TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/18 pp606-611
Volume 14, Number 5, October 2009

Improved Grapheme-to-Phoneme Conversion for Mandarin TTS

YI Lifu (易立夫)**, LI Jian(李 健), HAO Jie (郝 杰), XIONG Ziyu (熊子瑜)†

Toshiba (China) Research and Development Center, Beijing 100086, China;
† Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China

Abstract: Several methods were developed to improve grapheme-to-phoneme (G2P) conversion models for Chinese text-to-speech (TTS) systems. The critical problem of data sparsity was handled by combining approaches. First, a text-selection method was designed to cover as many G2P text corpus contexts as possible. Then, various data-driven modeling methods were used with comparisons to select the best method for each polyphonic word. Finally, independent models were used for some neutral tone words in addition to the normal G2P models to achieve more compact and flexible G2P models. Tests show that these methods reduce the relative errors by 50% for both normal polyphonic words and Chinese neutral tones.

Key words: grapheme-to-phoneme conversion; text design; Chinese neutral tone; speech synthesis

Introduction

Grapheme-to-phoneme (G2P) conversion is important for Chinese text-to-speed (TTS) systems. This paper focuses on the G2P conversion module after text analysis and normalization in a TTS system. The G2P conversion converts a target word from its written form (graphemes) to its pronunciation (phonemes) by giving the word segmentations after the text normalizations.

The main target words of English G2P conversion are the out-of-lexicon words. The target words of Chinese G2P conversion are the polyphonic words, whose pronunciations are determined by their contexts. Their proper pronunciation is beyond the scope of the TTS lexicon since the lexicon only stores the main pronunciation of each word. Chinese lexicons cannot use model training since the lexicon item cannot afford contextual alignment information besides part-of-speech (POS) and pronunciation. Thus, a text corpus with alignment information is necessary for Chinese G2P model training. Such a corpus is scarce and it must be carefully transcribed manually.

The Chinese language contains thousands of characters (the graphemes) and thousands of pinyins (the phonemes) combinations, many times more than those in English. This increases the possible contexts for each polyphonic word, which makes the task very complex not only for models, but also for the corpus. A large G2P corpus will require more complex G2P models. However, embedded TTS systems have limited memory and computational ability. Also, a large corpus requires excessive expense and time for the manual transcription. Therefore, the text design before transcription for the G2P corpus becomes very critical to balance the workload and model complexity. There are no known studies focusing on the balance problem for the Chinese G2P. Most recent studies on non-Chinese languages have concentrated on machine-learning techniques, such as decision tree (DT)[1,2], joint maximum entropy *n*-gram model[3], hidden Markov models (HMM)[4], and example-based methods[5], since they have enough existing lexicons or corpora for training. Some models have been proposed for Chinese G2P conversion, such as extended stochastic complexity (ESC)-based stochastic list[6], HMM[7], and transformation-based learning (TBL)[8]. These

Received: 2009-03-26; revised: 2009-04-30
** To whom correspondence should be addressed.
E-mail: yilifu@rdc.toshiba.com.cn; Tel: 86-10-85183111

techniques are powerful and perform well, however, studies on Chinese G2P conversion lack appropriate comparisons of the various modeling methods using similar feature sets.

The neutral tone is a special phonetic phenomenon in Chinese, which sounds quite weak with pitch contour changes that do not fit a normal lexical tone. These tone changes also sometimes lead to different meanings in the text. Thus the prediction of neutral tones is very important for TTS naturalness. Past studies[6-8] have treated the neutral tone as a common G2P question with little attention paid to the characteristics of the neutral tone itself. However, there are a set of some common rules for neutral tones, so word dependent models waste storage and are not convenient for model extension.

This paper presents a text search method for G2P, which increases the coverage for G2P models with an acceptable manual workload. A unified model is given to predict Chinese neutral tone words, which reduces the model complexity and increases the model reliability. For the Chinese polyphonic words, several data-driven methods, including generalized linear model (GLM) are used to simultaneously predict the pronunciation with an optimal method used to select the final sound.

1 Modeling Methods

Chinese G2P seeks to predict the pinyin sequence $\{Y=y_1y_2...y_n\}$ for a given word sequence $\{W=w_1w_2...w_n\}$ containing polyphonic words and the part-of-speech (POS) tag sequence $\{T=t_1t_2...t_n\}$. For the polyphonic words, the correct pronunciation is chosen from its two or more candidate pronunciations. The Chinese G2P conversion is then a common classification problem.

The most popular methods for G2P conversion are decision trees, such as CART and C4.5 methods[1]. Hence, four methods, C4.5, CART, GLM, and TBL were used for the G2P conversion. Since C4.5 and CART are widely used in other TTS researches, only GLM and TBL are introduced in this section.

1.1 Generalized linear model

GLM was previously developed for modeling[9]. It is a generalization of the multivariate linear regression model[10]. The GLM model predicts the pronunciation probability from the word features by

$$\mathrm{Pr}_i = \hat{\mathrm{Pr}}_i + e_i = h^{-1}(\beta_0 + \sum_{j=1}^{p}\beta_j C_{ij}) + e_i, \quad 0 < i \leqslant N \tag{1}$$

where Pr_i is the probability of a candidate pronunciation for the polyphonic word, e_i is the prediction error, h is a link function, β_j is a regression coefficient, C_{ij} is the context linguistic feature, and p is the dimension of the regression coefficient vector. N is the number of training samples, and i is the sample index. C_{ij} may not only be linear features, but also feature combinations or feature interactions. GLM treats a feature interaction as a new linear feature with only one regression coefficient. When h equals the identity function, the GLM is a plain GLM and the probability obeys the Gaussian distribution. When h equals the logit function, the GLM model is a logistic GLM model and the pronunciation probability obeys the Bernoulli distribution[8]. The logit function is defined as:

$$h^{-1}(z) = \mathrm{e}^z / (1+\mathrm{e}^z) \tag{2}$$

$$h(\hat{\mathrm{Pr}}_i) = \mathrm{logit}(\hat{\mathrm{Pr}}_i) = \log[\hat{\mathrm{Pr}}_i / (1-\hat{\mathrm{Pr}}_i)] = \beta_0 + \sum_{j=1}^{p}\beta_j C_{ij} \tag{3}$$

The logistic GLM guarantees that $\Pr(P|C)$ will range from 0 to 1 when $\Pr(P|C)$ is a nonlinear function of the context, since the logistic GLM can only process binary classification questions, it is used easily for polyphonic words with two pronunciations.

1.2 TBL

TBL is another successful error-driven machine learning technique[8]. TBL starts with a simple solution to the problem and applies transformations step by step. At each step, the transformation that gives the largest benefit is selected for the solution. The TBL steps when the selected transformation does not modify the data in enough places or there are no more transformations to be selected.

The open source toolkit fnTBL[11] was used for the G2P modeling. The fnTBL user needs to specify the rule templates, which describe the rules that the system will try to learn. The template descriptions define which features are checked for a proposed transformation.

The four methods, CART, C4.5, TBL, and GLM, are all simple for training and storage. So they were all compared for the G2P modeling.

2 G2P Text Design

The pronunciation (Pinyin) of a Chinese polyphonic word is almost determined by its context. The G2P text design selects appropriate contextual texts for important polyphonic words. The context design selects the most important Chinese polyphonic words and then analyzes it to select corresponding texts from a large corpus.

2.1 Polyphonic word analysis

The polyphonic words selected here are those with pronunciations that cannot be determined by the lexicon. These polyphonic words are unevenly distributed in round texts with some words occurring frequently whereas others rarely occur. Some polyphonic words have dominant pronunciations while others do not.

The TTS speech corpus contains about 10 000 utterances that are carefully transcribed pinyin descriptions of their pronunciations. The corpus contains a total of 713 polyphonic characters.

The uncertainty of the pronunciations is described by Shannon entropy, which is a measure of randomness of pronunciations. The importance weight of each polyphonic word is defined as the entropy times the occurrence frequency. 138 polyphonic words chosen with the highest weights gave a coverage ratio of the occurrences of all polyphonic words of 91.99%. These 138 polyphonic words were then the target words for the G2P study. The occurrences and weights of the polyphonic words are quite different as Table 1 shows.

Table 1 Weights and occurrence information for selected polyphonic words

Word	Main pronunciation	Entropy	Occurrence frequency	Weight
为	/wei2/	0.9894	611	604.5410
中	/zhong1/	0.0668	504	33.6569
将	/jiang1/	0.0399	467	18.6138
还	/hai2/	0.3185	316	100.6560

2.2 Text design

Well-designed context samples are essential for training balanced models for the key polyphonic words. Most Chinese G2P studies[6] have used common-used TTS corpora with sizes of 1 M - 2 M characters. However, these corpora may not be big enough to be a very rich context database. Therefore, the studies based on common corpora cannot be used for text design, so the studies did not analyze the text coverage. This study attempts to preserve a contextual balance for the polyphonic words. The most important features of polyphonic words are the two neighboring words and their POSs, as pointed out in past studies. Hence the two neighboring words and the polyphonic word itself are used as the contextual features for analysis and design. The design is based on a very large text corpus, the "People's Daily" newspapers with 25 years of articles, which contains about 544 M Chinese characters. The corpus covers various topics including politics, economics, technology, entertainment, art, and life.

A text analyzer and a text normalizer were used to process the corpus to obtain the word segmentation sequences and the identity context of the neighboring words and the polyphonic word. A total of 877 627 context samples of the 138 selected polyphonic words were extracted from the corpus with an average of about 1613 context samples for every 1 M characters. This shows that the common corpora of 1 M -2 M characters cannot possibly contain sufficient training data for polyphonic words.

The types of context samples were then analyzed. The very typical Chinese polyphonic word "为" has 51 693 usages with 31 377 distinct context samples with 26 502 of those occurring only once. There are certain high frequency samples that occur in the news, such as "为此" (since), "全心全意为人民"(Wholeheartedly for the people), "题为" (be entitled as), "为实现" (to achieve), and "为国家" (for the country). These high frequency samples are valuable and should be extracted when selecting the context samples. The percentage of polyphonic words included in a given number of context samples is shown in Table 2.

Table 2 Percentage of polyphonic words included in a given number of selected context samples

Number of selected contexts	Total coverage (%)
55 050	6.27
100 115	11.41
152 510	17.38
205 235	23.39
254 665	29.02
300 708	34.26
355 373	40.49
403 664	45.99
495 712	56.48
877 627	100.00

Each distinct context sample was included only once in the corpus to give a complete coverage with a limited number of sample texts. The texts were then sorted by their occurrence frequency. The text selection process is illustrated in Fig. 1.

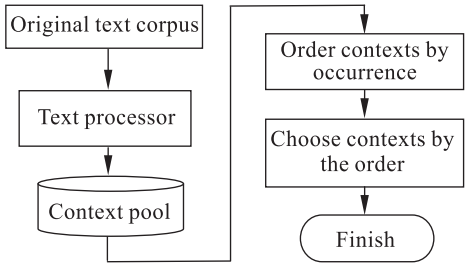

Fig. 1 Text selection process for the Chinese G2P

The various polyphonic words do not share any information in the model training for the Chinese G2P conversion. Therefore, separate G2P models were built for each polyphonic word. G2P modeling training has been found to require 3000-4000 contexts for each polyphonic word[6-8]. However, 3000 contexts for each of the 138 words will give about 400 K contexts to transcribe, which would be too expensive and time-consuming. However, many words rarely appear in common texts, so even 500 distinct contexts for such words would possible be too many. For example, there are only 186 distinct contexts for the word "折".

Therefore, there cannot be a fixed preset limitation on the number of context samples for all the polyphonic words, but the number of context samples will depend on these occurrence frequencies. In the end, about 150 thousand contexts were chosen for the text and analysis. The statistical contextual coverage was 17.08%. The number of selected context samples was from 200 to 3500. The manual transcriptions gave the word segmentations and pronunciations of the selected texts.

3 G2P Modeling Experiments

The transcribed polyphonic data gives the alignment information between the word sequence *W*, the POS sequence *T*, and the pinyin sequence *Y*. This information was then used in several data-driven methods to predict the appropriate pinyin for the Chinese polyphonic words as closely as possible and to choose the best method.

3.1 Feature settings

The features for the G2P conversion were the word sequence, *W*, and the POS tag sequence, *T*. These were similar feature settings for the four methods, but there were still some differences in the model characteristics and training requirements.

Since there are thousands of Chinese words, use of word feature in the GLM training will significantly increase the number of degrees of freedom in the GLM, so even 3500 samples preserved is insufficient for training. Therefore, a simplified feature fixed setting, (t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}) was used for each polyphonic word w_i, for the GLM training, to avoid overfitting.

The C4.5 and CART used the same feature set based on the two previous words and the following two words. Therefore, the feature set for each polyphonic word, w_i, was (w_{i-2}, t_{i-2}, w_{i-1}, t_{i-1}, w_{i+1}, t_{i+1}, w_{i+2}, t_{i+2}). The feature templates defined for TBL were

(t_{i-2}=>y_i), (w_{i-2}=>y_i), (t_{i-1}=>y_i), (w_{i-1}=>y_i), (t_{i+1}=>y_i), (w_{i+1}=>y_i), (t_{i+2}=>y_i), (w_{i+2}=>y_i), (t_{i-1}, t_{i+1}=>y_i), (w_{i-1}, t_{i-1}=>y_i), (w_{i-1}, t_{i+1}=>y_i), (t_{i-2}, t_{i-1}=>y_i), (t_{i-1}, w_{i+1}=>y_i), (t_{i+1}, w_{i+1}=>y_i), (t_{i-2}, t_{i-1}, t_{i+1}=>y_i), (w_{i+1}, w_{i-1}=>y_i).

3.2 Test results

The training data of each polyphonic word was split into two parts with 75% for training and 25% for testing. The results focus on the most important 46 polyphonic words (with a training data set size<=500 samples and main pronunciation probability<0.9) out of the 138 polyphonic words. Compared with using just each word's primary pronunciation, the average relative error rate reduction (RERR) for the 46 polyphonic words was 53.06%. The average error reduction of the 115 polyphonic words (including the 46 words, with error reduction>0) was 53.32%.

However, the G2P models did not improve the pronunciation accuracies of the 23 polyphonic words l. For these words, the average probability then using their primary pronunciations was 91.54%, and the average size of the training data was 763. That means there are only about 65 (=763×(1– 0.9154)) context samples for the secondary pronunciations for the model training, which is too few for the data-driven methods so the results are reasonable.

The predicting performance data for the important polyphonic words is listed in Table 3. The maximum accuracy in Table 3 is the maximum achieved of all four methods.

Table 3 Reduction performance of some polyphonic words

Word	Main pronunciation probability	Maximum accuracy	Best method
长	0.6352	0.8640	C4.5
攒	0.8865	0.9150	C4.5
朝阳	0.5348	0.7931	GLM
处	0.8760	0.9677	TBL
血	0.8806	0.9226	TBL
炸	0.8508	0.8851	TBL
只	0.7138	0.9793	CART

Comparing the 4 methods for the 46 most important polyphonic words showed that C4.5 was the best method for 17 polyphonic words, TBL was the best for 21 polyphonic words, and both CART and GLM were best for only 4 polyphonic words. Thus, TBL and C4.5 obviously outperform the other 2 methods.

TBL has been found to always perform better than the DT-based methods[8], but the present results are quite different. The polyphonic words have various regular linguistic combinations with different training data sets for each polyphonic word. The fitting method is normally the best, especially for small G2P training data sets.

4 Neutral Tone Methods

Lu[12] identified some regular patterns for neutral tones in Chinese speech. Hence, the neutral tone words were analyzed in the corpus with 10 000 utterances described in Section 2. Four types of neutral tone words were found with different strategies developed for each type.

(1) Some auxiliary and mood words, such as 的, 了, 吗, 吧, 哟, 呀, 哇, and 啊, are almost always neutral tone words. Therefore rule-based methods can be used for these words.

(2) Repeated words in Chinese normally contain neutral tone syllables. These can be classified into different types such as "AABB" (e.g., 仔仔细细) and "ABAB" (e.g., 痛快痛快). Rule-based methods can also be used for these words.

(3) Some special function words also frequently have neutral tones, such as 上, 下, 来, 着, 子, 过, 下, 地, 到, and 个. These words do not use their primary pronunciations, so prediction models are needed for them.

(4) Finally, there are other irregular neutral tone words that have very complex pronunciation patterns that require further analysis.

11 neutral tone Type 3 words, 到, 地, 个, 过, 上, 是, 头, 下, 在, 着, and 子, were selected which frequently occur in the corpus, for analysis. Among the total of 4258 samples extracted from the corpus, 3476 (81.63%) samples were not neutral tones, with only 782 (18.37%) samples being neutral tones. The modeling tests focused on these words that are neutralized when attached as dependent morphemes in polysyllabic word units.

Neutral tone prediction is a binary classification problem. The preliminary tests used only CART and GLM models for the 11 neutral tone words. The GLM feature set was $(t_{i-1}, w_i, t_i, t_{i+1})$ for word w_i. The CART tests used two feature sets with set A the same as for the GLM and set B including the 2 neighboring words in set A, i.e., $(w_{i-1}, t_{i-1}, w_i, t_i, t_{i+1}, w_{i+1})$. Use of the GLM method for a set is too complex to compute. Similar tests as in Section 3 gave the predicting accuracies listed in Table 4.

Table 4 Performance comparison for the neutral tone words

Method	Feature set	Accuracy (open test)
CART	Set A	0.9117
CART	Set B	0.9371
GLM	Set A	0.9540

The result in Table 4 shows that the GLM performance with set A is better than the CART result. The CART performance was better using set B than that with set A; thus addition of the neighboring words increases the goodness of fit. The accuracy using the GLM method in neutral tone prediction increases from 81.63% to 95.40%, which gives a relative error reduction of 74.96%.

5 Conclusions

Several approaches were developed to improve G2P conversion for Chinese text-to-speech systems. A simple efficient approach was developed for the G2P text selection. Furthermore, a generalized linear model was

then developed for the G2P conversion with comparisons to the TBL, C4.5, and CART methods. The best methods for each polyphonic word were selected based on their performance. Prediction models were also developed for Chinese neutral tones. These methods were developed to improve the performance of the Toshiba Mandarin TTS that won the Blizzard Challenge 2008.

This paper only presents a comparison of the 4 methods without an analysis of reasons for the performance differences. The G2P context text selection restricted the context definitions just to the neighboring words, while ignoring the influences of other words that may be important. In addition, data sparsity is still a problem for the polyphonic words even with the large corpus used here. Further work will improve the predictions for neutral tones with a larger corpus, with more tests to improve the analyses.

References

[1] Andersen O, Kuhn R, Lazarides A, et al. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proceedings of the 4th International Conference on Spoken Language Processing. Philadelphia, PA, USA, 1996: 1808-1811.

[2] Keinappel K, Kneser R. Designing very compact decision trees for grapheme-to-phoneme transcription. In: Proceedings of the 7th European Conference on Speech Communication and Technology. Aalborg, Denmark, 2001: 1911-1914.

[3] Chen S F. Conditional and joint models for grapheme-to-phoneme conversion. In: Proceedings of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland, 2003: 2033-2036.

[4] Taylor P. Hidden Markov models for grapheme-to-phoneme conversion. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005: 1973-1976.

[5] Charoenpornsawat P, Schultz T. Example-based grapheme-to-phoneme conversion for Thai. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, PA, USA, 2006: 1268-1271.

[6] Zhang Zirong, Chu Min, Chang Eric. An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese. In: Proceedings of 2002 International Symposium on Chinese Spoken Language Processing. Taipei, China, 2002: 59-62.

[7] Xu Jun, Fu Guohong, Li Haizhou. Grapheme-to-phoneme conversion for Chinese text-to-speech. In: Proceedings of the 8th International Conference on Spoken Language Processing. Jeju Island, Korea, 2004: 1885-1888.

[8] Zheng Min, Shi Qin, Zhang Wei, et al. Grapheme-to-phoneme conversion based on TBL algorithm in Mandarin TTS systems. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005: 1897-1898.

[9] Yi Lifu, Li Jian, Lou Xiaoyan, et al. Phrase break prediction using logistic generalized linear model. In: Proceedings of the 9th International Conference on Spoken Language Processing. Pittsburgh, PA, USA, 2006: 1308-1311.

[10] McCullagh P, Nelder J A. Generalized Linear Models. London, England: Chapman & Hall, 1989.

[11] Florian R, Ngai G. FnTBL toolkit. http://www.cs.jhu.edu/~rflorian/fntbl/index.html, 2001.

[12] Lu Yunzhong. Neutral Tone and R-Ending Retroflexion. Beijing, China: The Commercial Press, 2001. (in Chinese)