ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

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

Jun Xu, Guohong Fu and Haizhou Li

InfoTalk Technology
Republic of Singapore
jun.xu@infotalkcorp.com

Dept of Linguistics
The Univ. of Hong Kong
ghfu@hkucc.hku.hk

Institute for Infocomm
Research, Singapore
hli@i2r.a-star.edu.sg

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.

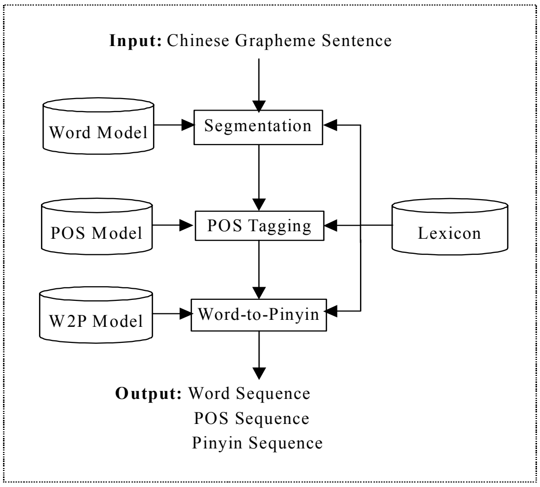

Figure 1: Workflow of G2P for Chinese TTS

In Chinese TTS, there are some additional complications because Chinese is an ideographical language. Unlike alphabetic languages, the correspondence between grapheme and sound in Chinese is not obvious. Furthermore, there are many polyphonic graphemes as well. A G2P for Chinese TTS system consists of three modules as shown in Figure 1, a sample is given in Table 1 to illustrate the steps. The sentence consists of six Chinese graphemes (characters) and one punctuation mark, which is further segmented into 5 words and one punctuation mark. In the rest of this paper, we refer Chinese grapheme to Chinese character.

Word Segmentation (WS): There are no spaces or other delimiters between words in Chinese text. A word segmentation algorithm is therefore required. The task is to identify the word boundary in the running text.

POS Tagging: For most of the polyphonic words, their sound can be disambiguated by their POS. As a word usually has multiple POS categories, the challenge here is to find the most likely POS tag for each of the words according to the context.

Word-to-Pinyin (W2P) Conversion: There may be more than one pinyin for a same word. The purpose of *word-to-pinyin* conversion is to select a correct pinyin for the input word from system lexicon in terms of its word-form and POS.

Table 1. An illustration of three steps Chinese G2P

An input Chinese sentence: 他连续敬了礼。						
WS	他	连续	敬	了	礼	。
POS	r	ad	v	u	n	w
W2P	TA1	LIAN2-XV4	JING4	LE5	LI3	

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

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

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. As shown 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 $\hat{X} = x_1 x_2 \cdots x_m$ that best explains the observations $O = o_1 o_2 \cdots o_m$. In general, the HMM can be computed as:

$$\hat{X} = \underset{X}{\operatorname{argmax}} P(X|O) \cong \underset{X}{\operatorname{argmax}} \prod_{i=1}^{m} P(o_i|x_i) P(x_i|x_{i-1}) \qquad (1)$$

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_{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 $C = c_1 c_2 \cdots c_n$, there may be multiple candidate word sequences $\{W = w_1 w_2 \cdots w_m\}$ according to a given lexicon. Word bigram segmentation aims to find the segmentation $\{W = w_1 w_2 \cdots w_m\}$ that maximizes the probability, i.e.

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|C) \cong \underset{W}{\operatorname{argmax}} \prod_{i=1}^{m} P(w_i|w_{i-1}) \qquad (2)$$

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

$$P(w_i | w_{i-1}) = \frac{Count\ (w_{i-1}, w_i)}{Count\ (w_{i-1})} \qquad (3)$$

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 $\{W = w_1 w_2 \cdots w_m\}$ 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 $\{W_U \mid W_U = x_1 x_2 \cdots x_m\}$ 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 thus to assign each word with a correct POS tag according 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 $\{W = w_1 w_2 \cdots w_m\}$, there may be more than one POS interpretation $\{T = t_1 t_2 \cdots t_m\}$. The aim of HMM tagging is to find the most probable sequence of POS tags, i.e.

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T \mid W) = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{m} P(w_i \mid t_i) P(t_i \mid t_{i-1}) \qquad (4)$$

where $P(w_i \mid t_i)$ 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 $P(t_i \mid t_{i-1})$ denotes POS tag bigram, the contextual probability, which is estimated from the training corpus using MLE:

$$P(t_i \mid t_{i-1}) = \frac{Count\ (t_{i-1}, t_i)}{Count\ (t_{i-1})} \qquad (5)$$

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 $\{W = w_1 w_2 \cdots w_m\}$ and the relative sequence of POS tags $\{T = t_1 t_2 \cdots t_m\}$, we start with constructing all potential pronunciations, or *pinyin* $\{Y = y_1 y_2 \cdots y_m\}$ 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.

$$\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y \mid (W,T)) = \underset{Y}{\operatorname{argmax}} \prod_{i=1}^{m} P((w_i, t_i) \mid y_i) P(y_i \mid y_{i-1}) \qquad (6)$$

where $P((w_i, t_i) \mid y_i)$ 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 $P(y_i \mid y_{i-1})$ denotes *pinyin* bigram, which is estimated from the training corpus using MLE:

$$P(y_i \mid y_{i-1}) = \frac{Count\ (y_{i-1}, y_i)}{Count\ (y_{i-1})} \qquad (7)$$

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 3. Sample part-of-speech categories

POS symbol	**Part-of-speech**
r	Pronoun
ad	Adjective used as adverbial modifier
v	Verb
u	Auxiliary
n	Noun
w	Punctuation mark

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 4: Experimental corpora for G2P

Corpus		Training Corpus	Test Corpus
Total sentences		37,051	3,705
Words	Total	920,089	91,862
	Lexicon	719,490	72,171
	Unknown	41,999	3,783
	Non-Standard	18,840	2,007
	Punctuation	139,760	13,901
Graphemes	Total	1,491,023	149,268
	Chinese	1,303,924	130,265
	Polyphone	168,816	16,998
	Non-Chinese	187,099	19,003

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)
(2) Bigram-WS/POS/DictionaryLookup-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

Items			Accuracy (%)		
			Exp.(1)	Exp.(2)	Exp.(3)
Closed Test	Segmentation		99.69	99.69	96.16
	POS tagging		94.51	94.51	-
	G2P	Overall	97.90	97.57	96.21
		Poly-Phone	**94.72**	**92.48**	**82.70**
Open Test	Segmentation		98.95	98.95	96.33
	POS tagging		94.33	94.33	-
	G2P	Overall	97.69	97.48	96.19
		Poly-Phone	**93.82**	**92.43**	**83.02**

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, 我们, 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, 维也纳 爱乐 乐团. 爱乐 should be pronounced AI4-YUE4. Since there is no 爱乐 in the system lexicon, the word will be segmented as 爱 and 乐, 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

[1] O. Andersen, R. Kuhn, et al, "Comparison of two tree-structured approaches for grapheme-to-phoneme conversion", ICSLP 1996, Philadephia.

[2] K. Torkaolla, "An efficient way to learn English grapheme-to-phoneme rules automatically", ICASSP 1993, Minneapolis.

[3] Daju Gou and Wanbo Luo, "Processing of polyphone character in Chinese TTS system", Chinese Information, No.1, pp.33-36.

[4] Zirong Zhang, Min Chu and Eric Chang, "An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese", ISCSLP 2002

[5] Jun Xu and Haizhou Li, "InfoTalk Speaker 2.0, The state of the art TTS system", Technical Report HK/SG-2002, InfoTalk Research and Development Center, Singapore.

[6] Guohong Fu, and Xiaolong Wang, "Unsupervised Chinese word segmentation and unknown word identification", NLPRS 1999, Beijing.

[7] Jianyuan Nie, M. L. Hannan and W.Y. Jin, "Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge", Communication of COLIPS, 1995.

[8] Jelinek Frederick and Robert L. Mercer, "Interpolated estimation of Markov source parameters from sparse data. In proceeding of workshop on Pattern Recognition in Practice 1980, Amsterdam.

[9] Guohong Fu and Kangkwong Luke, "Chinese unknown word identification using class-based LM", IJCNLP 2004, Sanya.

[10] Guohong Fu, "Statistic approaches to Chinese syntactic ambiguity resolution", Ph.D. thesis. Harbin Institute of Technology 2001, Harbin.

[11] Nanyuan Liang, "书面汉语自动分词系统－CDWS", in Journal of Chinese Information Processing, 1987.

[12] Haizhou Li, et al, "Chinese sentence tokenization using Viterbi decoder", International Symposium on Chinese Spoken Language Processing, Dec 1998, Singapore