漢語方言語音資料庫自動擴增補完方法

林居正[*]、王昱鈞[**]、蔡宗翰[***]

摘　要

在歷史語言學研究之中，以語言田野調查收集語音資料為其研究方法重要之依據，故針對漢語的聲韻學研究領域，漢字於現代各種漢語方言中之發音資訊是探討漢語語音之歷史演變及漢語方言間之親屬關係的重要研究材料。為利於資料的保存與處理，並結合資訊技術之應用，已有部分漢語方言語音已進行數位化資料庫的建置。然而漢語分為七至八種方言群，因文化傳播與使用人口等因素，各漢語方言語音資料的完整度和數位化程度各不相同，許多漢語方言的語音資料庫仍未臻齊備，諸多漢字尚未有語音資訊。因此我們提出一個新的基於機器學習方法的模型用以自動擴增補全漢語方言語音資料庫中缺少的漢字發音資訊，該模型利用既有的方言語音內容以及從中古漢語韻書中找出於各方言中語音對應的模式藉以預測未知的漢字發音。我們以《漢語方音字彙》資料庫作為評估資料集，以漢字語音的音韻特徵如聲母、韻母、聲調等來評估擴增結果之準確率。藉由我們的方法，不僅能擴增漢語方言資料庫內容，其預測結果能進一步協助研究者發現具研究價值的語音現象，從而發展更多研究議題的可能性。

關鍵字：資料增補、生成模型、漢語方言、語音資料庫

[*] 國立臺灣大學資訊工程學研究所碩士，E-mail: chu.cheng.lin@gmail.com。
[**] 國立臺灣大學資訊工程學研究所博士生，E-mail: d97023@csie.ntu.edu.tw。
[***]（通訊作者）元智大學資訊工程學系副教授，E-mail: thtsai@saturn.yzu.edu.tw。

An Automatic Augmentation Method for Chinese Dialect Pronunciation Databases

Chu-cheng Lin[*]、Yu-chun Wang[**]、Richard Tzong-han Tsai[***]

Abstract

Field phonetic data collection is an important basis for the methodology of historical linguistics. In Chinese historical phonology, the phonetic qualities of Chinese characters in modern dialects are valuable materials. These are crucial for investigating diachronic Chinese phonology and the phylogenetic relationship among Chinese dialects. Dialectal pronunciation databases have been established to employ information technology in preserving and further processing of these phonetic qualities. As a large language family, Chinese can be divided into 7 to 8 dialect groups. As population and issues of culture propagation differ from one dialect to another, the collection of their pronunciation also differs greatly in terms of completeness and degree of digitalization. Many dialects still lack satisfactory databases. Many character pronunciations have yet to be recorded. Therefore, we propose a new machine learning-based model to augment Chinese dialectal pronunciation databases automatically, filling out missing character pronunciations. This model uses existing dialectal pronunciations and medieval rhyme books to find patterns across dialects to predict unknown character pronunciations. The Project DOC dataset is then used for evaluation. Phonological features, such as initial, rhyme, and tone are used to evaluate the accuracy of the results. The proposed model augments the pronunciation database. Moreover, the predictions made by the model may facilitate the discovery of interesting phenomena.

Keywords: data augmentation, generative model, Chinese dialects, pronunciation database

[*] Master, Department of Computer Science and Information Engineering, National Taiwan University. E-mail: chu.cheng.lin@gmail.com

[**] Ph.D. student, Department of Computer Science and Information Engineering, National Taiwan University. E-mail: d97023@csie.ntu.edu.tw

[***] Associate Professor, Department of Computer Science and Engineering, Yuan Ze University. E-mail: thtsai@saturn.yzu.edu.tw