使用帶中文字幕的台語劇自動擴增台語語音辨識之訓練資料的技術

本院覽號	公告日期	智財權狀態
05T-1100924	2021-01-10	know-how

摘要

台語自動語音辨識的一個明顯問題是訓練資料量遠遠不足以構建實用的系統。 收集具有可靠轉錄文本的語音資料以訓練聲學模型是可行的，但成本高昂。此外，因為台語是一種口語，而不是常用的書面語言，用於語言模型訓練的文本資料亦極其稀缺，且難以收集。YouTube上有大量帶有中文字幕的台語劇集，我們發展出一項技術可以自動增加台語語音辨識的聲學模型和語言模型的訓練資料。基本想法是使用現有的台語語音辨識系統，通過對照台語語音，參考中文及台文同義詞辭典，將中文字幕轉換為最可能的台文詞序列。實驗結果顯示，台語語音辨識系統可以通過這個訓練資料增強技術得到顯著改善。

技術優勢

- 新穎性：這是第一個利用台語劇中文字幕自動擴增台語語音辨識之訓練資料的技術。
- 簡單：可利用網路上大量的台語劇快速擴充台語語音辨識之訓練資料。

應用範圍

- 台語語音辨識系統開發
- 華語與台語機器翻譯系統開發
- 有相似資料的其他語言(例如客語及原住民語言)的語音辨識系統開發
- 有相似資料的其他語言(例如客語及原住民語言)的華語與該語言的機器翻譯系統開發

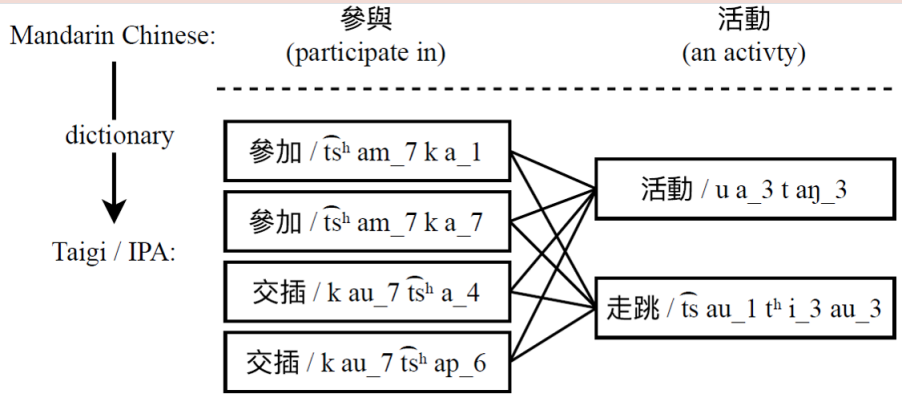

圖1.從中文單詞序列生成的示例台語單詞/音素圖。在 IPA 序列中，每個元音包含一個 IPA 符號，後跟一個音調代碼。

創作人

王新民、高明達

中央研究院 ACADEMIA SINICA