A Hybrid Approach to Enhance Task Portability of Acoustic Models in Chinese Speech Recognition

Jin-Song Zhang, Shu-Wa Zhang, Yoshinori Sagisaka and Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
{jzhang, szhang, sagisaka, nakamura}@slt.atr.co.jp

Abstract

This paper presents our approach to enhance the portability of acoustic models by mitigating the phonetic mismatch arising from a new testing task which is rather different from the training data. The approach is a hybrid one which combines knowledge-based context categorization to generate a context rich set of subword units, and data-driven-based acoustic model clustering on the level of context category. Compared with the conventional approach of only phonetic decision tree based model clustering and unseen model generation, the new approach improved greatly the desired subword coverage for the new testing domain, and achieved an error rate reduction by 10.8% for Chinese character accuracy in the recognition experiments. Together with the effect of the newly adopted basic units of 9 glottal stops, we achieved a total 23.5% error rate reduction in the testing compared to the baseline system.

1. Introduction

It is well-known that speech recognition system tends to degrade its performance when it is ported to a new testing task where exists considerable variabilities from the acoustic model development condition[1,2]. Among a number of variabilities from the different task specifications, phonetic context mismatch may not only lead to a low coverage of desired subword unit in the training data, but also result in biased acoustic models when the training data consists of repeated utterances over limited common scripts[2].

One efficient solution to the problem is data-sharing for the phones with similar phonetic context. Phonetic-decision-tree (PDT) based context clustering is the most popular data-driven method for training context dependent (CD) HMMs[1,5]. It is also regarded as one good method to solve the problem of unseen CD units. However, one easily neglected problem is that the data-driven nature of the PDT-based unseen model generation is less likely to generate robust models for unseen units as there are no data available. Also the data-driven developed models may suffer from lack of generalization when the variabilities (including phonetic context) in the training data are limited. Although these problems do not affect the performance of a task dependent (TDEP) recognition, it probably takes effects in a task independent (TIND) recognition where exists a notable phonetic mismatch from the training data.

This was just the situation we met when we develop the Mandarin Chinese speech recognizer for a hotel reservation dialogue system. What we have for acoustic model development is the 863 Putonghua database, acquired from mainland China, which contains the utterances of about 200 sexual-balanced speakers reading several common scripts from newspapers. Our task is to recognize daily speech conversations between customers and hotel receptionists for room reservations. When we use the traditional demisyllabic units: Initials and Finals (I/F), as the subword units[3,4,5,6], and develop triplet units (analogous to triphone), there are serious sparse training problem due to the vocabulary differences: a check of only 5 dialogs with 237 utterances showed over than 25% desired units have no samples in the training data. For the seen units, they also may be biased to the context environments of the limited newspaper sentences in 863 data.

In order to deal with the above-mentioned problems, we developed a hybrid approach in which we combined explicit-knowledge-based merge of subword units and data-driven-based model development. The power of reliable acoustic knowledge was exploited significantly in the first step: after unit design and context categorization, we reduced the size of triplet unit set from an original 111k to about 5.3k before any look into the data. This greatly increased the desired unit coverage for the new testing in the training data. The high degree of data sharing resulting from the knowledge-based unit merge is also assumed to improve the generalization of the models as the phonetic context environments became more generalized compared to the original ones. Then we apply the PDT to model clustering to exploit the efficiency of data-driven method. Since the context environments have changed from single phones into phone categories, it is necessary to design specifically phonetic questions appropriate for them.

The following sections are arranged as follows: after introduction of the problem of too many triplet units based on Initials/Finals, and the potential problem of PDT-based unseen model generation, we describe in details our knowledge-based context categorization and the design method of phonetic questions. Then we introduce the experimental set-up and results. Finally we give some discussions.

2. Number of I/F Triplet Units

A Chinese word is composed of one to several characters, and each character is pronounced as a monosyllable with a pitch tone. The totally phonetically differentiable tonal syllables are about 1,300, and the number of base syllables is about 410 when pitch tones are discarded. Traditional Chinese phonology [6] divides the base syllable into demi-syllabic units: an Initial and a Final. An Initial is a consonant, and a Final may be a vowel, a diphthong, or a vowel compound with a nasal ending. Besides, there are also 35 null-Initial syllables with only Finals.

Since a vowel in a compound Final is systematically different from that in isolation, Initial/Final based Chinese speech recognition systems were found to perform better than phone based systems [3,5]. We also use Initials and Finals as
the basic unit set. 21 Initials, 37 Finals (IFs) and a /sil/ for silence, were used to represent a 412 base syllable list in our speech recognition system. With this unit set, we found that the possible number of triplet units are as large as 111,625, including triplet Initials and Finals.

\[
\begin{array}{ccc}
21 \text{ Is} & 37 \text{ Fs} & 21 \text{ Is} \\
\text{or} & 35 \text{ Fs} & \text{or} \\
\text{or} & 1 \text{ sil} & \text{or} \\
\text{or} & 1 \text{ sil} & \text{or} \\
\end{array}
\]

syllable 1 syllable 2 syllable 3

Figure 1: The possible context environments of tri-syllable words for building triplet units, where 35 Fs indicates the 35 Final-only null-Initial syllables.

The calculation of the possible number of triplet units can be aided with the illustration in Figure 1. When we generate triplet Initials, we focus on the context of the Initial of syllable 2. The possible left contexts include all the possibilities in the ending segment of syllable 1, and the right contexts are the Finals inside syllable 2. Hence, the number is caculated as,

\[
(37 \text{ Fs} + 1 \text{ sil}) \times (412 - 35 \text{ Fs}) = 14,326
\]

where (412-35Fs) depicts the possible right context dependent intra-syllable Initials. For counting the triplet Finals, we focus on the Finals of syllable 2. The right contexts include all the possible beginning segment of syllable 3, and the left contexts are different when syllable 2 has a Initial or not. If there is one, the left context includes the 21 Initials, and if no, the ending segment of syllable 1 must be taken into consideration.

\[
(412 - 35Fs) \times (21 Is + 1 sil + 35Fs) + (37Fs + 1sil) \times 35Fs \times (21Is + 35Fs + 1sil) = 97,299
\]

where (412-35Fs) indicates the intra-syllable left dependent Finals. It is impossible to get enough data for robust estimation for all these over than 111k units separately.

3. PDT-based synthesis of unseen unit

Phonetic-decision-tree (PDT) is widely used to solve the sparse data problem in model training. It is regarded as one good method to solve the problem of unseen CD models by generating the unseen CD models through traversing a binary tree according to the phonetic question in each node [1].

Figure 2: Illustration of decision tree based synthesis of unseen units.

However, the robustness of the generation is not guaranteed. As illustrated in Fig. 2, the center state of the unseen CD Initial /a-s+s+u/ is synthesized using the parameters from /a-s+i/. This result is unwanted in that the /s/ in /a-s+s+u/ is probably produced with lip-rounding due to the articulatory anticipation for the succeeding /u/, whereas the /s/ in /a-s+i/ is not associated with the configuration. It may be more preferable to generate the state for /a-s+s+u/ using that from /u-s+s+u/, in which the /s/ similarly has lip-rounding. Since the tree is developed based on the data of seen CD units and a number of phonetic questions, the variability in the speech data and multiple choices of phonetic questions easily lead to different tree structures in different situations such as different training data or different phonetic questions used. As a result, consistent and robust synthesizes of unseen CD phones cannot be guaranteed in all cases.

4. Hybrid approach for model clustering

The hybrid approach to cluster models includes three steps: firstly, 9 glottal stops are adopted for a consideration of Chinese acoustics of null-Initial syllables. Secondly, based on the acoustic knowledge, coarticulaty effects of 21 Initials are categorized into 8 groups, and those of 37 Finals are categorized into 10 carryover and 12 anticipation groups. Thirdly, knowledge-based merged units are further data-driven clustered by the PDT method during model development.

4.1. Adoption of 9 glottal stops

Phonetic study shows that syllable boundary of null-Initial syllables are mainly manifested by glottal stops [6]. Speech synthesis also revealed that negligence of the glottal stops let the synthetic output for sequences such as ă-șă-ț (yi wu yi shi) difficult to be understood [7]. Since the glottal stops are produced by a momentary closure of the vocal folds, and approximately inherit the formant patterns of succeeding vowels. We adopt 9 symbols to represent the glottal stops preceding the 35 null-Initial syllables with different vowel onset (Table 1). A side effect of this adoption is that the number of possible triplet IF units can be reduced to about 27k ones, about 24.3% of the original 11k ones, as only bi-syllable combinations are needed to taken into account.

| 1.A1 | 1.hA1 | a |
| 2.A2 | 2.hA2 | ai, an |
| 3.A3 | 2.hA2 | ao, ang |
| 4.E1 | 3.hE1 | e, er |
| 5.E2 | 4.hE2 | ei |
| 6.E3 | 4.hE2 | En |
| 7.E4 | 4.hE2 | Eng |
| 8.O1 | 5.hO1 | O |
| 9.O2 | 6.hO2 | Ou |
| 10.V | 7.hV | v, van, ve, vn |
| 11.U | 8.hU | u, uu, uai, uan, uang, ui, un, iao, ione |
| 12.I | 9.hI | i1, i2, i3, ia, ian, iang, ia, iao, ije, in, ing, iu, ione |

Table 1. Anticipation groups of the 37 Finals. The second column gives the 9 glottal stops for the corresponding null-Initial Final-only syllables.

4.2. Knowledge based unit merge

Articulatory configurations are known to exert the major influences on the spectral patterns of phones. The phones which differ only on production manners approximately arouse the same formant transitions into its neighboring sounds. Therefore, from the view of coarticulatory effect, those phones can be categorized into the same context group. Several
previous studies have applied this idea to Chinese speech recognition [1,4].

We used this idea in the second step of our whole approach for two reasons: one is to reduce the number of possible triplet units and enhance the subword unit coverage of the testing in the training data, and the other is to generalize the phonetic context of one unit as the training database we used has limited context environments.

<table>
<thead>
<tr>
<th>Initials</th>
<th>Consonant Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a, ia, ua</td>
</tr>
<tr>
<td>E</td>
<td>e, ie, ve</td>
</tr>
<tr>
<td>I</td>
<td>ai, ei, uai, ui</td>
</tr>
<tr>
<td>I2</td>
<td>i1, i2, i3</td>
</tr>
<tr>
<td>V</td>
<td>v</td>
</tr>
<tr>
<td>U</td>
<td>u, ou, iu, ao, iao</td>
</tr>
<tr>
<td>O</td>
<td>o, uo</td>
</tr>
<tr>
<td>R</td>
<td>er</td>
</tr>
<tr>
<td>G</td>
<td>iang, ing, uang, eng, ang, iong</td>
</tr>
<tr>
<td>N</td>
<td>ian, in, uan, un, en, an, van, vn</td>
</tr>
</tbody>
</table>

Table 2. Carryover groups of 37 Finals.

Categorization of the coarticulatory effects needs to consider the carryover and anticipation effects of one unit respectively, as a Final may have different articulatory configurations at its onset and offset. Table 1 gives our categorization of anticipation effect of 37 Finals into 12 groups according to the onset articulatory configurations, and Table 2 gives 10 groups for the carryover effects according to offset configurations. Coarticulatory effects of Initials are categorized according to the constriction places that may range from individual to major places. We categorized the 21 Initials into 8 groups as in Table 3.

<table>
<thead>
<tr>
<th>Labial</th>
<th>1. Bilabial</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Labiodental</td>
<td>f</td>
</tr>
<tr>
<td>3. Apical dental</td>
<td>z, c, s</td>
</tr>
<tr>
<td>Coronal</td>
<td>4. Apical alveolar I</td>
</tr>
<tr>
<td>5. Apical alveolar II</td>
<td>n, l</td>
</tr>
<tr>
<td>Sub-apical</td>
<td>6. Sub-apical palatal</td>
</tr>
<tr>
<td>Dorsal</td>
<td>7. Palatal</td>
</tr>
<tr>
<td>8. velar</td>
<td>j, q, x</td>
</tr>
</tbody>
</table>

Table 3. Hierarchical configuration structure of 21 Initial consonants.

After the context effects are categorized, the possible triplet I/F units can be merged when they have the same right and left context categories. Finally, we got a subword unit set of 5.3k context-merged triplet I/F units, which is only about 4.8% of the original 111k ones.

### 4.3. PDT based model clustering

When we build acoustic models for the subword units, we applied the phonetic decision tree to the tying of HMM states in order to improve the data sharing of HMMs via data-driven method. Since the subword units are already knowledge-based merged units, specific phonetic questions should be designed for the context categories. The questions asked for the Initial groups are about the constriction place in the midsagittal plane and tongue movements. For example, “Labial-dental” indicates the Initials constricted in the front places including both labial and dental places. Tongue blade includes the consonant groups of palatal and sub-apical palatal. The questions asked for the Final categories are much like the normal ones for the phone targets, such as high-vowel, low-vowel, lip rounding, and etc.

### 5. Experiments and Results

#### 5.1. Experimental setup

Our system is a speaker independent, intra-word triplet unit, decision-tree based tied-state system. Each I/F unit is modeled by a left-to-right, three-emitting-state continuous density HMM without state skipping. Glottal stops are modeled by a one state HMM, and silence is modeled by two models: a three-emitting-state HMM and a one-shared-state HMM allowing state skipping.

Input speech, sampled at 16kHz, was initially pre-emphasized 1-0.97z⁻¹ and grouped into frames of 25ms with 10ms frame shift. For each frame, a Hamming window was applied followed by the computation of 12 MFCCs and normalized log-scaled energy. The first order time derivatives were added to the feature. Thus each speech frame was represented by a vector of 26 features.

The vocabulary consists of 9,754 words and the language model is a bi-gram word model trained from the mixed data of newspaper transcripts and Hotel-Reservation dialogue sentences, totally about 300k words. The recognition performance is evaluated according to the accurate rate of Chinese characters.

#### 5.2. Training and testing Data

We used all the data available at the 863 Putonghua corpus as the training data, including utterances of 83 male and 83 female speakers. We collected two testing databases to test the developed acoustic models. Two of the authors, both male native Chinese speakers, each read a different set of 650 sentences from the 863 scripts. This database is used for the task-dependent (TDEP) testing. The same speakers read another script of five Hotel Reservation clerk-customer dialogues (HRD) consisting of 237 sentences. Since the script is quite different from the newspaper style, this data is used as the task-independent (TIND) testing database.

Figure 3: The rates of unseen units for HRD testing. The left block in each group represent the unit kinds, and the right one for their accumulated occurrences in HRD.

#### 5.3. Subword unit coverage

The phonetic context mismatch of the two tasks can be examined by a look at the coverage of HRD units in the training data. Based on the lexicon, we generated the triplet I/F units from the 863 and HRD transcripts in three ways: A) the conventional triplet I/Fs, B) triplet I/Fs with glottal stops adopted, C) triplet I/Fs after the context effect categorized based on knowledge. Figure 3 shows the rates of the unit kinds.
and their accumulated occurrences to all those appeared in HRD data when they have no samples in the training data for the three different unit sets.

5.4. Speech recognition results

We developed HMMs for the three unit sets, each having 2.5k PDT-based tied states and 5 mixtures in each state. Unseen models are generated by PDT during model development. Recognition results are given in Table 4.

- Triplet A: The conventional triplet I/F HMMs. There are 9.6k triplets from the lexicon. A rich phonetic question set was designed to have a total of 232 right or left context pertaining questions. This serves the baseline system.
- Triplet B: Triplet I/F HMMs after glottal stops adopted. 8.8k triplets were generated from the lexicon. 14 extra questions about the glottal stops were added to the previous 232 question set for PDT-based HMM tying.
- Triplet C: Triplet I/F HMMs with context merged. 3.0k units were generated from the lexicon. A total of 94 questions were designed for PDT-based HMM tying.

<table>
<thead>
<tr>
<th>HMM</th>
<th>TDEP Decrease of CER</th>
<th>TIND Decrease of CER</th>
<th>TDEP – TIND</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8.1%</td>
<td>11.9%</td>
<td>-3.8%</td>
</tr>
<tr>
<td>B</td>
<td>6.9%</td>
<td>1.2%</td>
<td>10.2%</td>
</tr>
<tr>
<td>C</td>
<td>7.2%</td>
<td>-0.3%</td>
<td>9.1%</td>
</tr>
</tbody>
</table>

Table 4. TDEP and TIND testing results from the three HMM sets. Recognition results are evaluated in the average Chinese character error rates (CER).

6. Discussion

When we pay a check at the unseen unit rates in Figure 3, we can see the following phenomena:

1. In the case of conventional I/F triplet set, there is a considerable phonetic mismatch between the HRD testing and the training data. Over than 25% of the desired unit kinds do not appear once in the training data, and their accumulated occurrences amount to 15% of the HRD testing data.
2. After the adoption of the glottal stops to the basic unit set, there is about 5% decrease in the unseen units rates for the HRD testing compared to triplet set A.
3. Unseen unit rates were greatly reduced when the triplet set C is used, indicating it as a rich context representative unit set. Based on this unit set, the unseen units only account for less than 5% of the HRD testing data.

From the recognition results in Table 4, we see:

1. In the case of the baseline model, TIND testing got 3.8% (or a relative 46.9%) more errors than the TDEP testing. This indicates that the phonetic mismatch between the two tasks did decrease the portability of the acoustic models.
2. Adoption of the I-emitting-state HMMs for glottal stops reduced the error rates by 1.2% (a relative 14.8%) in TDEP testing and by 1.7% (a relative 14.3%) in the TIND testing when compared to the baseline approach, suggesting its efficiency.
3. However, TIND still has 3.3% (corresponding to a relative 47.8%) more errors than the TDEP testing for the B set HMMs. This phenomenon is a little hard to understand as we once expected a decrease in the gap between the TIND and TDEP testings, because the unseen unit rates have decreased by 5% compared to the baseline. The result might be explainable when we regard it resulting from model generalization. Since the training data has limited phonetic context environments, the learned models tend to be biased to the context specific to not only neighboring units but also other ones such as prosodic environments. The lack of generalization affects the recognition even though the unit coverage is to some extent improved.

4. The C set HMMs increased slightly the recognition error by 0.3% (a relative 4%) in the TDEP testing compared to the B set of HMMs. But it still achieved 0.9% (a relative 11.1%) less errors than the baseline model. In the TIND testing, this set HMMs degraded the error rates by 0.9% (a relative 10.8%) compared to B set, and by 2.8% (a relative 23.5%) compared to the baseline A set. Thus, it achieved the best portability for the HRD task among the three HMM sets. The gap between TIND and TDEP testing is also reduced to 26.4%, much less than those of the other two approaches, indicating its better generalization property.

5. Although the unseen units only stands for less than 5% of the TIND data, the error rate is still 26.4% more for TIND than TDEP testing. This indicates some domain specific training data may still be necessary to calibrate the acoustic models in order to achieve the similar performance of TDEP testing.

7. Conclusion

We presented here our hybrid approach to develop a context-rich subword unit set for Chinese speech recognition. Experiments showed that the approach can not only theoretically improve greatly the subword coverage rate in the training data for a TIND testing condition, but also reduced by a total of 23.5% error rate for the testing when compared to the conventional approach.

Acknowledgement

We appreciate very much Dr. Seiichi Yamamoto, the head of Spoken Language Translation Laboratories, for his kind support to this study.

References