別紙４（課程博士）

Date of Submission:
平成 28 年 1月 6日

Department 情報・知能工学専攻	Student ID Number 学籍番号	第 129302 号
Applicant's name 氏名	KHEANG Seng	

Supervisors 指導教員	Kouichi KATSURADA Shigeru KURIYAMA

Abstract

論文内容の要旨 (博士)

Title of Thesis 博士学位論文名	A Study on Two-Stage-based Architecture for Grapheme-to-Phoneme Conversion （書記素－音素変換のための 2 ステージアーキテクチャに関する研究）

(Approx. 800 words)
(要旨 1,200 字程度)

A modern speech synthesis (Text-to-Speech or TTS) system usually generates output speech through phonological information (or phonetic transcription) rather than direct representation of textual information. The phonemic transcription of a written word could be possibly generated by consulting a pronunciation dictionary available inside the system for the in-vocabulary words or predicted through a data-driven Grapheme-to-Phoneme (G2P) conversion for the unknown or out-of-vocabulary (OOV) words. Due to the variability in the pronunciation rules, there is no strict correspondence between graphemes and phonemes, especially in English language. In order to improve the prediction performance of the G2P conversion model, we propose several approaches based on a two-stage architecture which allows to treat the problems occurred in the conversion using two different steps: graphemes-to-phoneme and phonemes-to-phonemes.

Our first approach is called "a two-stage neural network-based G2P conversion" which is designed for dealing with the problem of conflicting phonemes, where an input grapheme could, in the same context, produce many possible output phonemes at the same time. For example, if a neural network model takes a sequence of seven graphemes as input, the grapheme 'A' on sequence "HEMATIC" can produce the phoneme /AE/ when it belongs to the word "SCHEMATIC", and also /AH/ when it is within another "MATHEMATICIAN". To solve such a problem, our proposed model first converts the input text/word into multiple phoneme substrings and then uses a combination of the obtained phoneme substrings as a new input pattern to predict the output phoneme corresponding to each input grapheme in a given word.

Since the performance of the neural network-based model for G2P conversion is limited, we use an existing weighted finite-state transducer (WFST)-based method implemented in the Phonetisaurus toolkit to implement our second proposed model. Except the acronyms and words with special pronunciations, we have figured out that most of the error words in G2P conversion are caused by the wrong prediction of their own vowel

graphemes. Therefore, we design several grapheme generation rules, which enable extra details (or sensitive information) for the vowel graphemes appearing within a word. These rules are applied to the input text/words at the first-stage of our proposed model. The evaluation results have shown that a G2P model using different rules can produce different output results that allow each rule to tackle different problems which may occur in different contexts during a conversion. This shows that a single approach does not suffice when addressing all the problems encountered by G2P conversion. Considering this fact, a combination of various approaches using different techniques is a reasonable strategy for treating the problems in a flexible manner.

Combining various techniques can both lend flexibility to the conversion and improve its predictive performance. Therefore, we present a phoneme transition network (PTN)-based architecture for G2P conversion. First, it converts a target word into multiple phoneme strings using different existing data-driven methods. Then, it aligns the obtained results—the phoneme-sequence hypotheses—using dynamic programming algorithm, combines them into a confusion network (or PTN), and determines the final output phoneme sequence by selecting the best phonemes from all the PTN bins—blocks of phonemes/transitions between two nodes in the PTN. Moreover, in order to extend the feasibility and improve the performance of the proposed PTN-based model to another higher level, we introduce a novel use of right-to-left (reversed) grapheme-phoneme sequences along with grapheme generation rules. Both techniques are helpful not only for minimizing the number of required methods or source models in the proposed architecture but also for increasing the number of phoneme-sequence hypotheses as well as new phoneme candidates, without increasing the number of methods. Therefore, the techniques serve to minimize the risk from combining accurate and inaccurate methods that can readily decrease the performance of phoneme prediction.

Various model combinations have been conducted and tested. Evaluation results using various word-based pronunciation dictionaries or datasets (such as NETtalk, Brulex, CMUDict and CMUDict_noisy) and K-fold cross-validation techniques show that our proposed PTN-based model, when trained using the reversed grapheme-phoneme sequences, often outperforms conventional left-to-right grapheme-phoneme sequences. In addition, the evaluation also demonstrates that the PTN-based method for G2P conversion is more accurate than all the baseline approaches that are tested in terms of both phoneme and word accuracy.

In the future, we plan to create new and effective grapheme generation rules to further improve our proposed approach, enabling a trained model to generate more accurately output phoneme-sequence hypotheses, such that only two models (using conventional and reversed grapheme-phoneme sequences) will be sufficient for our PTN model.