Context Dependent Syllable Acoustic Model for Continuous Chinese Speech Recognition

Hao Wu, Xihong Wu
Speech and Hearing Research Center, State Key Laboratory on Machine Perception, Peking University, P.R.China
wuhaocis.pku.edu.cn, wxh@cis.pku.edu.cn

Abstract
The choice of basic modeling unit in building acoustic model for a continuous Mandarin speech recognition task is a very important issue [1]. Unlike traditional phoneme or Initial/Finals (IFs) units based acoustic modeling methods, which usually suffer from the limitations of less accuracy in modeling intra-syllable variations and long scale temporal dependencies, in this paper, a practicable syllable based approach is presented. In contrast with IFs, syllable can implicitly model the intra-syllable variations in good accuracy. Also, by carefully choosing context modeling schemes and parameter tying methods, syllable based acoustic model can capture longer temporal variations while keeping the complexity of model well controlled. Meanwhile, considering the data unbalanced problem, multiple sized unit model based approaches are also implemented in this research. The result experiment shows the acoustic model based on the presented syllable based approach is effective in improving the performance of the Chinese continuous speech recognition.

1. Introduction
Acoustic model is one of the most important knowledge sources for automatic speech recognition system, which represents acoustic features for phonetic units to be recognized. In building a acoustic model, one fundamental and important issue is choosing of basic modeling units [1]. Generally speaking, when the target language of the speech is specified, there are several types of subword phonetic unit can be used for acoustic modeling. Different acoustic model unit set can behaves dramatic differently on the performance.

Chinese is naturally a syllabic language and each basic language unit (Chinese character) can be phonetically represented by a syllable. Furthermore, each Chinese syllable also has Initial-Final structure. According to the official released scheme for Chinese phonetic alphabet, each syllable is regarded as the combination of these aspects are very helpful for the design of acoustic models.

Most Chinese speech recognition systems use Initial/Finals (IFs) as basic acoustic modeling unit set [2, 3, 4, 5, 6], which is mainly due to the low complexity of model and fair requirement on the amount of training data. There are only 59 standard IFs in Chinese altogether, even when tone is considered for finals, there is no more than 210 such units in the unit set. Also, property of the initial-final Chinese syllable structure can greatly decrease the real number of context dependent models in implantations.

Previous experiments have shown that the acoustic modeling based on IFs can achieves good performances on reading speech; however, when the evaluation scheme shifting to the spontaneous speech, the performance becomes unsatisfactory. In the case of spontaneous speech recognition, IFs based triphone models lack long scale temporal dependencies, and behave weak in capturing some effect in spontaneous speech such as reduction and skipping of phones etc. Moreover, when applying decision tree based parameter tying in a less accurate single Gaussian framework, inter and intra context modeling could become inaccurate.

Compared with IFs, syllable encompass phone clusters that contains the most strong co-articulation, which been considered as the most important cause of variations Mandarin speech. For this reason syllables usually have more stable acoustic realizations than initials and finals [7]. Besides of this, the construction of syllable in Chinese language is strictly regulated by a series of rules. For example, unless under very unusual conditions, the structures of all syllables are restricted to a few basic ones (e.g. V, VC, CV, and CVC ), successive consonants are not allowed within a syllable, and the permutation of the components in a syllable is restricted too. Such restrictions dramatically decreased the total number of syllables. Compared with the huge number of English Syllables (over 30,000), the number of Chinese syllables is less than 1,300 (with tone considered) [1]. At last, being a syllable centric language, Chinese has the property that every basic semantic unit, say, Chinese characters, has one or more corresponding single syllable pronunciations as its phonetic baseforms, so the syllables can be used to construct the pronunciation lexicon without bringing generalization problems. These properties make syllable a suitable acoustic modeling unit candidate for Chinese speech recognition system.

During the past decades, several researches have been made on syllable unit based Mandarin acoustic modeling[2, 5]. However, almost all these implementation are based on context independent modeling framework, and the performances of these implementations are hardly outgo those of the context dependent IF based ones. Although the idea of implementing context dependent syllable is intuitive and attractive, it was just impractical due to following reasons.

Firstly, to implement a syllable unit based context dependent acoustic model, a very large training corpus is needed, and the occurrences of syllables in the corpus should be even. These situations are hardly met on early released Mandarin corpora (such as 863 standard continuous speech corpus and HUB4 broadcast news corpus). In fact, these early released corpora contain only less than 150 hours speech altogether, and the occurrences of syllables in which are seriously uneven. For ex-
ample, the 863 corpus contains about 110 hours speech, and the transcriptions of utterances are severely repeated. Statistics shows on 863 corpus transcription shows that some high-frequency syllables (such as "[d_0]") hold more than 5% of occurrence, while there are some syllables only hold less than 0.001% in the whole corpus. These facts brought crucial problems in model training (especially on forward and backward algorithm derived from maximum likelihood criteria).

Secondly, although usually being a very efficient method to improve the model precision, standard triphone based context dependent modeling method is impractical for syllable based acoustic models. A common syllable set (with tones and natural tone considered) usually contains about 1,200 to 1,500 units (1,429 in our case, according to pronunciation lexicon), which may derives a tri-syllable set holding more than 1,700,000,000 context dependent units. Even a toneless syllable unit (usually contains about 400 syllables, 413 in our cases) may derives a context dependent unit set that contains more than 64,000,000 tri-syllable units. With this, heavy problems are brought to modeling and decoding implementation issues of model management, indexing and parameter clustering. Even with today’s computer hardware systems, these problems hardly be solved.

Fortunately, problems mentioned above have been partly solve by the recently works on speech corpora collection and progresses on computer hardware/software systems. Some lately released standard Mandarin speech corpora contains well portioned huge amount of speech data. For example, the Intel database (released around 2003), contains over 250 hours speech by both male speakers and female speakers; the recently developed SAMSUNG database collected by 4 major institutes in Peking, contains over 200 hours speech by about 1,000 speakers; another corpus, the 863 phase II database contains 150 hours speech. Also, by carefully designed transcriptions ensure these corpora usually have better consideration on syllable coverage than old corpora. With these progresses, it becomes possible to make a context dependent syllable based model been sufficiently trained.

In this paper, we carried out preliminary implementations and experiments of using syllable as the basic unit of acoustic model. While the insufficiency of training data is solved by and large, there are still practical problems of handling huge context dependent unit set. Besides, although been alleviate in the new training corpus, the unevenness of syllable is still a considerable issue. To solve the first problem, we took a intuitive solution of simplifying the left and right contexts from syllable to shorter unit; to further solve the problem of data unevenness problem, multiple sized units set based model was implemented as one tradeoff scheme between complete syllable units model and complete initial-final units model.

The rest part of the paper will give more detailed description on our implementations and experiments. In section 2, the proposed models will be discussed from two aspects: the design of model unit set and the specifying of context for syllable units. In section 3, the training procedure is briefly described; In section 4, the result of experiments was presented and discussed; And the conclusions and future works will be addressed in section 5.

2. Definition of unit sets and context modeling

In this work, five unit sets were selected as acoustic modeling unit sets including: the baseline initial-final set (ITF-BL) set; complete tonal syllable set (TS); complete toneless syllable set (TLS), and multiple sized unit sets with ITFs and 40 or 100 high frequency syllables (H40 / H100). The detailed configuration and corresponding context modeling scheme of each unit set will be introduced in following paragraphs.

2.1. Initial-tonal final baseline unit set (ITF-BL)

Initial-tonal final set has been the dominant unit for Mandarin speech acoustic modeling in recent years. In this work, an ITFs set used as the baseline unit set.

By appending four dummy-initial ("ga", "go", "ge", "ger"), two semivowel ("y", "w") and some natural-pronounced finals (on which tone is marked as "0") to the standard ITF set, ITF-BL set includes 27 initials and 187 tonal finals. The detailed category of unit is listed in table 1.

<table>
<thead>
<tr>
<th>Unit type</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>b c h d f g ga ge ger go h j k l m n p q r s t w x y z zh</td>
</tr>
<tr>
<td>Tonal Final</td>
<td>a0 a1 a2 a3 a4 a0 a1 a2 a3 a4 a0 a1 a2 a3 a4 a0 a1 a2 a3 a4 a0 a1</td>
</tr>
<tr>
<td>Others</td>
<td>silence</td>
</tr>
</tbody>
</table>

The context dependency modeling method used in this unit set follows the standard triphone framework. After appending dummy initial and semivowel into unit set, each triphones has either a initial-tonal final-initial pattern or a tonal-final-initial-tonal-final pattern, the final model set includes over 344,000 triphones.

The left-to-right model topology was used to represent every ITF, and one state skipping allowed within each model. The number of states in each model was set to 2~3 for initials and 4~5 for tonal finals, according to the corresponding phonetic structures.

2.2. Tonal syllable set

A complete tonal syllable set was distilled from the transcription of the training set. The set includes 1,429 toned syllables. This syllable set can meet the dictionary on the whole. Yet there are a few syllables in dictionary such as "[L"le3]" did not appear in the train set. When encountering this situations, each missing syllable in dictionary is replaced by the closest one.

Left and right context was simplified to reduce the number of context dependent units. Considering that every Chinese syllable have fixed "initial-final" structure, the left context can be reduced from previous syllable to correspondent final part (tone is discarded); this decreased the number of left context from 413 to 38 (37 finals including 2 variations of final 1 and silence), and right context can be reduced from following syllable set to correspondent initial set, which makes a reduction of
right context from 413 to 28 (27 initials includes dummy initial, and silence). With this simplification, the total number of context dependent toned syllable units in the set dramatically decreased from 243,743,101 (413 × 1, 429 × 413) to 1,520,456 (28 × 1, 429 × 38). The complete initial-final context set is shown in Table 2.

Table 2: Definition of toneless initial-final context set

<table>
<thead>
<tr>
<th>Unit type</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>a b c d e f g h i j k l m n o p q r s t u v w x y z</td>
</tr>
<tr>
<td>Toneless Final</td>
<td>a a i an ang ao e ei en eng er i ia ian iao ia o ie if in</td>
</tr>
<tr>
<td></td>
<td>ing iong iu o uong ou u uu uan uang ui un uo v van</td>
</tr>
<tr>
<td>Others</td>
<td>silence</td>
</tr>
</tbody>
</table>

The left-to-right model topology was used to represent every syllable, and one state skipping allowed for state propagating within the model. The number of states in each model was set as same as the summation of state numbers of corresponding initial and final.

2.3. Toneless syllable set

The toneless syllable set is also distilled from the transcription of the training set but with a toneless version. This set consisted of 413 syllables. It was observed that all tone-less syllable in the lexicon has been included in these set. And the statistics of shows that training samples for every unit are more even portioned in the training corpus, compared with that in the case of toned syllable unit set.

Toneless syllable uses same context set as tonal syllable unit. The simplified context set again decreased the number of total context dependent units from 70,444,997 (413 × 1, 429 × 413) to 439,432 (28 × 1, 429 × 38).

The topology of each toneless syllable model was set as same as the tonal ones.

2.4. Multiple sized unit set with IFs and high frequency syllables

While syllable based unit set have significant advantage over the sub-syllable ones, it still suffers from the problems of data unevenness. Statistics in training corpus shows that when a unit has occurred too few, both state tying and tied model parameter estimation becomes unreliable. To solve this problem, we proposed a multiple sized unit set as the tradeoff of complete syllable set and initial-final set. Taking initial-final set as the basic unit set, we added most frequently occurred syllables (according to the statistics of training corpus) into the set. By this compromise, we expect that both the original IF units and the complementary high frequency unit can get enough training data and be well trained.

Two multiple set was designed. The first set includes the initial-final units and top 40 high frequently occurred syllables. These 40 syllables hold about 25% occurrences of syllables in the whole corpus. The second set includes the initial-final units and top 100 high frequently occurred syllables. These 100 syllables hold over 50% occurrences of syllables in the whole corpus.

The definition of context set is coherent with previous several unit set, so is the topology of models.

3. Training of Acoustic Models

In this section the detailed implementation of acoustic models will be introduced. To compare the performance of different units, a uniform framework was used to train all the acoustic models. In this training framework, acoustic model unit is represented by continuous mixture density HMM, Baum-Welch algorithm is used to estimate the parameters of the model, and decision tree based state tying technique is used to reduce the total number of parameters in model [1].

3.1. TRAINING CORPORA

Training of syllable based acoustic models requires a large amount of data. In our experiments, all models are trained with a training data set which contains about 360 hours speech from over 750 male speakers, which is picked up from three widely used continuous Mandarin speech corpora: The 863-I, 863-II and Intel corpus. The brief information about these three speech corpora is listed in Table 3.

Table 3: Training corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Speakers</th>
<th>Amount of Speech (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>863-I (male)</td>
<td>83</td>
<td>56.67</td>
</tr>
<tr>
<td>863-II(male)</td>
<td>120</td>
<td>78.08</td>
</tr>
<tr>
<td>Intel (male)</td>
<td>556</td>
<td>227.30</td>
</tr>
<tr>
<td>total</td>
<td>759</td>
<td>362.05</td>
</tr>
</tbody>
</table>

Most speech in these three corpora is reading style speech. Still, since the amount of data is quite large, we can expect that most inter/intra syllable variations could occurs in the training set and be captured in acoustic model.

3.2. Acoustic Model training

3.2.1. Features extracting

The acoustic features here used are the Mel-frequency cepstral coefficients (MFCCs), which includes 12 cepstral coefficients together with the logarithmic energy. Instead of generally used standard mean and variance normalization, histogram equalization based cepstral normalization was applied to compensate the convolving and additive noise. By including the first and second derivatives of the parameters, the final feature vectors have 39 elements.

3.2.2. Training procedure

In this section the detailed implementation procedure of acoustic models will be introduced. To compare the performance of different units, a uniform framework was used to train all the acoustic models.

In this training framework, acoustic model unit is represented by continuous mixture density HMM, Baum-Welch algorithm is used to estimate the parameters of the model, and decision tree based state tying technique is used to reduce the total number of parameters in model. In this section the detailed implementation of acoustic models will be introduced. To compare the performance of different units, a uniform framework was used to train all the acoustic models. In this training framework, acoustic model unit is represented by continuous mixture density HMM.

The whole training process consists of four phases:
1) **Context independent (CI) model training:** In this phase, context independent HMM model for each unit is trained by Baum-Welch algorithm, which is used to give initializations to context dependent unit models.

2) **Untied context dependent (untied-CD) model training:**

   During this phase, a set of untied HMM was trained, which includes all the context dependent units which have occurred at least 3 times in the training corpus. Single Gaussian is used as the emitting distribution for each HMM state, and probabilistic count of each untied state is recorded during the training process.

3) **Decision tree based state tying:** During this phase, a decision tree [3, 8] is build for each CI HMM states. For each node in decision trees, the corresponding impurity function is computed with the distributions and statistics of involved untied states, which was gathered in phase 2. Number of tied state was set to 8,000 for initial-final model, and 10,000 for syllable based models. For multiple sized unit models, the tied number is chosen to make the number of tied states in initial-final models same as that of baseline initial-final model, with this consideration, the tied state number is set to 10,000 for H40 model and 12,000 for H100.

4) **Tied context dependent (untied-CD) model training:**

   The tied model is then trained under Baum-Welch framework, and at this stage, the emitting distribution of each state was split from single Gaussian to Gaussian mixture distribution to improve the precision of model. For initial-final models, the number of components for each state is set to 32, while for syllable models the corresponding component number is set to 24. This makes the models to have same total parameter numbers. For multiple sized unit models, the component number is set to 32.

### 4. Experimental Results

The proposed models is evaluated on the male speech from the test set of 2004 continuous Chinese speech recognition evaluation held by 863 project[9]. This set includes 100 sentences, uttered by 10 male speakers, each speaking 10 sentences. Each sentence includes 15 to 45 Chinese characters, the speech in the whole test set sum up to 15 minutes, includes approximately 2,700 Chinese characters.

Pulsar 1.0, which is the large vocabulary continuous speech recognition system developed in Speech and Hearing Research Center, Peking University, was used to evaluate the performance on the test set. A trigram based language model, which include 64,277 Chinese words, was trained on CMU LM tools and used in evaluation.

The result of experiment is shown in table 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Corr(%)</th>
<th>Sub(%)</th>
<th>Del(%)</th>
<th>Ins(%)</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>64.71</td>
<td>33.28</td>
<td>2.01</td>
<td>2.98</td>
<td>38.27</td>
</tr>
<tr>
<td>TS</td>
<td>62.26</td>
<td>34.43</td>
<td>3.30</td>
<td>2.30</td>
<td>40.04</td>
</tr>
<tr>
<td>TLS</td>
<td>64.79</td>
<td>33.17</td>
<td>2.04</td>
<td>2.39</td>
<td>39.60</td>
</tr>
<tr>
<td>H40</td>
<td>65.74</td>
<td>30.95</td>
<td>3.30</td>
<td>1.95</td>
<td>36.21</td>
</tr>
<tr>
<td>H100</td>
<td>64.06</td>
<td>32.96</td>
<td>2.98</td>
<td>2.04</td>
<td>37.98</td>
</tr>
</tbody>
</table>

In our preliminary experiments, the tonal syllable model did not show advantage over initial-final models, while toneless syllables, even lacking of tonal information, still achieved slightly better performance than baseline system on word accuracy rate. This fact implies that while syllable based unit has great potential on improving the precision of model, data unevenness is still the key problem to be handled.

Furthermore, in these preliminary implementations, one path left-to-right topology is used for syllable models. Such simple model is not suitable for presenting the long scale temporal dependency, thus lost a major advantage of using syllable units. This issue is going to be considered in our future works.

Multiple sized model shows good performance when containing 40 high frequency syllables. However, when more high frequency syllables were added to model, the performance becomes worse. This may contribute to the fact that when more syllable added to model, training data for corresponding initial and final models are reduced. This problem is especially serious when the high frequency took very large portion of occurrences for corresponding initial and final.

### 5. Conclusion And Future Work

In this paper, we carried out several preliminary implementation and corresponding experiments on syllable based context dependent Chinese acoustic model. The results shows that while syllable based unit set does have great potential of increase the performance of Chinese acoustic model, there are still problems to be handled.

While the above works are still preliminary and coarse, there is several works need to be done in next phase. First, better training corpus is expected to used in model training, and data reusing methods should be considered in multiple sized unit model training. Than, better context cluster method may be used. Better syllable selection methods for multiple sized unit model also should be considered. At last, the training and experiment in this paper is still carried out on a reading style speech corpus. With more spontaneous style speech corpora available, the research in this paper is expected to be extended.

### 6. References


