

5.3 Significance test

We have conducted a significance test (difference of two proportions) on the experimental results for the proposed methods. Please refer to [6] for details of the significance test. With the level of significance \( \alpha \) set at 0.01, the word accuracy improvements were statistically significant for the case of RFs-PDTs over baseline as well as for the case of RFs with semi-tied transforms over RFs-PDTs.

6. SUMMARY

In this paper, we have investigated combining semi-tied covariance matrices with RFs-PDTs based acoustic models. The semi-tied covariance matrices were applied to each set of acoustic models in RFs-PDTs, and the effects of five definitions of semi-tied classes were evaluated. On a large vocabulary conversational speech recognition task of telehealth automatic captioning, our experimental results have shown that the proposed method significantly improved the word accuracy performance, which is attributed to the strengthening of the individual sets of acoustic models as well as the reduction of their correlations.

Acknowledgement

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC004340-05A2.

The authors would like to thank Mr. Xin Chen for a helpful discussion on the decoding parameters.

REFERENCES