Tied-State Multi-path HMnet Model using Three-Domain Successive State Splitting

Soo-Young Suk, Hiroaki Kojima

Speech Processing Group, Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology, Japan
{sy.suk, h.kojima}@aist.go.jp

Abstract

In this paper, we address the improvement of an acoustic model using the multi-path Hidden Markov network (HMnet) model for automatically creating non-uniform tied-state, context-dependent hidden markov model topologies. Recent research has achieved multi-path model topologies in order to improve the recognition performance in gender-independent, spontaneous-speaking applications. However, the multi-path acoustic model size may increase and require more training samples depending on the increased number of paths. To solve this problem, we used a tied-state multi-path topology by which we can create a three-domain successive state splitting method to which environmental splitting is added. This method can obtain a suitable model topology with small mixture components. Experiments demonstrated that the proposed multi-path HMnet model performs better than single-path models for the same number of states.

Index Terms: Acoustic model, Multi-path HMM, Model Topology, PDT-SSS, Environmental Splitting.

1. Introduction

The Hidden Markov Model (HMM) is the most widely used technique in automatic speech recognition systems. However, various HMM models did not perform well enough in real-environment. One reason is that the acoustic models of an HMM-based classifier have many hidden factors such as speaker-specific characteristics that include gender types and speaking styles, and acoustic environments including communication channel characteristics and ambient noise.

Recent research has achieved multi-path model topologies that have several methods for different targets [1][2]. One is more appropriate for modeling pronunciation variations with syllable-length acoustic models than with context-dependent phonemes. Pronunciation variant-based multi-path HMMs were proposed for frequent syllables for which variation may be high. This approach combines knowledge-based and data-driven techniques by using phonetic knowledge to initialize the parallel paths of the syllable models [3]. Also, multi-condition HMMs have recently been proposed for dealing with various kinds of background noise [4]. Some methods have also been explored for data separation such as short- or long-duration phonemes [5]. Each path is trained to generate a multi-path model using isolated speech samples. At this time, the separation was performed through recognition experiments or data analysis in order to divide speech data.

Performance is improved by using a multi-path model. However, the number of states also increases, and parallel paths of the multi-path models increase the lexical confusability. Thus, the high variation states only perform the splitting to generate HMM topology with minimum number of states.

Several methods have been reported to reduce the total number of states that is applied the selective multi-path algorithm or a bottom-up state tying algorithm. Selective multi-path algorithm is based on likelihood analysis of database by automatic viterbi segmentation[6]. Generated multi-path models is defined that have another one-state path for fast and distorted part of speech, only for selected set of phones. However, these method are a determined model topology that is a three-state single-path or six-state multi-path at each phoneme. Although state splitting is necessary for a multi-path topology because of acoustic differences. However splitting does not require all the state sequences of phonemes. Therefore, selective state splitting is more suitable for constructing a non-uniform multi-path model where the state split in high acoustic variation case.

This paper investigates how to create a tied-state, non-uniform multi-path Hidden Markov network (HMnet) model that is a partially parallel state sequence. To automatically generate multi-path HMnet model, our research used an extended method based on Phonetic Decision Tree - Successive State Splitting (PDT-SSS) [7]. PDT-SSS clustering originally used the maximum likelihood criterion to choose the phonetic question with which each state was split. The SSS algorithm can create contextual and temporal tied-state topologies through state splitting in two domains. In this research, we propose three-domain state splitting methods to estimate contextual temporal and environment tied-state multi-path topologies. Our method adds a new environment domain that is applicable to distinguishing gender, speaking duration or background noise. In this paper, the environment domain is adopted gender to generates a separated male and female path model. This paper is organized as follows. Section 2 introduces various acoustic HMM topologies. Next, Section 3 describes the proposed multi-path HMnet model. In Section 4, we evaluate the performance of the HMM topology by phoneme recognition. Finally, we offer our conclusions in Section 5.

2. Acoustic HMM Topology

To create acoustic models of speech recognition using a large speech database, it is important to design the HMM topology as well as parameter estimation. Context-independent phonemes (monophones) have sufficient training data to generate robust models. However, monophones occurring in different contexts are different. Instead, using context-dependent phonemes (triphones) rather than monophones is known to provide higher recognition accuracy. So, continuous speech recognition systems employ a large number of triphones and continuous density HMMs.
An HMM topology using $N$ phoneme, $s$ states (each using $m$ mixtures), and $P$ multi-paths requires $(Nsm)P$ Gaussians for the models. In Table 1, a single-path monophone model with $N = 40$ phonemes including silence, $s = 3$ states, and $m = 1$ mixture requires 120 to 3840 Gaussians for models. There are 3 states 7378 phonemes for triphones, and an average of 4.5 states for extraphones. An extraphone with $P = 2$ paths and $m = 1$ mixtures thus requires more than 66,000 Gaussians. Here, the number of phonemes for triphones is varied by training database.

An HMnet model using PDT-SSS has a fixed number of states in spite of the number of phonemes because all phoneme states are shared. For example, the i-a-u phoneme is shown to be included in two paths in Fig. 1(d). The number of states in a multi-path model is proportional to the number of paths in monophones, triphones and extraphones. However, a multi-path HMnet model can be generated that adds only a small number of states; the total number of states used is generally 2000 to 5000.

3. Three-Domain Successive State Splitting

To automatically generate a multi-path topology, we used an expanded method based on the phonetic decision tree SSS. The PDT-SSS algorithm generates a suitable model topology of phoneme context classes and optimal parameters for HMMs commonly using successive state splitting with a question list.

Unlike the PDT-SSS algorithm that generates a single-path context-dependent tied-state model, the proposed algorithm constructs a multi-path context-dependent tied-state model with a suitable number of states by splitting three-domain in addition to environmental splitting. In this paper, the environment domain is applied to split male and female states for generating a suitable gender-dependent model. The environment domain is applicable to both gender and age, as well as region or microphone or noise conditions. Also, it is possible to split the utterance duration to express multi-style or the speed-variant nature of spontaneous speech.

Figure 1 illustrates various acoustic HMM model topologies with single-path and multi-path schemes. The monophone (Fig. 1(a)) and triphone (Fig. 1(b)) have the same three-state left-to-right model for all phonemes. An extended triphone (extraphone) topology selectively extends phoneme states that have high acoustic variation or a long duration (Fig. 1(c)). The HMnet model (Fig. 1(d)) can be used to solve an unseen triphone and to generate a suitable state model topology of the triphone types that are constructed by tying the state parameter. A phonetic-tied HMnet model enables state splitting for the contextual and temporal domains by PDT-SSS, which is modified from the original SSS. The PDT-SSS algorithm can generate the optimal HMnet topology according to the database analysis. However, the usual acoustic model has a single-path topology, which includes not only HMnet but also other models.

Single-path topology is limited to modeling using only Gaussian mixture components when a spontaneous speech database includes pronunciation variation and various speech styles. It is therefore necessary to create a multi-path topology for spontaneous speech if the gender or acoustic environment is not determined previously. A multi-path model has several HMM state sequences in parallel within a phone unit. During speech recognition, the HMM path that outputs the highest likelihood will be chosen for each phone by Viterbi processing.

Table 1: HMM configuration parameters (N phoneme $\times s$ stats $\times P$ paths).

<table>
<thead>
<tr>
<th>Model</th>
<th>Single-path state</th>
<th>Multi-path state</th>
<th>Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monophone</td>
<td>$40 \times 3 \times 1$</td>
<td>$40 \times 3 \times 2$</td>
<td>$1 \sim 32$</td>
</tr>
<tr>
<td>Triphone</td>
<td>$7378 \times 3 \times 1$</td>
<td>$7378 \times 3 \times 2$</td>
<td>$1 \sim 32$</td>
</tr>
<tr>
<td>Extraphone</td>
<td>$7378 \times 4.5 \times 1$</td>
<td>$7378 \times 4.5 \times 2$</td>
<td>$1 \sim 32$</td>
</tr>
<tr>
<td>PDT-HMnet</td>
<td>$500 \sim 7000 \times 1$</td>
<td>$500 \sim 8000 \times 1$</td>
<td>$1 \sim 32$</td>
</tr>
</tbody>
</table>
1) Train the gender separated context-dependent models.
2) Generate linguistic questions manually and prepare the prototype context-independent acoustic models.
3) Choose one state with maximum distribution and split the selected state into three domains. The distance is defined as the sum of Bhattacharyya distances between corresponding states.
4) Re-estimate the model topology after selecting the best splitting domain.
5) Return to step 3 until the predefined number of states or no more states to split.

The iteration is stopped when either there are no more states to split the gain in likelihood is smaller than a preset threshold or the desired number of states has been reached. After the construction of model topology, the models were then re-trained using Baum-Welch algorithm and the number of mixtures at each state increased to 32. Figure 3 illustrates example of the generated central phoneme “a” using proposed method. Each triphone has 3–4 states respectively, and it is multi-path topology which the same triphone is split into the path of male and female. For example, i-a-o triphone was generated gender dependent path which male is 1-2-3-5 state sequences and female is 1-4-5 state sequences in table 2. The other triphone shows gender independent path, example i-a-k triphone is 1-4-7 a path.

The proposed model performs better because the model creates only a necessary path, which is increased in likelihood after the state split, except for a low likelihood path and a path of insufficient training data. Also, the best performance is shown in the database in which the various environmental factors are included, because splitting is performed according to the necessary environment priority.

4. Experiments

4.1. Conditions

This section evaluates the performance of the proposed multi-path HMnet model through phoneme recognition experiments with a phone-pair model. The proposed models were compared to four conventional models (monophone, triphone, extriphone, and HMnet). The maximum state length is limited to six in the extriphone and HMnet topologies. We used 41 phonemes including one silence phoneme and one short-pause phoneme; a language model was not used. Each acoustic model was trained on the Japanese Newspaper Article Sentences (JNAS) corpus [8]. There were 100 male training speakers and 100 female training speakers (about 100 sentences per speaker); there were five male and five female speakers for testing. Table 3 lists the analysis conditions.

<table>
<thead>
<tr>
<th>Database</th>
<th>JNAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing</td>
<td>16 KHz sampling, 16 bits, 25 ms hamming window, 10 ms frame shift</td>
</tr>
<tr>
<td>Feature</td>
<td>12 MFCC + 12 delta MFCC + power + 1 delta power (total 26 dim.)</td>
</tr>
<tr>
<td>Time domain</td>
<td></td>
</tr>
<tr>
<td>Context domain</td>
<td></td>
</tr>
<tr>
<td>Environment domain</td>
<td></td>
</tr>
</tbody>
</table>

We use the Phoneme Accuracy Rate (PAR) defined in equation 1.

\[
PAR(\%) = \left( \frac{PhNo - Sub - Del - Ins}{PhNo} \right) \times 100
\]  

PhNo is the number of phonemes in the reference string, Sub is the number of phonemes substituted in the recognized string, Del is the number of phonemes deleted in the recognized string, and Ins is the number of phonemes falsely inserted into the recognized string.

4.2. Results

To demonstrate the effectiveness of the proposed model using three-domain splitting, we compare it with a conventional model and a multi-path HMnet model. Figure 4 presents the performances of the baseline model and the gender-separated multi-path model.
The results indicate that the PAR of the multi-path (MP) model increases faster than that of the single-path (SP) model for the monophone, triphone, and extraphone. Here, each extraphone adopted three to six states according to the likelihood. In case of monophone, it can confirm that the performance is improved model with all number of Gaussian mixture. For triphone and extraphone, the performance improvement can be demonstrated using the multi-path topology for the one to eight Gaussian mixture. However, the effect is decreased for 16 or more Gaussian mixture components. Also, an incremented number of states (22,252-→42747 - see Table 1) generates unnecessary computations for a 1% performance improvement.

![Figure 5: Comparison of phoneme recognition accuracy of various HMM models. (a) Tied-state Triphone. (b) Single-path HMMnet. (c) Multi-path HMMnet.](image)

Figure 5 compares the phoneme recognition accuracy of various HMMnet models. The maximum PAR of these models is near 83%. The MP-HMMnet model performs better than the Tied-triphone and SP-HMMnet models for the same number of states. Interestingly, multi-path models are more effective with one to four Gaussian mixtures. This can be attributed to the fact that the Multi-path HMMnet model does not require a high number of mixture components because the essential state performs the split first.

Table 4: Recognition accuracy of various HMMnet models for a 2000-state 4-mixture and a triphone of 22252-state 1-mixture.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triphone</td>
<td>67.73%</td>
</tr>
<tr>
<td>Tied-triphone</td>
<td>75.87%</td>
</tr>
<tr>
<td>SP-HMMnet</td>
<td>77.77%</td>
</tr>
<tr>
<td>MP-HMMnet</td>
<td>79.73%</td>
</tr>
</tbody>
</table>

Table 4 presents the recognition accuracies of various HMMnet and triphone models with less than 10000 Gaussian Output Probability Density Distribution (GOPDD) or the minimum number of GOPDD for real-time processing. The various HMMnet models have the same number of GOPDD (8000). However, the triphone must generate over 22,252 GOPDD even though the model uses a mixture. Using the MP-ExHMMnet model leads to a 9% phoneme error rate reduction for ExHMMnet and a 16% phoneme error rate reduction for HMMnet models with the same 2000-state 4-mixture topology. The triphone model achieves a 67.73% phoneme recognition accuracy using a 22,252-state 1-mixture topology. Our results indicate that the proposed multi-path model using contextual, time, and gender dimensional splitting methods improves the recognition accuracy compared to the HMMnet that uses only contextual domain splitting for the same number of GOPDD.

5. Conclusions

We proposed a multi-path generation method to improve the speech acoustic model. The proposed model created non-uniform, tied-state HMMnet topologies by three-domain splitting. Experimental results using phoneme recognition demonstrate that the proposed method can automatically create an appropriate model and obtain better performance, especially with a small number of mixture components, than the original method. We are planning to apply our method for generating a multi-duration model to represent the multiple styles or speed-variant nature of spontaneous speech.

6. Acknowledgement

This work was supported by Japan Society for Promotion of Science (JSPS) Research Grant.

7. References