Strategies to Improve Children's Speech Recognition
改善孩童語音辨識之策略

鄭詠、李丞恩
指導老師： 劉奕汶 教授

Abstract

隨著科技進步，語音辨識的應用益發廣泛。從手機的語音助理、聊天機器人到智慧家庭的聲控模式，語音辨識已無所不在，近幾年由於深度學習的理論成熟，使得語音辨識系統如虎添翼，更加精準。但遇到一些特殊狀況，語音辨識的表現差強人意，比如孩童語音，孩童語音至今的辨識效果仍然不如大人語音，有以下原因：（一）孩童常有不正確發音、(二)孩童語言架構常有不合邏輯或顛倒語法的情況、(三)孩童語音有較大的音高變化、(四)缺乏足夠資料訓練模型。為了改善孩童語音辨識，我們研究出三種策略來改善辨識效果：其一為 pitch adaptation，藉由弭平孩童語音較大的音高變化以增加辨識率；其二為 speaker embedding，這種方法透過說話人識別（speaker identification）來加強孩童語音辨識系統的表現。最後為 language model interpolation，利用大人語言模型與孩童語言模型的混合達到降低錯誤率的效果。

為了實驗以上三種作法，我們搭建了語音辨識系統並使用深度網絡訓練模型。系統建立在開源語音辨識軟體 Kaldi 之上，Kaldi 具有整合語音辨識模型的功能，同時也支援深度網絡。最後，我們將分析三種方法的實驗結果並與基線系統（Baseline system）做比較。

Introduction

語音辨識系統的架構如圖 1 所示，通常由兩個部分組成。第一為 Acoustic model，亦即「聲學模型」，聲學模型掌控了音素的分類，也就是給定一個文字之後，發出這個語音的機率；第二為 Language model，我們稱為「語言模型」，語言模型將一個句子出現的機率拆解成其中每個單字出現的機率之積，得到一句話出現的機率大小 [1]。

前文提到，孩童因為發音咬字不成熟等原因，導致孩童語音在辨識上有許多困難，其錯誤率通常約在大人的 2-5 倍。我們研究了三種改善孩童語音辨識的策略，以下將分別介紹：

1. Pitch Adaptation [2]

我們分別提取大人與孩童語料庫中語者的音高變化，經由音高變化的變異數分析，我們發現，孩童在說話時明顯具有比大人更大的音高變化，如圖 2 所示。我們採用圖 3 的演算法來弭平孩童語音較大的音高變化。

2. Speaker embedding

Speaker embedding 是將每個 frame 中語者的 MFCC 隱藏的資訊，比如：音高、年齡、性別、情緒等額外特徵，利用統計的手法壓縮成一個向量，再將此向量和特徵一起作為神經網絡的訓練資料。我們使用的向量有 i-vector [3]與 xvector [4]，其中 x-vector 以圖 4 的神經網絡產生。

3. Language model interpolation

此方法利用大人語言模型與孩童語言模型的混合達到降低錯誤率的效果，我們將 baseline 的語言模型和孩童文本建立的語言模型利用 IRSTLM [5]由

$$\hat{P}(\omega_k|\omega_1 \cdots \omega_{k-1}) = \lambda \hat{P_1}(\omega_k|\omega_1 \cdots \omega_{k-1}) + (1-\lambda)\hat{P_2}(\omega_k|\omega_1 \cdots \omega_{k-1})$$

進行插值（interpolation）。

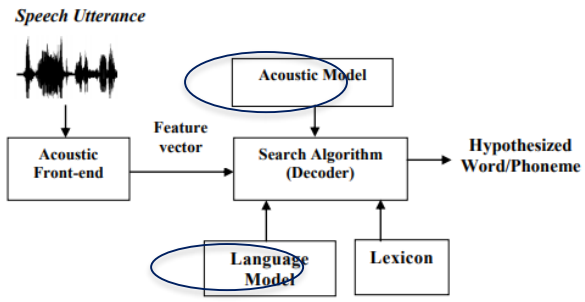

圖 1　語音辨識模型的架構[1]

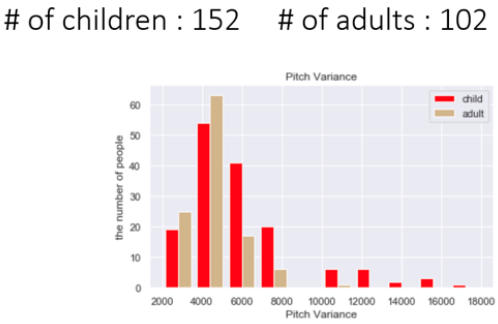

圖 2 大人與小孩語者的 Variance 分布圖

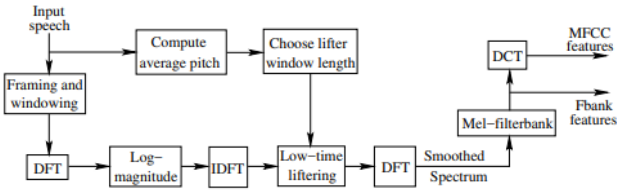

圖 3 Pitch-adaptation 之 MFCC 計算流程圖[2]

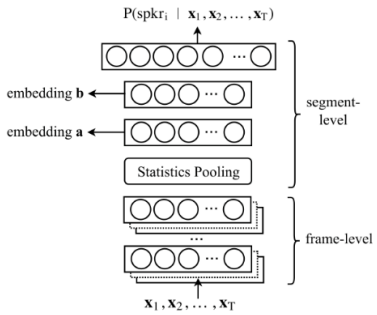

圖 4 抽取 xvector 之神經網路架構[3]

參考資料

[1] 圖片取自[1]。
[2] 圖片取自[2]。
[3] 圖片取自[4]。

[1] Karpagavalli S, Chandra E, “A review on Automatic Speech Recognition Architecture and Approaches”, International Journal Of Signal Processing, Image Processing and Pattern Recognition, Vol.9, Page 4, pp. 393-404, 2016.

[2] S. Shahnawazuddin, A. Dey, and R. Sinha, “Pitch-adaptive front-end features for robust children’s ASR,” in *Proc. INTERSPEECH*, 2016, pp. 3459–3463.

[3] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” in *IEEE Speech Recognition and Understanding Workshop*, pp. 55–59, 2013.

[4] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Proc. Interspeech*, 2017, pp. 999–1003.

[5] M. Federico, N. Bertoldi, and M. Cettolo, “IRSTLM: an open source toolkit for handling large scale language models,” in *Interspeech*, 2008, pp. 1618–1621.

心得感想

丞恩：本學年度的專題實作的是改善孩童語音辨識的策略研究。我主要學習到的項目有：（一）語音辨識的基本理論、（二）計算語音特徵向量 MFCC 的演算法實作，以及 Pitch Adaptation MFCC 演算法之實作、（三）i-vector 的數學理論以及（四）x-vector 的理論。雖然基於實作的觀點，必須在理論與實際操作中做出取捨，但我還是盡可能理解整個語音辨識系統的理論架構。語音辨識是一門機率統計以及數位訊號處理的交叉學科，尤其 i-vector 是利用高維度上的統計學、線性代數以及矩陣微積分的理論所發現的，是我個人認為研究的策略中最漂亮的部分，值得仔細探究，也因此激發我對統計學的興趣。我們所使用的工具是 Kaldi，但可能由於是較新興的工具，因此缺乏實體的教材，而網路上的資源大多較為零散，偏向解釋既有程式碼之教學。另外也由於不熟悉 Linux 系統，對我而言，依自己實際需求編寫 Kaldi 的程式碼(shell script)仍不大可行，初期在理解 Kaldi 的程式碼上也遭遇不少挫折。因此，我要特別感謝劉奕汶教授以及戴強麟學長的指導和安排，以及組員鄭詠的協助。

鄭詠:首先，特別感謝劉奕汶教授以及戴強麟學長這段期間不厭其煩的指導，雖然一開始我完全不懂語音辨識，但是教授與學長給了我們許多引導，讓這次的專題不但能順利完成，還收穫滿滿。回顧這一年實作專題，我遇到了不少困難，語音辨識的門檻極高，因此網路資源相對稀少，再加上語音辨識背後牽涉龐大的數學推導，對於學習上喜歡追根究底的我，真的花了很多時間理解背後的原理，記得有一次讀一篇論文，我花了整個週末研究論文裡面的一則數學式，儘管論文上只是短短的一行，在上網查詢資料時，我卻發現證明背後還有推導，推導裡面還有證明，不知不覺竟然因為一篇語音辨識的論文把機率、微積分還有線性代數複習了一回，現在回想，覺得自己執著於一則數學式有點好笑，但也因此知道這一年的實作專題沒有留白。近日剛好看到一則新聞，因為新冠病毒在全球爆發，人們突然間身處在害怕觸摸的世界，許多以接觸來控制的裝置，在未來可能以語音控制取代，新聞作者還下了一個有趣的標題：新冠病毒的大流行成為語音辨識「大流行」的強大推動力。看到這則新聞有些開心，我好

像也跟上了流行!再次感謝教授與學長的耐心與包容，以及一同奮鬥的隊友丞恩，希望未來我能用上這次實作專題所學!