Modeling Pronunciation Variation Using Context-Dependent Weighting and B/S Refined Acoustic Modeling

Fang Zheng †, Zhanjiang Song ‡, Pascale Fung †, William Byrne ¶

† Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. ‡ Univ. of Science and Technology, Hong Kong. ¶ CLSP, The Johns Hopkins University, USA.

Abstract

The pronunciation variability is an important issue that must be faced with when developing practical automatic spontaneous speech recognition systems. By studying the initial/final (IF) characteristics of Chinese language and developing the Bayesian equation, we propose the concepts of generalized initial/final (GIF) and generalized syllable (GS), the GIF modeling method and the IF-GIF modeling method, as well as the context-dependent pronunciation weighting method. By using these approaches, the IF-GIF modeling reduces the Chinese syllable error rate (SER) by 6.3% and 4.2% compared with the GIF modeling and IF modeling respectively when the language modeling, such as syllable or word N-gram, is not used.

1. Introduction

When people speak casually in daily life, it is quite common to find many different pronunciations of individual words, which may make speech recognizer’s performance degrade a lot. In casual speech, phone change and sound change phenomena are common, and there are often sound or phone deletions and/or insertions. These problems are made especially severe in Mandarin casual speech since most Chinese are non-native Mandarin speakers, and there is an even larger pronunciation variation due to the influence of speakers' native language.

To find a solution to this problem in acoustic modeling stage, a speech recognition unit (SRU) set should be well defined so that it can well describe the phone/sound changes, including insertions and deletions. An annotated spontaneous speech corpus should also be available, which at least has the base form (canonical) and surface form (actual) strings of SRUs.

In this paper, the Chinese Annotated Spontaneous Speech (CASS) corpus will be briefly introduced in Section 2. Based on the transcription and statistics of CASS corpus, the generalized initials/finals (GIFs) is proposed to be the SRUs in Section 3. In the following section, we construct the framework for the pronunciation modeling, where an adaptation method is used for the refined acoustic modeling and a context-dependent weighting method is used to estimate the output probability of any surface form given its corresponding base form. Section 5 lists the experimental results while summaries are given in Section 6.

2. CASS Corpus

A Chinese Annotated Spontaneous Speech (CASS) corpus was created to collect samples of most of the phonetic variations in Mandarin spontaneous speech due to pronunciation effects [1].

Made in ordinary classrooms, amphitheatres, or school studios without the benefit of high quality tape recorders or microphones, the recordings are of university lectures by professors and invited speakers, student colloquia, and other public meetings. The collection consists primarily of impromptu addresses, and were delivered in an informal style without prompts or written aids. As a result the recordings are of uneven quality and contain significant background noises. The recordings were delivered in audiocassettes and digitized into single-channel audio files at 16kHz rate and with 16-bit precision. A subset of over 3 hours’ speech was chosen for detailed annotation, which formed the CASS corpus. This corpus contains the utterances of 7 speakers at a speed as fast as about 4.57 syllables per second on an average, and in standard Chinese with slight dialectal backgrounds.

The CASS corpus was transcribed into a five-level annotation, including canonical Character Level, canonical Toned Pinyin (or Syllable) Level, Initial/Final Level with time boundary information, surface SAMPA-C [2] Level for observed IPA sequences, and Miscellaneous Level with labels for spontaneous phenomena that can be used for garbage/filler modeling.

3. Generalized Initials/Finals

In spontaneous speech, there are two kinds of differences between the canonical initials/finals (IFs) and their surface forms if the deletion and insertion are not considered. One is the sound change from one IF to a similar SAMPA-C sequence, such as nasalization, centralization, voiceless, voiced, rounding, syllabic, pharyngealization, and aspiration. The other is the phone change directly from one IF to another quite different SAMPA-C sequence. We refer to the surface form of an IF as its generalized IF (GIF). Obviously, the IFs are special GIFs.

3.1. Definition of GIF Set

The canonical IF set consists of 21 initials and 38 finals, totally 59 IFs. By searching in the CASS corpus, we initially obtain a GIF set containing over 140 possible IF-like SAMPA-C sequences. However, some of them occur for only a couple of times which can be regarded as least frequently observed sound variability forms therefore they are merged into the most similar canonical IFs. Finally we have 86 GIFs, taken as the SRUs.

In order to well model the spontaneous speech, additional garbage models are also built for breathing, coughing, crying,
disfluency, laughing, lengthening, modal, murmur, non-Chinese, smacking, noise, and silence.

3.2. Generalized Syllable (GS)

Similar to the GIF, we refer to any possible pronunciation of a given canonical syllable as one of its generalized syllables (GSs). According to the CASS corpus with the GIF transcription, it is easy to find all possible GSs. Each GS will be associated with an output probability \( P (\text{GS}_s | \text{IF}_k, \text{GIF}_k) \) defined as the probability of the GIF sequence (one generalized initial followed by one generalized final) given its corresponding canonical syllable, and it can be learned from the CASS corpus. This gives a probabilistic multi-entry syllable-to-GIF lexicon.

4. Pronunciation Modeling

Given an acoustic signal \( A \) of spontaneous speech, the goal of the recognizer is to find the canonical/baseform syllable string \( B \) that maximize the probability \( P(B|A) \). According to the Bayes' Rule, the recognition result is

\[
B^* = \arg \max_B P(B|A) = \arg \max_B P(A|B)P(B) \tag{1}
\]

In Equation (1), \( P(A|B) \) is the acoustic modeling part and \( P(B) \) is the language modeling part. We focus only on the acoustic modeling to propose some approaches.

4.1. Theory

Assume \( B \) is a string of \( N \) canonical syllables, i.e., \( B=b_1, b_2, ..., b_N \). For simplification, we apply the independence assumption to the acoustic probability,

\[
P(A|B) = \prod_{n=1}^{N} P(a_n | b_n) \tag{2}
\]

where \( a_n \) is the partial signal related to syllable \( b_n \). In general, by developing any term in right hand of Equation (2) we have

\[
P(a|b) = \sum_s P(a|b,s)P(s|b) \tag{3}
\]

where \( s \) is any surface form, i.e. GS, of syllable \( b \). Therefore, the acoustic modeling is divided into two parts; the first part \( P(a|b,s) \) is the refined acoustic model while the second part \( P(b) \) is the conditional probability of GS \( s \). Equation (3) provides a solution to the sound variability modeling by introducing a surface form term. We will present methods for these two parts.

4.2. IF-GIF Modeling

According to the characteristics of Chinese language, any syllable consists of an initial and a final. Because our speech recognizer is designed to take semi-syllables as SRUs, term \( P(a|b,s) \) should be rewritten in terms of semi-syllables. Assume \( b=(i, f) \) and \( s=(g_i, g_f) \), where \( i \) and \( gi \) are the canonical initial and the generalized initial respectively, while \( f \) and \( gf \) are the canonical final and the generalized final respectively. Accordingly, the independence assumption results in

\[
P(a|b,s) = P(a | i, g_i) \cdot P(a | f, g_f) \tag{4}
\]

More generally, the key point of acoustic modeling is how to model the IF and GIF related semi-syllable, i.e., how to estimate \( P(a|\text{IF}, \text{GIF}) \). There are three different choices:

- Use \( P(a|\text{IF}) \) to approximate - independent IF modeling.
- Use \( P(a|\text{GIF}) \) to approximate[3][4] - independent GIF modeling.
- Estimate \( P(a|\text{IF}, \text{GIF}) \) directly - IF-GIF modeling.

It is obvious that the IF-GIF modeling should be the best choice among these three kinds of modeling methods if there are sufficient training data, and it needs the IF-GIF transcription.

The IF Transcription is directly obtained from the Syllable Level transcription via a simple syllable-to-IF dictionary and is canonical. The GIF transcription is obtained once the GIF set is determined. By comparing the IF and GIF transcriptions, an actual observed IF transcription, named IF-A Transcription, is generated, where both deleted and inserted GIFs are considered. Finally the IF-GIF Transcription is generated directly from both the IF-A and GIF Transcriptions.

However, if the training data are not sufficient, the IF-GIF modeling will not work well or even work worse due to the data sparseness issue.

A reasonable method is to generate the IF-GIF models from their associated models; the adaptation techniques can be used[5]. There are at least two approaches. The IF-GIF models can be transformed either from the IF models or from the GIF models. The former method is called the base form GIF (B-GIF) modeling and the later the surface form GIF (S-GIF) modeling.

In the B-GIF modeling, we use the IF transcription to build all IF models. Suppose each IF corresponds to \( K \) GIFs, say \( GIF_k \), \( 1 \leq k \leq K \), then all the \( IF_k \) models are cloned from the base form model \( IF \). Finally by using the IF-GIF transcription and the adaptation technique, the \( IF-GIF_k \) models will be obtained. This can be regarded as “adapting \( P(a|b) \) to \( P(a|b,s) \)”.

In the S-GIF modeling, the GIF transcription is used to build all GIF models. Similarly, suppose \( K \) IFs, say \( IF_k \), \( 1 \leq k \leq K \) share \( GIF \), then all the \( IF-GIF_k \) models are cloned from model \( GIF \). The same adaptation method is used to obtain the \( IF_k-GIF \) models. This can be regarded as “adapting \( P(a|s) \) to \( P(a|b,s) \)”.

The difference between the S-GIF and the B-GIF methods lies only in how we initialize IF-GIF models; the former copies from the base form models while the latter the surface models. By comparing the two methods, it is straightforward to conclude that the initial IF-GIF models using B-GIF method will have bigger within-model scatters than those using the S-GIF method. This analysis shows S-GIF method will outperform B-GIF method.

The IF-GIF modeling enables multi-entry for each canonical syllable. Considering the multi-pronunciation probabilistic syllable lexicon, each entry in HTK[5] has the form of \( (b, i_{g_i}, f_{g_f}) \), where \( b=(i,f) \) is the base form and \( (g_i,g_f)=s \) its surface form.

4.3. Context-Dependent Weighting

In Equation (3), the second part \( P(s|b) \) stands for the output probability of a surface form given its corresponding base form.

A simple way to estimate \( P(s|b) \) is to directly learn from the database with base/surface form transcriptions. The resulting probability is referred to as the Direct Output Probability (DOP).

The problem is that the DOP estimation will not be so accurate if the training database is not big enough. Actually, what we are considering in the pronunciation weighting \( P(s|b) \)
are the base form and surface form of Chinese syllables, and the data sparseness in the syllable level remains a problem, therefore many weights are often not well estimated.

It is true that the syllable level data sparseness DOESN’T mean the semi-syllable level data sparseness, which suggests us to estimate the output probability via the semi-syllable statistics.

According to the Bayes’ Rule, the semi-syllable level output probability of a surface form GIF given its corresponding base form IF can be rewritten according to the context as

$$P(GIF \mid IF) = \sum_{C} P(GIF \mid IF, C) P(C \mid IF)$$

(5)

where C is the context of IF, it can be a bigram, a trigram or whatever related to IF. Suppose C includes the current IF and its left context $IF_L$. Equation (5) can be rewritten as

$$P(GIF \mid IF) = \sum_{IF_L} P(GIF \mid (IF_L, IF)) P(IF_L \mid IF)$$

(6)

In the sum on the right hand side of Equation (6), term $P(GIF(IF_L, IF))$ is the output probability given the context and term $P(IF_L \mid IF)$ is similar to the IF transition probability. These two terms can be learned from the database directly; hence Equation (6) is easy to be calculated offline. Based on the way of developing $P(GIF(IF))$, this method is called Context-Dependent Weighting (CDW) and the estimated probability is called the Context-Dependent Weight (CDW).

If we define

$$M_L(GIF \mid IF) = P(GIF \mid (L, IF)) P(L \mid IF)$$

(7)

Equation (6) can be rewritten as

$$P(GIF \mid IF) = \sum_{IF_L} M_{IF_L}(GIF \mid IF)$$

(8)

and according to Equation (6), we define another function:

$$Q(GIF \mid IF) = \max_{IF_L} M_{IF_L}(GIF \mid IF)$$

(9)

The above equations are focused on the initial and final, and the IF pair (IF$_L$, IF) could be either a (initial, final) pair or a (final, initial) pair.

To give the syllable level output probability estimation $P(s|b)$ as in Equation (3), we have three different estimations:

CDW-M: $P(s|b) = P(g_i | i) \cdot M_s(g_i | f)$

(10)

CDW-P: $P(s|b) = P(g_i | i) \cdot P(g_f | f)$

(11)

CDW-Q: $P(s|b) = Q(g_i | i) \cdot Q(g_f | f)$

(12)

Obviously Equation (10) considers the inner-syllable constraints, which is believed to be more useful. If Equation (10) or Equation (12) is used, the sum of approximated $P(s|b)$ over all possible $s$ for $b$ is often less than 1.0, that’s the reason we call it a weight instead of a probability.

If we do not consider the IF-GIF modeling, instead we assume that in Equation (3) $P(a_i|b,s) = P(a_i|b)$, in other words the acoustic modeling is exactly the GIF modeling. In this case the use of the CDW results in that the multi- pronunciation probabilistic syllable lexicon will have entries in the form of $(b, g_i, g_f, w_{s,t},)$, where the weight $w_{s,t}$ can be taken as one from Equations (10), (11), or (12), and nothing taken for the weight means the equal probability or equal weight.

4.4. Integrating IF-GIF modeling and CDW

When we consider both the CDW and the IF-GIF modeling, we have the multi- pronunciation syllable lexicon with entry in the form of $(b, i_{g_i}, f_{g_f}, w_{s,t},)$. 1.

5. Experimental Results

Experiments are done across the CASS corpus. The corpus is divided into two parts, the first part is the training set with about 3.0 hours’ spontaneous speech data and the second is the testing set with about 15 minutes’ data. The HTK is used for the training, adaptation and testing [5]. A 3-state 16-gaussian HMM is used to model each IF, GIF or IF-GIF. The feature is 39-dimension MFCC_E_D_A_Z with 25ms frame size every 15ms.

Experimental results include (1) UO: unit (IF, GIF or IF-GIF) level comparison without the syllable lexicon constraint; (2) UL: unit level comparison with the syllable lexicon constraint; and (3) SL: syllable level comparison with the syllable lexicon constraint. Listed are the correctness percentage (%Cor) and the accuracy (%Acc) as defined in HTK [5].

**Experiment 1. Independent IF modeling.** This is for comparison only. The result is listed in Table 1. The lexicon used here is a single-entry syllable-to-IF lexicon with equal weights.

<table>
<thead>
<tr>
<th>Item</th>
<th>IF</th>
<th>GIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%Cor</td>
<td>%Acc</td>
</tr>
<tr>
<td>UO</td>
<td>46.28</td>
<td>41.70</td>
</tr>
<tr>
<td>UL</td>
<td>50.34</td>
<td>42.30</td>
</tr>
<tr>
<td>SL</td>
<td>34.92</td>
<td>30.48</td>
</tr>
</tbody>
</table>

**Table 1. Results of independent IF and GIF modeling.**

**Experiment 2. Independent GIF modeling.** This is the baseline system where an equal probability or weight is provided for the multi-entry syllable-to-GIF lexicon. Also see Table 1 for results. We can find that in general the performance of independent GIF modeling is worse than independent IF modeling. It is obvious, because the GIF set is bigger than the IF set and hence GIFs will not be better trained than IFs on the same training database.

**Experiment 3. IF-GIF modeling.** This experiment is designed to test the IF-GIF modeling, $P(a_i|b,s)$. Except the acoustic models themselves, the experiment condition is similar to that in Experiment 2. The B-GIF and S-GIF modeling results are given in Table 2. The mean updating, MAP adaptation and MLLR adaptation methods are tried, and listed are the best results.

From the table, it is seen that S-GIF outperforms B-GIF. Compared with the GIF modeling, the S-GIF modeling achieves a syllable error rate (SER) reduction of 3.6%.

<table>
<thead>
<tr>
<th>Item</th>
<th>B-GIF</th>
<th>S-GIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%Cor</td>
<td>%Acc</td>
</tr>
<tr>
<td>UO</td>
<td>43.31</td>
<td>38.67</td>
</tr>
<tr>
<td>UL</td>
<td>46.67</td>
<td>38.25</td>
</tr>
<tr>
<td>SL</td>
<td>36.07</td>
<td>31.39</td>
</tr>
</tbody>
</table>

**Table 2. Results of B-GIF and S-GIF modeling.**

**Experiment 4. Pronunciation weighting.** This
experiment is designed to find a best way to estimate the pronunciation weight $P(s|b)$. To avoid the influence from the IF-GIF modeling, we use GIF modeling only, in other words we assume $P(a|b,s)=P(a|b)$, $P(a|b,s)=P(a|b)$ is not listed because it is much worse. In the syllable lexicon, two kinds of pronunciation weighting, i.e. DOP and CDW, are used for each entry. The results for DOP and CDW methods are listed in Table 3. Though for CDW $\sum P(s|b)\le 1$ and mostly it does not meet $\sum P(s|b)=1$, CDW outperforms DOP. Compared with IF-GIF modeling, the pure pronunciation weighting method CDW achieves a SER reduction of 5.1%.

Table 3. Results of DOP or CDW w/ GIF modeling.

<table>
<thead>
<tr>
<th>Item</th>
<th>DOP w/ GIF</th>
<th>CDW w/ GIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%Cor</td>
<td>%Acc</td>
</tr>
<tr>
<td>UL</td>
<td>47.78</td>
<td>40.53</td>
</tr>
<tr>
<td>SL</td>
<td>35.85</td>
<td>31.15</td>
</tr>
</tbody>
</table>

Experiments 5. Integrated pronunciation modeling. Either IF-GIF modeling or CDW pronunciation weighting improves the system performance individually; we have reason to believe that the integration of CDW and IF-GIF modeling will improve the performance much better. The result is given in Table 4. The SER reduction is 6.3% totally compared with the GIF modeling.

Table 4. Results of integrating CDW w/ IF-GIF modeling and when integrating syllable bigram.

<table>
<thead>
<tr>
<th>Item</th>
<th>CDW w/ IF-GIF</th>
<th>Plus Syllable N-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%Cor</td>
<td>%Acc</td>
</tr>
<tr>
<td>UL</td>
<td>47.28</td>
<td>40.72</td>
</tr>
<tr>
<td>SL</td>
<td>37.87</td>
<td>33.39</td>
</tr>
</tbody>
</table>

Experiments 6. Integration of syllable N-gram. Though language modeling is not the focus of pronunciation modeling, to make Equation (1) a complete one, we borrow a cross-domain syllable N-gram. This syllable bigram is trained using both read texts from Broadcast News (BN) and spontaneous texts from CASS, the amount of texts from BN is much bigger, therefore we call it a borrowed cross-domain syllable bigram. The result listed in Table 4 shows this cross-domain syllable N-gram is helpful and it reduces the SER by 10.7%.

Figure 1 gives an outline of all above experimental results. The overall SER reduction compared with GIF modeling and IF modeling is 6.3% and 4.2% (all without syllable N-gram).

6. Summaries

In order to model the pronunciation variability in spontaneous speech, firstly we propose the concept of generalized initial/final (GIF) and generalized syllable (GS) with or without probability, secondly we propose the GIF modeling and IF-GIF modeling aiming at refining the acoustic models, thirdly we propose the context-dependent weighting method to estimate the pronunciation weights, and finally we integrate the cross-domain syllable N-gram into the whole system.

Although the introduction of the IF-GIF modeling and the pronunciation weighting leads to performance reduction on the unit level compared with the IF modeling, but the syllable level overall performance for IF-GIF modeling greatly outperforms the IF modeling. From the experimental results, we conclude that

- The overall GIF modeling is better than the IF modeling.
- By refining the IF and GIF, the resulting IF-GIF modeling $P(a,b|s)$ is better than both the IF modeling $P(a|b)$ and the GIF modeling $P(a|s)$, even if data is sparse, when the S-GIF/B-GIF adaptation techniques can be used to provide a solution to data sparseness.
- The S-GIF method outperforms the B-GIF method because of the well-chosen adaptation initial models.
- The context-dependent weighting (CDW) is more helpful for sparse data than direct output probability (DOP) estimating.
- The cross-domain syllable N-Gram is useful.

7. References