Grapheme-to- Phoneme Conversion for Chinese Text-to-Speech

Jun Xu, Guohong Fu and Haizhou Li

Abstract

This paper reports a study of grapheme-to-phoneme (G2P) conversion for Chinese text-to-speech (TTS) system. As Chinese is a syllabic language, syllable is commonly adopted as the phonetic unit in TTS, which is represented by pinyin, the standard Chinese romanization. A Chinese G2P conversion is to find correct pinyin for polyphonic graphemes in the input text. In this paper, a complete G2P framework is presented, which includes a two-stage statistical word segmentation module, a hidden Markov model (HMM) based part-of-speech (POS) tagging module and a word-to-pinyin conversion module. In the word-to-pinyin conversion, a word grapheme is augmented by its POS tag in an effort to resolve the pronunciation disambiguation in G2P. The G2P experiments show that the polyphone G2P accuracy is improved by 9.41% after introducing POS module and further improved by 1.39% while applying the proposed word-to-pinyin method.

1. Introduction

Typically, there are several challenges in building a TTS system, such as text corpus design, speech corpus recording and labeling, prosody modeling, synthesizer development and importantly grapheme-to-phoneme (G2P) conversion.

In most of the alphabetic language such as English, the main objective of the grapheme-to-phoneme is to generate pronunciations for words that are out of vocabulary (OOV). Many approaches for letter-to-phoneme conversion have been proposed [1][2]. However, unlike the OOV problem in the alphabetic languages, the challenge in Chinese G2P conversion is to resolve the correct pinyin from polyphonic candidates. To help understand the issue in Chinese text, let’s exam a lexicon of 57,090 unique grapheme entries, as in Table 2. There are 453 polyphonic words out of 57,090 words in the Chinese lexicon that account for 0.8% of the lexicon in terms of word entries. However, they make up 11.3% in terms of grapheme counts in a real text corpus, as shown in Table 4 where we find 168,816 polyphonic graphemes out of a corpus of 1,491,023 graphemes. Therefore the issue of polyphone disambiguation has great impact on the quality of Chinese TTS system.

To implement the three modules in Chinese G2P, many approaches have been proposed [3,4,6,7,11,12]. Most of them...
rely on a Chinese lexicon. One of them is lexicon-based approach to word segmentation is FMM, forward maximum match [11], followed by dictionary lookup. A Chinese lexicon typically includes all the polyphones for an orthographic entry. During lookup, when multiple pronunciation choices, or pinyin, are available, the most frequently used pinyin is chosen. It is noted that the most used sound does not mean the correct sound.

Table 2: The distributes of words in Chinese lexicon

<table>
<thead>
<tr>
<th>Word Length (Grapheme number)</th>
<th>Number of Word</th>
<th>No. of Words With Poly-POS</th>
<th>Polyphonic Word Number</th>
<th>Resolvable by POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6763</td>
<td>1940</td>
<td>301</td>
<td>257</td>
</tr>
<tr>
<td>2</td>
<td>34357</td>
<td>4007</td>
<td>131</td>
<td>127</td>
</tr>
<tr>
<td>3</td>
<td>8457</td>
<td>98</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>7513</td>
<td>16</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>57090</td>
<td>6061</td>
<td>453</td>
<td>404</td>
</tr>
</tbody>
</table>

In this paper, we propose a statistical approach under an unified framework of hidden Markov model or HMM, that implements the three steps of G2P as described above. It makes use of contextual information of word syntactic statistics in addition to lexical information to effectively resolve polyphone disambiguation.

In Chinese, it is common that a polyphonic word is pronounced differently due to different POS. In Table 2, 404 polyphonic words out of total 453 could be disambiguated by their POS. We have good reason to believe that a word-to-pinyin method incorporating POS information is to outperform a simple dictionary lookup.

HMM is used as a basic solution to handle the G2P problems in this paper. That is how to find the state sequence \( \hat{X} = x_1, x_2, \ldots, x_m \) that best explains the observations \( O = o_1, o_2, \ldots, o_m \). In general, the HMM can be computed as:

\[
\hat{X} = \arg \max \prod_{i=1}^{m} P(x_i | x_{i-1}) P(o_i | x_i)
\]

where \( P(o_i | x_i) \) denotes the probability of \( o_i \) given state \( x_i \) in the current position \( i \), and \( P(x_i | x_{i-1}) \) denotes the probability of \( x_i \) given the history \( x_{1:i-1} \), i.e. bigram. It is noted that first order Markov dependency represented by bigram is assumed in the formulation. Given a set of training data, the parameters in equation (1) can be estimated by Baum-Welch algorithm under the maximum likelihood estimation (MLE) framework. The Viterbi algorithm is subsequently applied to resolve the best state sequence \( \hat{X} \).

This paper is organized as followings, in Section 2, we present a two-stage statistical word segmentation solution; Section 3 gives a HMM approach for POS tagging. Section 4 describes a word-to-pinyin method. Section 5 reports the G2P results in InfoTalk Speaker, a commercial multi-lingual TTS system [5]. Section 6 discusses the test results. Finally, we conclude with discussions.

2. A two-stage word segmentation

During the past decades, great success has been achieved in Chinese word segmentation [6][7]. However, two issues are still far from perfect, that is, the ambiguity resolution and unknown word identification. In Chinese, a word consists of one or more graphemes. Compound words of multiple graphemes are very common as shown in Table 2. Word entries kept in a lexicon are called known words. New words outside the lexicon, such as personal names, are called unknown words. One should notice that all single Chinese grapheme are all valid words. Therefore, identifying unknown word means chunking up consecutive small words to form a larger word. A two-stage statistical word segmentation solution is proposed in this paper.

2.1. Stage I: Segmentation of known words

Known word segmentation is a process of disambiguation of word boundary. In our system, we use word bigram language models to resolve word boundary ambiguities in word segmentation process.

For a Chinese grapheme (character) string \( C = c_1, c_2, \ldots, c_n \), there may be multiple candidate word sequences \( \{W = w_1, w_2, \ldots, w_n\} \) according to a given lexicon. Word bigram segmentation aims to find the segmentation \( \hat{W} = \arg \max P(W | C) \) that maximizes the probability, i.e.

\[
\hat{W} = \arg \max_{w} P(W | C) = \arg \max_{w} \prod_{i=1}^{n} P(w_i | W_{i-1})
\]

where \( P(w_i | W_{i-1}) \) denotes the word bigram, which is estimated from segmented corpus using MLE:

\[
P(w_i | W_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}
\]

To cope with data sparseness in MLE, linear interpolation technique [8] is used for word bigram smoothing.

2.2. Stage II: Unknown words identification

The unknown word identification consists of three major steps: (1) Firstly, an unknown word extractor extracts a sequence of known words \( \{W = w_1, w_2, \ldots, w_n\} \) that is likely to contain unknown words based on the related word-formation rules, word chunking probability and its left and right contextual word \( w_L, w_R \) . (2) A candidate word constructor then generates a lattice of all possible new segmentation \( \{W_U = \hat{x}_1, \hat{x}_2, \ldots, x_u\} \) that suggests unknown words from the extracted sequence. (3) A Viterbi decoder finally incorporates word chunking model, word-formation patterns and word bigram probability to score these candidates [9].

In this paper we will only discuss the results on the segmentation of known words. Much further research is still needed to study the polyphone disambiguation ability for unknown word.

3. POS tagging

After word segmentation, POS tagging is to resolve POS
ambiguity. Some words may have different syntactic categories in different context. Given a sequence of words, the task of POS tagging is to assign each word with a correct POS tag according to its context. Similar to word segmentation, a statistical method, HMM tagging, is applied to process POS tagging in this paper. Details can also be found in [10].

Given a word sequence \( \{W = w_1,w_2 \cdots w_n\} \), there may be more than one POS interpretation \( \{T = t_1,t_2 \cdots t_n\} \). The aim of HMM tagging is to find the most probable sequence of POS tags, i.e.

\[
\hat{T} = \text{argmax}_T P(T|W) = \text{argmax}_T \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})
\]  

(4)

where \( P(w_i|t_i) \) denotes the probability of word \( w_i \) given POS tag \( t_i \), the lexical probability, which can be estimated by computing the relative frequencies of the corresponding events in the training data, and \( P(t_i|t_{i-1}) \) denotes POS tag bigram, the contextual probability, which is estimated from the training corpus using MLE:

\[
P(t_i|t_{i-1}) = \frac{\text{Count}(t_{i-1},t_i)}{\text{Count}(t_{i-1})}
\]  

(5)

To cope with data sparseness in MLE, linear interpolation technique is used to smooth the estimated POS bigram.

4. Word-to-Pinyin Conversion

After POS tagging, each word in a sentence is associated with a POS tag. One implementation could be dictionary lookup if the POS-tagged word could single out a pronunciation. However, it is not always the case. Next we proposed a word-to-pinyin conversion approach based on contextual information. The word-to-pinyin conversion here is to incorporate POS tag into a HMM to disambiguate polyphonic words.

Given an input sequence of Words \( \{W = w_1,w_2 \cdots w_n\} \) and the relative sequence of POS tags \( \{T = t_1,t_2 \cdots t_n\} \), we start with constructing all potential pronunciations, or pinyin \( \{Y = y_1,y_2 \cdots y_n\} \) in a lattice. Each word \( w_i \) is augmented with its POS tag \( t_i \) to become an extended word \( (w_i,t_i) \). The aim of HMM is to find the most probable sequence of candidate pronunciations, i.e.

\[
\hat{Y} = \text{argmax}_Y P(Y|W,T) = \text{argmax}_Y \prod_{i=1}^{n} P(w_i|t_i)P(y_i|y_{i-1})
\]  

(6)

where \( P((w_i,t_i)|y_i) \) denotes the probability of word \( (w_i,t_i) \) given pinyin \( y_i \), which can be estimated by the relative frequencies of the corresponding events in the training data, and \( P(y_i|y_{i-1}) \) denotes pinyin bigram, which is estimated from the training corpus using MLE:

\[
P(y_i|y_{i-1}) = \frac{\text{Count}(y_{i-1},y_i)}{\text{Count}(y_{i-1})}
\]  

(7)

To cope with data sparseness in MLE, linear interpolation technique is used to smooth the pinyin bigram.

5. Experiments

This section reports the experiments using the G2P framework proposed in this paper, which is also implemented in InfoTalk Speaker. The test is intended to illustrate InfoTalk Speaker’s ability to perform accurate Chinese G2P in unrestricted text domains. Three aspects of test are reported, segmentation test, POS tagging test and word-to-pinyin conversion test. Furthermore, some comparison tests between the proposed methods and other methods are also included.

5.1. Test database

5.1.1. Lexicon

The InfoTalk Speaker engine involves three lexicons, namely, the system lexicon, the default user lexicon and user specified lexicon. Current test are conducted only on the first two lexicons, which contain 57,090 unique Chinese word entries. Each of the word entries is tagged with one or more POS categories. There are 38 categories. Table 3 lists the symbols that are used for the sample sentence in Table 1. A pinyin is associated with a Chinese word together with its POS.

<table>
<thead>
<tr>
<th>POS symbol</th>
<th>Part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>Pronoun</td>
</tr>
<tr>
<td>ad</td>
<td>Adjective used as adverbial modifier</td>
</tr>
<tr>
<td>v</td>
<td>Verb</td>
</tr>
<tr>
<td>u</td>
<td>Auxiliary</td>
</tr>
<tr>
<td>n</td>
<td>Noun</td>
</tr>
<tr>
<td>w</td>
<td>Punctuation mark</td>
</tr>
</tbody>
</table>

5.1.2. Corpus

In order to train models for G2P, we have built semi-automatically a word-POS-pinyin aligned data from Peking University corpus. This test is conducted against this corpus. As shown in Table 4, the training corpus contains 37,051 sentences or 920,089 words while the test corpus has 3,705 sentences or 91,862 words.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Training Corpus</th>
<th>Test Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>37,051</td>
<td>3,705</td>
</tr>
<tr>
<td>Lexicon</td>
<td>719,490</td>
<td>72,171</td>
</tr>
<tr>
<td>Unknown</td>
<td>41,999</td>
<td>3,783</td>
</tr>
<tr>
<td>Non-Standard</td>
<td>18,840</td>
<td>2,007</td>
</tr>
<tr>
<td>Punctuation</td>
<td>139,760</td>
<td>13,901</td>
</tr>
<tr>
<td>Graphemes</td>
<td>1,491,023</td>
<td>149,268</td>
</tr>
</tbody>
</table>

5.2. Results

We measure the performance by accuracy which is defined to be the number of correct hits in automatic analysis divided by
the total number of items in standard test-corpus, which is manually validated in terms of word segmentation, POS tagging and pinyin transcripts. An item in automatic analysis is considered to be correct if and only if it matches exactly the one in standard test-corpus. We also compare our results with other reported approaches in three experiments:

1. InfoTalk Speaker (bigram-WS/POS/bigram-W2P)
3. FMM-WS/DictionaryLookup-W2P

The test results are summarized in Table 5, where for each of the closed and open tests, the accuracy of word segmentation, POS tagging and G2P conversion are reported separately. It is observed in Table 4 that there are 16,998 polyphones out of 149,268 graphemes in test corpus. To demonstrate the effectiveness of the proposed approach over polyphone disambiguation, we report both the overall G2P accuracy and polyphone G2P accuracy.

Table 5: Performance testing results for G2P

<table>
<thead>
<tr>
<th>Items</th>
<th>Accuracy (%)</th>
<th>Exp.(1)</th>
<th>Exp.(2)</th>
<th>Exp.(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation</td>
<td>99.69</td>
<td>99.69</td>
<td>96.16</td>
<td></td>
</tr>
<tr>
<td>POS tagging</td>
<td>94.51</td>
<td>94.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2P</td>
<td>Overall</td>
<td>97.90</td>
<td>97.57</td>
<td>96.21</td>
</tr>
<tr>
<td></td>
<td>Poly-Phone</td>
<td>94.72</td>
<td>92.48</td>
<td>82.70</td>
</tr>
<tr>
<td>Segmentation</td>
<td>98.95</td>
<td>98.95</td>
<td>96.33</td>
<td></td>
</tr>
<tr>
<td>POS tagging</td>
<td>94.33</td>
<td>94.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2P</td>
<td>Overall</td>
<td>97.69</td>
<td>97.48</td>
<td>96.19</td>
</tr>
<tr>
<td></td>
<td>Poly-Phone</td>
<td>93.82</td>
<td>92.43</td>
<td>83.02</td>
</tr>
</tbody>
</table>

5.3. Results Analysis

It is observed in Table 5 that the proposed approach is the most effective. In open test, it achieves 98.95% for word-segmentation, 94.33% for POS tagging, 97.69% for G2P and 93.82% for polyphone disambiguation respectively. Again in open test, comparing experiment (2) and (3), we find that POS tags contribute to 9.41% absolute G2P accuracy improvement for polyphones. When comparing experiment (1) and (2), we find that the proposed bigram word-to-pinyin model contributes another 1.39% to G2P accuracy improvement.

6. Discussion

Further investigation also suggests several sources of error in G2P. Firstly, it is found that inconsistency between lexicon and test database, from different sources, leads to some errors. For example, the Chinese word, 爱乐乐团, is manually transcribed into pinyin, WO3-MEN2, in the standard test corpus. However, it is found WO3-MEN5 in TTS lexicon. Secondly, Chinese word definition is always a debating subject, it is not easy to have a complete lexicon. However, from G2P point of view, the wider coverage, the better the lexicon is. For example, there is an unknown word in the Chinese grapheme sentence, 爱乐乐团, 爱乐 should be pronounced A14-YUE4. Since there is no 爱乐 in the system lexicon, the word will be segmented as 爱爱, and pronounced as AI4 LE4, that the bigram-WS and POS could not help. Lastly, there are many undesired errors in the standard test corpus, which has great impact on the test results. We will refine the database in the future.

7. Conclusions

This paper presents a statistical framework for Chinese G2P. It’s shown that the HMM statistical method is effective in word segmentation, POS tagging and word-to-pinyin conversion. The framework allows us to incorporate word syntactic statistics using bigram in addition to lexical information in the process. Without loss of generality, the framework can be extended to benefit from higher order statistics, such as trigram. The proposed three steps G2P approach is especially effective for polyphone disambiguation. The experiments also suggest using POS tagging to improve Chinese G2P performance. The proposed approach has been integrated into InfoTalk Speaker, a popular TTS product for Chinese Mandarin, Cantonese in the speech industry.

8. References