探究語言模型合併策略應用於中英文語碼轉換語音辨識

Exploring Disparate Language Model Combination Strategies for Mandarin-English Code-Switching ASR

林韋廷　Wei-Ting Lin, 陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{60347014S, berlin}@ntnu.edu.tw

摘要

語碼轉換 (Code-Switching, CS) 在多語言社會中是一種常見的現象；例如，在台灣的官方語言是中文，但居民們日常對話時而會夾雜一些英文詞彙、片語或語句。語碼轉換語音的轉寫，在自動語言辨識(Automatic Speech Recognition, ASR) 上仍被視為一個重要且具有挑戰性的任務。而為了提升 CS ASR 效能，改進其語言模型是最直接且有效的方法之一。有鑒於此，我們提出多種不同階段的語言模型合併策略以用於中英文語碼轉換自動語言辨識。在本篇論文的實驗設定中，會有兩種中英文 CS 語言模型和一種中文的單語言模型，其中 CS 語言模型使用的訓練資料與測試集同一領域(Domain)，而單語言模型是用大量一般中文語料訓練而成。我們透過多種不同階段的語言模型合併策略以探究 ASR 是否能結合不同的語言模型其各自的優勢以在不同任務上都有好的表現。在本篇論文中有三種語言模型合併策略，分別為 *N*-gram 語言模型合併、解碼圖 (Decoding Graph) 合併和詞圖 (Word Lattice) 合併。經由一系列在企業應用領域的多種語料之實驗結證實，透過語言模型的合併的確能讓 CS ASR 對不同的測試集都有好的表現。

關鍵詞：語碼轉換、語言模型、語音辨識、解碼圖、詞圖

Abstract

Code-switching (CS) speech is a common language phenomenon in multilingual societies. For example, the official language in Taiwan is Mandarin Chinese, but the daily conversations of the ordinary populace are often mingled with English words, phrases or sentences. It is

generally agreed that transcription of CS speech remains an important challenge for the current development of automatic speech recognition (ASR). One of the straightforward and feasible ways to promote the efficacy of CS ASR is to improve the language model (LM) involved in ASR. Given these observations, we put forward disparate strategies that conduct combination of various language models at different stages of the ASR process. Our experimental configuration consists of two CS (i.e., mixing of Mandarin Chinese and English) language models and one monolingual (i.e. Mandarin Chinese) language models, where the two CS language models are domain-specific and the monolingual language model is trained on a general text collection. Through the language model combination at different stages of the ASR process, we purport to know if the ASR system could integrate the strengths of various language models to achieve improved performance across different tasks. More specifically, three strategies for combining language models are investigated, namely simple *N*-gram language model combination, decoding graph combination and word lattice combination. A series of ASR experiments conduct on CS speech corpora complied from different industrial application scenarios have confirm the utility of the aforementioned LM combination strategies.

Keywords: code-switching, language model, automatic speech recognition, decoding graph, word lattice

一、緒論

當在對話中使用兩種以上的語言時，這種現象稱為語碼轉換，根據語言切換情形又可細分為兩種類型：句子間的語碼轉換 (inter-sentential CS) 和句子內的語碼轉換 (intra-sentential CS)。其中又以 intra-sentential CS 的語音辨識任務較為困難，因為其語言切換的情形更多種，更難訓練出一個好的 CS ASR。

語碼轉換有幾種方法，可以簡單地分為四個方面：語言識別 (Language identification, LID)、資料增強 (data augmentation)、模型調適 (model adaptation) 和模型改進 (model improvement)。

首先，最常見的方法是用 LID 標記每個句子或每個單詞，然後分別使用每個單語言 ASR 系統進行語音識別。在 [1-7] 中提出了使用 LID 的相關任務。但使用這種方法的缺點是若是前端的語言辨識錯誤，後端的語音辨識就會錯誤，產生錯誤傳導的問題。

CS 訓練資料的缺乏也是訓練 ASR 模型的瓶頸之一。因此，資料增強在 CS ASR 中

也起到重要作用。資料增強包括聲學 (acoustic) 和文本 (textual) 資料增強，透過資料增強能訓練更加可靠的聲學和語言模型。在聲學資料增強方面，音檔可另外混和噪音 (noise) 或對其進行速度擾動 (speech perturbation) [8]，以及針對大量未標記資料的自動標記方法[9, 10] 或文本轉語音 (text-to-speech，TTS) 技術[11-13]產生更多的訓練資料被應用於聲學模型訓練。在文本資料增強方面，CS 語句可透過句子生成技術[9, 10, 14-17]產生而被用於語言模型訓練。

與資源較少的 CS 資料相比，單語言訓練資料更多。要如何利用大量的單語訓練數據來改善 CS ASR，常用的方法為使用 CS 資料來調適預訓練模型 (pre-trained model)，讓模型保有原本預訓練模型的效能外，也能有好的 CS ASR 效果。因此，以前有一些任務採用遷移式學習[5, 18-23]以利用大量的單語言資料來彌補 CS ASR 的資料稀疏性問題。

除了模型調適之外，還有一些任務是改進模型結構[4, 7, 24, 25]，讓模型可以學習處理多語言 ASR。此外，還引入了多任務學習[3-6]，使模型能同時學習 LID 和 CS ASR。期望 LID 資訊可以幫助 CS ASR 的提升。

CS ASR 還有其他較新穎的方法。在[26]中，他們提出了一種只使用單語言資料來訓練端到端 (End-to-end, E2E) CS ASR 的方法。對不同語言的輸出向量加上限制，讓每種語言的輸出向量的分布相近以達到語言轉換的效果。在[27]中也使用相同方法來訓練 CS 語言模型。在[28, 29]中，他們提出了一種用多解碼圖 (Multi-graph)進行解碼 (decode) 的方法應用於 CS ASR。 其中，多解碼圖為多語言解碼圖和單語言解碼圖結合而成。此種解碼方法可讓每個單語言或多語言語音辨識任務有平行且獨立的搜索空間，以更有效地使用每種語言的文本資源。

在本篇論文也將用到論文[28]的方法，和其他種語言模型的合併方法作比較。採用語言模型的合併技術其目的為結合 CS 語言模型和單語言模型各自的優勢以在不同任務上都能有好的表現。在接下來的章節二將介紹聲學模型的模型架構及方法，章節三將介紹三種不同階段的語言模型的合併方法，章節四則會簡述實驗設定及對實驗結果進行探討與分析。

二、聲學模型

(一) TDNN-F

時延神經網路 (Time Delay Neural Network, TDNN) 最早用於音素辨識[30]。因為難

以對語音訊號的時間定位有精確的標記，所以 TDNN 有時移不變性的特性，在語音辨識上會與時間位置無關。另外，TDNN 在建模時也會考慮上下文關係，每層隱藏層會接收到前層不同時間的輸出，舉例來說，若時間延遲 (time delay) 為 2，那就會考慮連續 3 個音框 (frame) 的特徵。藉由這種特性，讓 TDNN 可以表現出語音在時間上的連續關係，也可以考慮特徵序列的長時間相關性。

在論文[31]將 TDNN 應用於語音辨識，和原始的 TDNN 不同，多引進了子採樣 (sub-sampling)，只保留神經網路的部分連接 (connection)，因此降低計算量，加快訓練速度，也沒影響到原始的模型效能。在[31]實驗中也證明其訓練速度比深度神經網路(Deep Neural Network, DNN)、遞歸神經網路(Recurrent Neural Network, RNN)快，效能也相對提升。

TDNN-F[32]之後也被提出，和 TDNN 的差別主要有兩點，第一點就是將原來的權重矩陣分解成兩個矩陣，而且第二個矩陣需為半正交矩陣(semi-orthogonal matrix)。如此一來，不僅降低參數量，加快計算速度，也能保持相同的建模能力。第二點則是多了跳層連接(skip connections)，將前一層的輸出和當前層的輸出相加當作下一層的輸入，讓模型架構可以架得更深，且避免有梯度消失的問題。

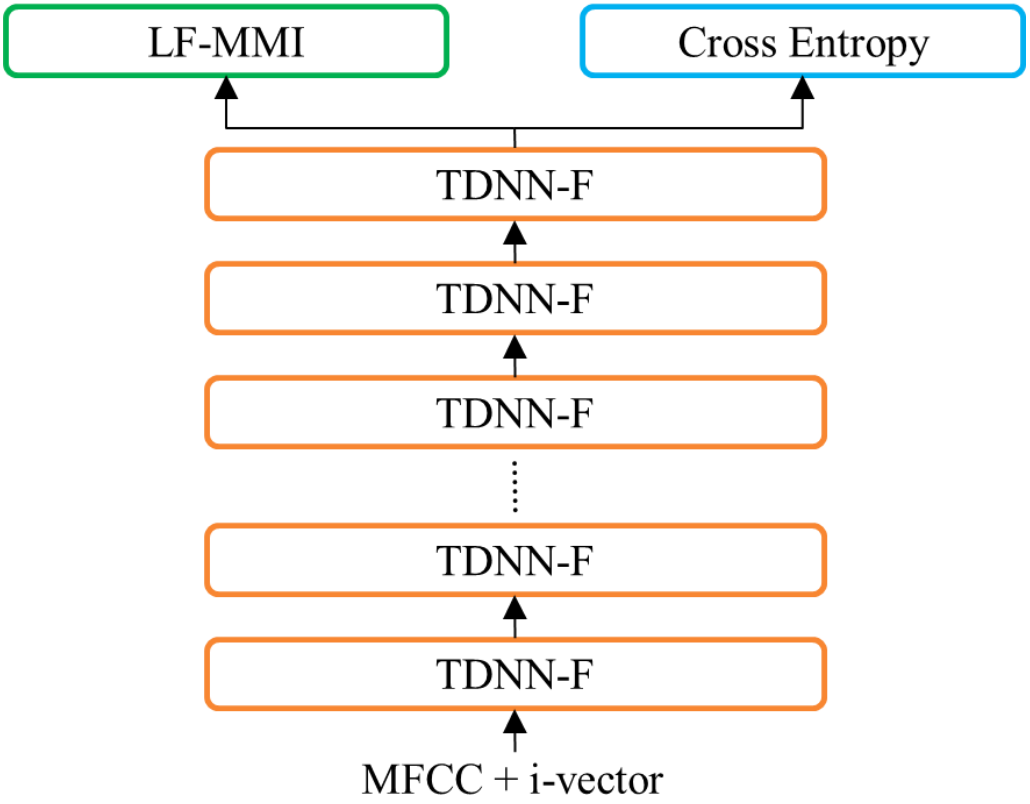
(二) LF-MMI

由於進行鑑別式訓練會提升聲學模型的效能，所以在語音辨識上除了以交叉熵(Cross entropy)作為損失函數訓練模型外，也會加上鑑別式訓練。進行鑑別式訓練前需要先進行交叉熵訓練，配合語言模型以產生 lattice，然而產生 lattice 是一個解碼的過程，會耗費不小的時間和空間的複雜度。

之後就有論文[33]提出了 Lattice-free Maximum Mutual Information (LF-MMI) 的方法，以音素或狀態 (state) 取代詞 (word) 作為語言模型的單元，使產生 lattice 的計算可以在 GPU 上進行，除了降低空間複雜度，也加快了 MMI 的訓練速度。

(三) Chain model

本篇論文實作的語音辨識工具為 Kaldi[34]，在 Kaldi 中的 chain model 使用了 LF-MMI 的鑑別式訓練，且另外加入了一些技巧使模型訓練更穩定、快速，例如：將隱藏式馬可夫模型 (Hidden Markov Model, HMM) 從三狀態改為單狀態的 HMM；使用音素作為語言模型的單元；加入交叉熵正規化 (cross entropy normalization) 進行多任務學習



圖一、TDNN-F 聲學模型架構

(multi-task learning)。其模型架構圖如圖一。

三、語言模型合併

(一) *N*-gram 語言模型合併

N-gram 是一種統計語言模型。假設一段 M 個詞組成的句子其機率為 $P(w_1, w_2, \dots, w_M)$，根據連鎖律 (chain rule) 可展開成 $\prod_{i=1}^{M} P(w_i|w_{i-1}, \dots, w_1)$，因為第 i 個字需考慮到前 $i-1$ 個字，若遇到較長的句子時，計算量會變大，所以會根據$n-1$ 階馬可夫假設 ($n-1$ order Markov assumption) 簡化，只需考慮前 n 個字，公式如下：

$$P(w_1, w_2, \dots, w_M) \approx \prod_{i=1}^{M} P(w_i|w_{i-n+1}, \dots, w_{i-1}) \quad (1)$$

不同的 *n*-gram 語言模型進行插值合併：

$$LM_{fusion} = \lambda LM_1 + (1-\lambda)LM_2 \quad (2)$$

其中 LM_* 為 *n*-gram 的機率，*n* = 3，即 trigram 語言模型。

(二) Graph 合併 (WFST 合併)

在 Kaldi 工具會使用 WFST 表示詞序列 (word sequences) 對應的 HMM state。在解碼時即可通過聲學模型計算出的 HMM 發射機率，搜尋出最佳路徑得出解碼結果。在 Kaldi 工具中以 HCLG[35]當作搜尋空間，其可分為四個部分：(1) G：從語言模型抽取的詞序列資訊；(2) L：存有每個詞對應的所有音素的發音辭典；(3) C： 音素的上下文關係；(4) H：HMM 的拓撲結構。融合成 HCLG 的過程中都會進行確定化 (determinization) 和最小化 (minimization)：

$$HCLG = \min(\det(H \circ \min(\det(C \circ \min(\det(L \circ G)))))) \quad (3)$$

其中 ∘ 為 composition，det 為 determinization，min 為 minimization。

如論文[28]，提出在 decode 端利用多個 graph 結合的 multi-graph 進行解碼。將不同的 graph 用聯集 (union) 結合成一個 graph，讓各個 graph 都有各自的搜索空間(searching space)，而不互相影響。

(三) Lattice 合併

Lattice 合併[36-38] 是將不同系統解碼後產生的 lattice，用不同的權重和損失函數進行合併：

$$W^* = argmin_W \frac{1}{N} \sum_{n=1}^{N} \sum_{W'} \lambda_n P_n(W'|\mathcal{O}) L(W, W') \quad (3)$$

其中 $P_n(W'|\mathcal{O})$ 為第 n 個 lattice 的後驗機率，λ_n為第 n 個 lattice 的結合權重，$L(W, W')$ 為詞序列W和W'間的 Levenshtein 編輯距離 (Levenshtein edit distance)。W^*為對各個系統的$\sum_{W'} P(W'|\mathcal{O}) L(W, W')$ 的平均透過最小化貝式決策風險 (Minimum Bayes Risk) 估計其最小值。

四、實驗設定與結果

(一) 資料集

本篇論文使用的語料為中英文混合會議語音資料，共有 230 小時，是國內某企業會議時錄製的語料庫。由於錄製內容都是實際對話狀況，所以會出現不一樣的說話腔調、頓點、速度等等。對話中除了一般對話外，也會出現一些專有名詞，甚至會突然出現英文專有名詞或日常用語。這語料庫因為沒經過特別設計，而且出現 CS 的問題 (包含 inter-sentential CS 和 intra-sentential CS)，所以挑戰性較大。

除了原本的訓練資料，另外加上了台灣的中文語料 Formosa 資料集和 YouTube 上收集的語料，除了擴增訓練集，也增加其豐富性。其中 Formosa 語料庫為從廣播、電視、開放課程等收集而來的真實的台灣自發性語音 (spontaneous speech) 中文語料，我們選了其中的 NER-Trs-Vol1、NER-Trs-Vol2 和 NER-Trs-Vol3 資料集和原有訓練集合併。另外也從 YouTube 上的開放課程或演講收集語料，其語料使用的語言為以中文為主，英文為輔。因為我們會把訓練資料中過長或過短句刪除，所以整合出的訓練集共為 635 小時。

而實驗中測試集有三種： (1) 會議的某一場錄音，是以中文為主的中英文 intra-sentential CS，共有 1 小時，資料集被命名為會議錄音測試集。(2) 短句測試集，內容為一些較少見的專有名詞，包含中文、英文和 CS 的專有名詞，共有 3 小時。(3) 長句測試集，由 33 位語者利用手機或平板錄製而成。100 句皆為經過設計過的語句，語句為

以中文為主的中英文 intra-sentential CS，共有 6 小時。

另外實驗中會用三種不同資料集來訓練 N-gram 語言模型，包括：(1) Meeting LM，用包含於 230 小時訓練集中的一部分會議錄音的文本資料訓練而成的語言模型。(2) General LM，用 Chinese Gigaword 資料集中的大部分繁體語料訓練而成的語言模型。(3) Keyword LM，用短句測試集的文本資料訓練而成的語言模型。其中，Meeting LM 和 Keyword LM 為 CS 語言模型，General LM 為中文單語言模型。其訓練資料細節如表一。

表一、語言模型的訓練資料細節

	詞彙數	字數
Meeting LM	551,141	1,605,545
General LM	627,819,651	1,534,226,867
Keyword LM	3,043	16,184

選擇用短句測試集的文本資料訓練一個語言模型，是為了實驗若結合和測試資料集同領域 (domain) 的語言模型，是否能在提升短句測試集辨識正確率時，也不會降低在其他測試集的辨識正確率。

(二) 實驗設定

在聲學特徵方面，會對音檔抽取 40 維的梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients, MFCC) 和 3 維的音調 (pitch)，並另外加上 100 維的 i-vector。

關於 TDNN-F 的模型訓練，參照了 Kaldi 的腳本 (script) 進行訓練。模型包含了 17 層 1536 維的 TDNN-F，每層的矩陣分解瓶頸 (bottleneck) 皆為 160 維。

實驗結果的評估方法採用混和錯誤率 (Mixed error rate, MER)，即英文採用詞錯誤率 (Word error rate, WER)，中文採用字錯誤率 (Character error rate, CER)。

(三) 實驗結果與分析

實驗結果如表二，方法(1)－(3)為只用單個語言模型進行解碼得出的結果，從數據可發現 Meeting LM、General LM 和 Keyword LM 分別在會議錄音測試集、長句測試集和短句測試集和其他語言模型相比都有較好的表現。

方法(4)－(6)為 Meeting LM 和 Keyword LM 在三種不同層次的合(結合比例為 1：

表二、不同層次的語言模型合併於測試集的 MER

測試集 / 方法	會議錄音測試集	短句測試集	長句測試集	平均
Meeting LM (M)－(1)	**27.85**	40.48	18.75	29.03
General LM (G)－(2)	41.72	39.83	18.54	33.36
Keyword LM (K)－(3)	94.31	**2.46**	80.05	58.94
N-gram LM 合併(M+K)－(4)	28.61	2.97	18.61	16.73
Graph 合併(M+K)－(5)	28.06	3.02	19.29	16.79
Lattice 合併(M+K)－(6)	70.07	3.25	49.28	40.87
N-gram LM 合併(G+K)－(7)	42.46	2.92	19.47	21.62
Graph 合併(G+K)－(8)	42.04	2.81	19.25	21.37
Lattice 合併(G+K)－(9)	84.66	3.28	46.98	44.97
N-gram LM 合併(M+G+K)－(10)	29.68	3.19	**15.82**	**16.23**
Graph 合併(M+G+K)－(11)	28.18	3.00	18.80	16.66
Lattice 合併(M+G+K)－(12)	68.06	3.68	40.44	37.39
Mixed LM (M+G+K)－(13)	35.57	6.55	16.70	19.61

1)；方法(7)－(9)為 General LM 和 Keyword LM 在三種不同層次的合併 (結合比例為 1：1)。由方法(1)可發現 Meeting LM 其實在長句測試集的 MER 和 General LM 只有些微差距，所以方法(4)－(5)能在三個測試集都有好的效果；反之，General LM 因為在會議錄音測試集的表現不好，所以方法(7)－(8)的平均 MER 比方法(4)－(5)差。另外透過方法(6)和(9)可發現 Lattice 合併只在短句測試集有好效果，其他測試集的 MER 都很高，我們推測因為 Keyword LM 是用特定領域的資料訓練而成，透過 Lattice 合併的方法相比其他方法較容易影響到其他語言模型的效能。

方法(10)－(12)為 Meeting LM、General LM 和 Keyword LM 在三種不同層次的合併 (結合比例為 1：1：1)。方法(12)和方法(6)、(9)一樣只在短句測試集有好效果，而方法(10)和(11)的平均 MER 比只用兩個語言模型合併還要低，證實了合併三個語言模型能讓 ASR 系統在三個測試集都有好的表現。

另外，方法(13)為將三個語言模型的訓練資料合併在一起再訓練成一個語言模型進行解碼得出的結果，從平均 MER 的比較結果可證明透過方法(10)和(11)的合併方法會比

方法(13)好。而在長句測試集上，方法(10)和(13)比(11)好的原因，我們推測在語言模型比較早的階段開始合併，會對長句測試集比較有幫助，因為方法(10)和(13)都是在建成解碼圖的階段之前合併，在其產生的解碼圖上，各個語言模型的搜索空間並非各自獨立的，彼此間會互相影響，所以會較易解碼出混合各個語言模型詞語的句子。但是在其他測試集上反而因為這種性質而導致 MER 較高。

五、結論

本篇論文採用多種不同層次的語言模型合併於 CS ASR 上，目的為結合不同語言模型的優勢以在不同的任務上都能有好的表現。透過實驗數據可發現 *N*-gram 語言模型的合併和 Graph 的合併被證實能有效地結合不同語言模型的優勢，並能在各個測試集上都有好的表現，最後以結合三種不同的語言模型的效果最好，也證實了比直接混合三種訓練資料訓練的語言模型表現得還要好。另外 Lattice 合併於實驗中因為較易受到一個特定領域資料訓練的語言模型所影響，而沒有預期的好表現於不同的測試集上。未來我們也將對這部分進行深入探討。

參考文獻

[1] Shinji Watanabe, Takaaki Hori, and John R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *Proc. ASRU*, 2017.

[2] Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey, “An end-to-end language-tracking speech recognizer for mixed-language speech,” in *Proc. ICASSP*, 2018.

[3] Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, et al., “Towards end-to-end code-switching speech recognition,” in *Proc. ICASSP*, 2019.

[4] Ke Li , Jinyu Li , Guoli Ye , Rui Zhao , and Yifan Gong, “Towards code-switching asr for end-to-end ctc models,” in *Proc. ICASSP*, 2019.

[5] Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, et al., “Investigating end-to-end speech recognition for mandarin-english code-switching,” in *Proc. ICASSP*, 2019.

[6] Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, et al., "On the End-to-End Solution to Mandarin-English Code-switching Speech Recognition," in *Proc. INTERSPEECH*, 2019.

[7] Metilda Sagaya Mary N J, Vishwas M. Shetty, and S. Umesh, "Investigation of Methods to Improve the Recognition Performance of Tamil-English Code-Switched Data in Transformer Framework," in *Proc. ICASSP*, 2020.

[8] Duo Ma, Guanyu Li, Haihua Xu, and Eng Siong Chng, "Improving code-switching speech recognition with data augmentation and system combination," in *Proc. APSIPA* 2019.

[9] Emre Yılmaz, Henk van den Heuvel, and David A. van Leeuwen, "Acoustic and Textual Data Augmentation for Improved ASR of Code-Switching Speech," in *Proc. INTERSPEECH*, 2018.

[10] Emre Yılmaz, Henk van den Heuvel, and David A. van Leeuwen, "Code-Switching Detection with Data-Augmented Acoustic and Language Models," in *Proc. SLTU*, 2018.

[11] Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Zero-Shot Code-Switching ASR and TTS with Multilingual Machine Speech Chain," in *Proc. ASRU*, 2019.

[12] Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li , et al., "End-to-end Code-switched TTS with Mix of Monolingual Recordings," in *Proc. ICASSP*, 2019.

[13] Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li, "End-to-End Code-Switching TTS with Cross-Lingual Language Model," in *Proc. ICASSP*, 2020.

[14] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung, "Learn to Code-Switch: Data Augmentation using Copy Mechanism on Language Modeling" in *Proc. ICASSP*, 2019.

[15] Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee, "Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation," in *Proc. INTERSPEECH*, 2019.

[16] Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee, "Code-switching Sentence

Generation by Generative Adversarial Networks and its Application to Data Augmentation," in *Proc. INTERSPEECH*, 2019.

[17] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che, "CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP," in *Proc. IJCAI*, 2020.

[18] Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung, "Towards End-to-end Automatic Code-Switching Speech Recognition," in *Proc. ICASSP*, 2019.

[19] Min Ma, Bhuvana Ramabhadran, Jesse Emond, Andrew Rosenberg, and Fadi Biadsy, "Comparison of Data Augmentation and Adaptation Strategies for Code-switched Automatic Speech Recognition," in *Proc. ICASSP*, 2019.

[20] Gustavo Aguilar and Thamar Solorio, "From English to Code-Switching: Transfer Learning with Strong Morphological Clues," in *Proc. ACL*, 2020.

[21] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, et al., "Meta-Transfer Learning for Code-Switched Speech Recognition," in *Proc. ACL*, 2020.

[22] Sanket Shah, Basil Abraham, Gurunath Reddy M, Sunayana Sitaram, and Vikas Joshi, "Learning to Recognize Code-switched Speech Without Forgetting Monolingual Speech Recognition," *arXiv:2006.00782*, 2020.

[23] Gurunath Reddy Madhumani, Sanket Shah, Basil Abraham, Vikas Joshi, and Sunayana Sitaram, "Learning not to Discriminate: Task Agnostic Learning for Improving Monolingual and Code-switched Speech Recognition," *arXiv:2006.05257*, 2020.

[24] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W. Black, "Sequence-based Multi-lingual Low Resource Speech Recognition," , in *Proc. ICASSP*, 2018.

[25] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta Learning for End-to-End Low-Resource Speech Recognition," in *Proc. ICASSP*, 2020.

[26] Yerbolat Khassanov, Haihua Xu, Van Tung Pham, Zhiping Zeng, Eng Siong Chng, et al., "Constrained Output Embeddings for End-to-End Code-Switching Speech Recognition with Only Monolingual Data," in *Proc. INTERSPEECH*, 2019.

[27] Shun-Po Chuang, Tzu-Wei Sung, and Hung-Yi Lee, "Training a code-switching language model with monolingual data," in *Proc. ICASSP*, 2020.

[28] Emre Yılmaz, Samuel Cohen, Xianghu Yue, David van Leeuwen, and Haizhou Li, "Multi-Graph Decoding for Code-Switching ASR," in *Proc. INTERSPEECH*, 2019.

[29] Xianghu Yue, Grandee Lee, Emre Yılmaz, Fang Deng, and Haizhou Li, "End-to-End Code-Switching ASR for Low-Resourced Language Pairs," in *Proc. ASRU*, 2019.

[30] Alexander H. Waibel, Toshiyuki Hanazawa, Geoffrey E. Hinton, Kiyohiro Shikano, and Kevin J. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[31] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015.

[32] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, et.al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. INTERSPEECH*, 2018.

[33] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016.

[34] Daniel Povey, Arnab Ghoshal, Ghoshal Boulianne, Lukas Burget, Ondrej Glembek, et al., "The kaldi speech recognition toolkit,"in *Proc. ASRU*, 2011.

[35] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukas Burget, Arnab Ghoshal, et al., "Generating exact lattices in the WFST framework," in *Proc. ICASSP*, 2012.

[36] Haihua Xu , Daniel Povey , Lidia Mangu , and Jie Zhu, "An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination," in *Proc. ICASSP*, 2010.

[37] Haihua Xua , Daniel Poveyb , Lidia Manguc , and Jie Zhua, "Minimum Bayes Risk Decoding and System Combination Based on a Recursion for Edit Distance," in *Proc. CSL*, 2011.

[38] Tien-Hong Lo and Berlin Chen, "Leveraging Discriminative Training and Model Combination for Semi-supervised Speech Recognition," in *IJCLCLP*, 2018.