使用元學習技術於語碼轉換語音辨識之初步研究

A Preliminary Study on Leveraging Meta Learning Technique for Code-switching Speech Recognition

余福浩 Fu-Hao Yu
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
M10815004@mail.ntust.edu.tw

陳冠宇 Kuan-Yu Chen
國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

摘要

語碼轉換(Code-switching)的語音辨識近年來愈來愈受到研究學者的重視，雖然人們在日常生活中使用語碼轉換的情形逐漸增加，但是可以用來訓練語碼轉換之語音辨識器的語音資料，相較於主流的單語言（如：英語或中文）語料，更是少之又少，這是因為語音語料的標註(Labeling)相當耗時費工。為了提升語碼轉換之語音辨識器的效能，在本研究中，我們提出利用近期逐漸興起的元學習(Meta Learning)方法，希望可以利用資料量較多的單語言語料，提升語碼轉換的語音辨識任務之成效。實驗中，我們採用 SEAME (South East Asia Mandarin-English)資料集，透過元學習中的與模型無關之元學習法(Model-Agnostic Meta-Learning, MAML)進行語音辨識器訓練，實驗結果證明，模型能夠快速適應於最終的語碼轉換語音辨識任務。

關鍵詞：語碼轉換，語音辨識，元學習

Abstract

In recent years, code-switching speech recognition has become an important research topic. Code-switching in conversation speech is gradually increasing in our daily lives. However, compared with monolingual languages (e.g., English or Chinese), only a few resources can be obtained for training a code-switch speech recognizer. To mitigate the deficiency, in this paper, we propose a meta-learning approach for code-switching speech recognition. In other words, following the model-agnostic meta-learning (MAML) procedure, we first train the speech recognizer by using monolingual corpora, and then a fine-tune stage is performed to obtain the final code-switching speech recognizer by using code-switching data. We evaluate the proposed method on the SEAME (South East Asia Mandarin-English) dataset. A series of experiments show that the meta-learning method can improve the performance of the low-resource code-switching speech recognition task.

Keywords: Code-switching, Speech Recognition, Meta-learning

一、緒論

語音辨識中語碼轉換(Code-switching)的議題目前愈來愈受到重視。在全球化的基礎之下，人類學習多種語言的情形開始蔚為風潮，如非英語系國家通常會學習英語或學習使用頻率較高的主流語言作為第二外語來使用，在這樣的時空背景下造成人們在說話時產生語碼轉換的可能性日益增加，以亞洲地區為例，於許多不同的場景中，人們在對話或溝通時開始產生容易在中文間穿插英語等外語的情形，從大專院校校園甚至到工作職場中，這樣的狀況逐漸普及，因此含有語碼轉換的語音辨識在自然語言領域開始成為一個重要的問題。

近年來，由於深度神經網路的興起，深度學習在影像辨識任務上獲得了非常優秀的成績，辨識準確率甚至能夠超越人類，因此研究學者們開始傾向使用深度學習來進行機器翻譯和語音辨識等自然語言方面的任務[1, 2, 3, 4, 5]，研究成果顯示，基於深度學習的模型可以獲得比傳統模型還要傑出的成績，因而此類方法逐漸變成了近年自然語言研究的主流與趨勢。

在使用深度學習方法來訓練模型時，通常必須藉由大量的訓練資料進行訓練，才能夠成功訓練出良好的模型，若模型的訓練資料過於稀少，則容易產生擬合不足

(Underfitting)的問題，使模型在訓練集和測試集的表現都不夠良好。在語音辨識的任務中，含有語碼轉換的語音資料相當稀少，這使得要訓練一個成功的語碼轉換語音辨識器變得不太容易，因此有許多研究開始將含有語碼轉換的語音辨識視為低資源(Low-resource)的任務，嘗試用不同的訓練方式或模型方法來解決這個問題。有鑑於此，本論文提出使用元學習(Meta Learning)的方式來進行含有語碼轉換的語音辨識，並嘗試使用目前主流的元學習方法中與模型無關之元學習法(Model-Agnostic Meta-Learning, MAML)[8]於語碼轉換之語音辨識，期望可以提升語音辨識的準確率。

二、相關研究

（一） 元學習(Meta Learning)

元學習(Meta Learning)是近年來逐漸興起的一種深度學習方式，元學習的理論背景建立在希望所訓練出的深度學習模型，能夠有如同人類的學習行為，人類對於學習一件新事物的能力很強，通常不用像深度學習模型一樣要看過大量學習資料才能做得很好，例如可以只看過一幅特定畫家的畫作就能大致判斷大多數的畫作是否出自於此畫家，或是只須看過一個器物的樣貌，就能夠成功辨認出相同器物的不同樣式或形狀。元學習理論認為人類之所以能夠學習得這麼快速，是因為已經累積了很多先前學習的經驗，所以才能夠達到快速學習(Rapid Learning)的能力。因此，在元學習理論中，為了使模型能過獲得快速學習的能力，將透過額外的訓練資料產生不同的訓練任務(Task)，讓模型「學習如何學習(Learning-to-learn)」，使模型在面對目標任務時並不是從頭開始學起，而是有了過往學習的經驗或知識，也就是具備了一定的先驗知識，並學會了「如何學習」的經驗與技巧，成為了更厲害的學習者，藉由過往的學習經驗，深度學習模型在未來遇到目標任務時，便可以利用少量的資料達到快速學習的成果。

近期由於元學習逐漸受到重視，在研究方面發展出三大不同的主流方法，分別為：以含有記憶性功能的神經網路，如使用具有長短期記憶遞迴神經網路[18]、神經網路圖靈機(Neural Turing Machines, NTM)[19]來實踐的黑箱適應方法(Black-box Adaptation)[6, 7]；以及基於最佳化(Optimization-based)的元學習方法[8, 9]，像是與模型無關之元學習(Model-Agnostic Meta-Learning, MAML)[8]和可擴展的元學習演算法(Reptile)[9]等方法；最後則是基於測度(Metric-based)理論的非參數化(Non-parametric)方法，例如著名的孿生網路(Siamese Network)[10]等。其中，與模型無關之元學習方法更成為目前最主流的方

法之一，與模型無關之元學習方法屬於一種基於最佳化方式來達成元學習概念的方法。元學習旨在希望能夠訓練出一個模型成為好的學習者（即元學習者(Meta-learner)），在各種學習任務下只需要利用少量的訓練資料，就可以快速地解決或適應一個新的學習任務，也就是希望我們訓練出的元學習者能夠像是人類一樣，能夠在少量的訓練下達成快速學習的目標；而與模型無關之元學習演算法便發展出了一種與模型無關的元學習方式，由於他在演算法的設計上是完全與模型無關，因此我們便能夠使用梯度下降的訓練方式，直接地應用在任何一種學習問題和模型上，例如：分類(Classification)、迴歸(Regression)和強化學習(Reinforcement Learning)等常見的不同任務中，和以往發展的元學習方法之不同點在於與模型無關之元學習方法不會增加模型的參數量，也不需要限制模型的架構，因此模型不受限於各種遞迴神經網路，也可以和全連接(Fully-connected)神經網路與卷積神經網路(Convolutional Neural Network, CNN)等不同神經網路進行組合。透過與模型無關之元學習演算法的訓練，可以使訓練後的模型在給定一個新的學習任務時，能夠快速地適應於新任務且在新任務上能有較好的結果。

（二） 語碼轉換語音辨識模型之現況

傳統上常用來進行語碼轉換語音辨識任務的模型有像是基於高斯混合模型結合隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)[20]或是深度類神經網路結合隱藏式馬可夫模型(Deep Neural Network-Hidden Markov Model, DNN-HMM)的語音辨識器 [15, 16]，而近年來由於端對端(End-to-end)的語音辨識器成為研究主流，因此近期亦有基於端到端的語碼轉換語音辨識器模型，例如基於 CTC 以及注意力機制的混合模型(Hybrid CTC-Attention based Models)[12, 17]，並嘗試結合加入語言辨識(Language Identification, LID)進行多任務學習(Multitask Learning)的訓練策略，來改善語碼轉換語音辨識的準確度。

三、方法

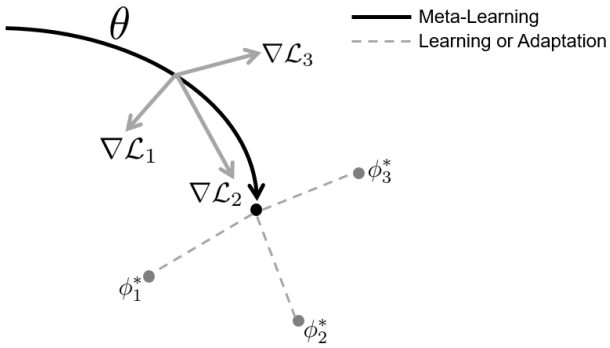
由於目前並不容易取得擁有大量訓練資料的語碼轉換語音資料集，這使得基於深度學習的語音辨識模型難以在語碼轉換的任務上有良好的表現。所幸的是，近期開始興盛起的元學習方法，目標便是使得訓練出的模型能如同人類一般，在少量的資料下能夠做到快速學習的學習行為，也正因為如此元學習方法十分適合使用在訓練資料不多、低資源(Low-resource)的學習任務[13, 14]。在語碼轉換的語音辨識任務上，我們認為在宏觀的

概念上，可以將含有語碼轉換的語音視為一種新的語言，若我們將它當成一種新的語言來看待，當然也就能夠將語碼轉換這個語言和單語言當作不同的訓練任務，基於這個想法，我們就可以利用元學習的概念，希望藉由單語言語料訓練出的模型，能夠快速地適應於目標任務，也就是含有語碼轉換的語音辨識任務上。有鑑於此，本論文提出使用元學習中，與模型無關之元學習來進行語音辨識器的訓練，期望在單語言語料的訓練下，可以獲得一個更好的語碼轉換語音辨識器，進一步地提升語碼轉換語音辨識的準確率。

（一） 與模型無關之元學習(Model-Agnostic Meta-Learning, MAML)

在與模型無關的元學習方法中[8]，我們首先定義模型為f，負責進行預測任務，在給定輸入資料x即可映射到輸出資料y，與模型無關的元學習演算法和傳統元學習的概念相同，訓練過程可分為元訓練(Meta-training)與元測試(Meta-testing)兩大部分，在進行元訓練時，可將原本的訓練資料集$\mathcal{D}$經過隨機抽樣產生許多不同的子資料集$\mathcal{D}_i = \{(x_1, y_1), \cdots, (x_n, y_n)\}$，此時每一個子資料集都可以切分成訓練資料（或稱為支撐集(Support Set)）$\mathcal{D}_i^{tr} = \{(x_1, y_1), \cdots, (x_k, y_k)\}$ 以及測試資料（或稱為查詢集(Query Set)）$\mathcal{D}_i^{ts} = \{(x_1, y_1), \cdots, (x_l, y_l)\}$，此時便可以將每一個子資料集視為一個不同的任務$\mathcal{T}_i$，也就是說$\mathcal{D}_{meta-train} = \{\mathcal{D}_1, \cdots, \mathcal{D}_n\} = \{(\mathcal{D}_1^{tr}, \mathcal{D}_1^{ts}), \cdots, (\mathcal{D}_n^{tr}, \mathcal{D}_n^{ts})\} = \{\mathcal{T}_1, \cdots, \mathcal{T}_n\}$，此外若是每一個任務的訓練資料集$\mathcal{D}_i^{tr}$都有$k$筆訓練資料，我們也可稱之為 K 樣本學習(K-shot Learning)。而在元測試時也可定義出$\mathcal{D}_{meta-test} = (\mathcal{D}^{tr}, \mathcal{D}^{ts})$，也就是我們想適應於目的領域或目的任務的訓練與測試資料，我們會利用其中的訓練資料$\mathcal{D}^{tr}$訓練模型，使模型能夠快速適應在新的目標任務中，最後便可以將模型應用於新任務的測試資料中進行預測，以達到比較好的預測結果。

更明確地，我們希望能夠在元訓練資料集$\mathcal{D}_{meta-train}$上找到一個較好的元參數θ，利用此參數在新的任務上得到一個好的模型參數ϕ，在與模型無關的元學習演算法中，我們將模型的初始化參數視為元參數θ，將模型記為f_θ，當要訓練一個新任務$\mathcal{T}_i$時，透過訓練資料與誤差函數$\mathcal{L}$計算出的損失$\mathcal{L}_{\mathcal{T}_i}$，可以利用梯度下降的方式來更新模型的參數（通常只進行一次，但也可直觀地擴展為多次更新），在與模型無關的元學習方法中，經過參數更新適應於新任務的模型參數即為ϕ_i，也就是說$\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$，其中$\alpha$為梯度下降的可調整參數。與模型無關的元學習方法希望模型的初始化參數θ更新變為



圖一、與模型無關的元學習方法參數優化路徑示意圖。

ϕ_i後，能夠在新任務$\mathcal{T}_i$上有良好的表現，因此我們可以訂定出元學習的元目標函數(Meta Objective Function)為：

$$\min_{\theta} \sum_{i} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i}) = \sum_{i} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\theta})}) \quad \text{式 (1)}$$

意即希望能找到一個最好的初始化參數，使得此參數在新任務上只需透過一次梯度下降進行參數更新，就能在新任務上有最好的表現，也就是使損失最小化。此外，要優化元學習目標函數同樣是一個最佳化問題，我們可以再次利用梯度下降的方式對元學習的目標函數進行最佳化（又稱為元最佳化(Meta Optimization)），代表元參數θ可依照公式$\theta \leftarrow \theta - \beta \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$進行更新，其中$\beta$亦為梯度下降的可調整參數。與模型無關的元學習方法的優化過程可以參考圖一之視覺化示意圖。

完整的與模型無關的元學習方法之演算法可參考以下演算法 1 之虛擬碼，值得一提的是，與模型無關的元學習方法如同上述介紹會有兩種最佳化，第一種最佳化於演算法 1 中第 6 行，是負責進行讓模型適應於新任務之中的最佳化，又稱為內層迴圈最佳化(Inner Loop Optimization)；第二種最佳化於演算法 1 中第 8 行，則是負責讓模型能夠快速適應於新任務，使模型能成為一個更好的學習者的最佳化，又稱為外層迴圈最佳化(Outer Loop Optimization)。

雖然與模型無關的元學習演算法不會增加模型的參數量，但是卻有著在進行外層迴圈最佳化時需要計算$\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$的缺點，而$\nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$的計算則是牽涉到需要對損失函數進行二階微分(Second-order Derivatives)，要在高維度的向量空間中進行二階微分則必須用 損失建立出其完整的黑塞矩陣(Hessian Matrix)，代表在進行損失函數優化時計算量

演算法 1：Model-Agnostic Meta-Learning（MAML）

輸入：α, β

1: 隨機初始化模型參數 θ
2: **while** 訓練尚未完成 **do**
3: 產生一個（或多個）新任務 $\mathcal{T}_i$
4: **for all** $\mathcal{T}_i$ **do**
5: 利用 $\mathcal{D}_i^{tr}$ 計算 $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
6: 計算參數 $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: 更新模型初始化參數 $\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$
9: **end while**

演算法 1、與模型無關的元學習方法之演算法。

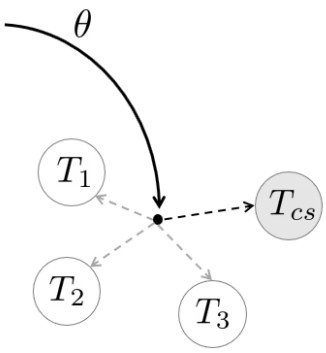
將會大幅提升，並消耗更多記憶體空間，使得模型的訓練速度大幅下降。為了避免計算出完整的黑塞矩陣，我們也可考慮利用一階近似(First-order Approximation)的方式來計算出$\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$的近似值，以對演算法加速。在原本的梯度計算中，由於$\phi = \theta - \alpha \nabla_\theta \mathcal{L}(f_\theta)$，我們可先利用連鎖律對其展開，接著則會出現一項$\nabla_\theta\big(\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta)\big)$的梯度計算，由於$0 < \alpha < 1$使$\alpha \nabla_\theta \mathcal{L}(f_\theta)$很接近 0，此項會十分接近$\nabla_\theta(\theta)$，因此可以選擇忽略計算此項內部的梯度，直接以$\nabla_\theta(\theta)$進行近似，而又因$\nabla_\theta(\theta) = I$所以經過一階近似後可得$\nabla_\theta \mathcal{L}(f_\phi) \approx \nabla_\phi \mathcal{L}(f_\phi)$：

$$\begin{aligned} \nabla_\theta \mathcal{L}(f_\phi) &= \big(\nabla_\phi \mathcal{L}(f_\phi)\big)(\nabla_\theta \phi) \\ &= \big(\nabla_\phi \mathcal{L}(f_\phi)\big)\big(\nabla_\theta(\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta))\big) \\ &\approx \big(\nabla_\phi \mathcal{L}(f_\phi)\big)(\nabla_\theta(\theta)) = \nabla_\phi \mathcal{L}(f_\phi) \end{aligned} \qquad \text{式 (2)}$$

利用一階近似方法實作的與模型無關的元學習演算法又特稱為一階與模型無關的元學習演算法(First-order MAML, FOMAML)，即在計算上以$\theta \leftarrow \theta - \beta \nabla_{\phi_i} \mathcal{L}_{\mathcal{T}_i}(f_{\phi_i})$取代原本於演算法 1 中第 8 行的更新式，即可達成一階近似[8]。

（二）基於元學習的語碼轉換語音辨識系統

由於希望能夠藉由使用元學習技術改善語碼轉換任務的語音辨識準確度，我們將語碼轉換的語音資料作為目標任務，即用來進行元測試階段的訓練，而訓練集中僅含有中英文之單語言資料則用於元訓練階段的訓練和測試中。在進行元訓練階段時，每一次都將從



圖二、使用元學習的概念於語碼轉換語音辨識的任務T_{cs}。

單語言資料中隨機抽取出部分資料，於元學習中可將其視為一個新的語音辨識任務，接著使用與模型無關之元學習方法來訓練、更新模型，則可使模型學習到要如何快速適應於一個新的語音辨識任務之中。最後在元測試階段，則使用含有語碼轉換情形的語料對模型進行訓練，使其能夠適應於我們想解決之目標任務，由於模型已經學會如何快速適應於不同的語音辨識任務，因此就算語碼轉換的訓練資料不如單語言語音資料多，模型依然能夠快速適應於語碼轉換的語音辨識任務，且能夠有較好的表現。圖二為使用元學習的概念於語碼轉換語音辨識系統的示意圖。

四、實驗

為了瞭解使用元學習方法是否能在語碼轉換的資料集上帶來改善語音辨識的效果，我們採用 LAS(Listen, Attend and Spell)[4]架構作為語音辨識器，並使用與模型無關之元學習演算法的方式進行元學習實驗，並與傳統語音辨識系統的結果進行比較。

（一） 資料集

我們使用 SEAME (South East Asia Mandarin-English)[11, 12]作為實驗的資料集，由於想了解元學習對語碼轉換語音辨識是否帶來改進，我們將資料集中的訓練資料進行分類，根據語音資料文字中是否為語碼轉換資料來將訓練資料分成兩大部分，分別為只有中文或只有英文的單語言資料以及含有中英文夾雜的語碼轉換資料，測試集則使用 SEAME 中的 dev_man 與 dev_sge 資料合併進行測試，相關統計資訊如表一所示。最後，我們使用字錯誤率(Character Error Rate, CER)與詞錯誤率(Word Error Rate, WER)作為評估標準。

	小時數	小時(比例)		
		純中文	純英文	語碼轉換
$train$	101.13	16.18 (16%)	16.18 (16%)	68.76 (68%)
dev_{man}	7.49	1.04 (14%)	0.52 (7%)	5.91 (79%)
dev_{sge}	3.93	0.23 (6%)	1.61 (41%)	2.08 (53%)

表一、SEAME 資料集統計資訊。

Model	CER	WER
LAS	55.0%	63.4%
LAS+MAML	**49.7%**	**58.9%**

表二、實驗結果。

（二） 語音辨識模型

實驗時我們採用 Google 於 2015 年提出之 LAS(Listen, Attend and Spell, LAS)模型[4]作為語音辨識器使用，LAS 模型由金字塔式堆疊的雙向長短期記憶(Long Short-term Memory)作為編碼器(Encoder)，以及含有注意力(Attention)機制的解碼器(Decoder)組合而成，結合以上兩種特殊機制在語音辨識上取得了非常好的結果。實驗中 LAS 模型依照以下設定進行設置，在編碼器中使用三層雙向各 180 維的長短期記憶，在解碼器中使用兩層 360 維的長短期記憶，詞嵌入(Word Embedding)的維度大小為 180 維，以 Uniform Distribution [-0.1,0.1]進行模型初始化，其餘設置參考原模型設定。

（三） 實驗結果

實驗中，基礎系統使用 LAS 模型訓練於僅有中文或英文之單語言資料 5 個世代(Epoch)後，再更新(Fine-tune)於語碼轉換的資料 5 個世代，最後的字錯誤率約為 55.0%、詞錯誤率為 63.4%。當我們將元學習方法運用於語碼轉換之語音辨識系統時，首先是使用單語言資料作為元訓練資料集$\mathcal{D}_{meta-train}$，每一次都將隨機從元訓練集中隨機取出 8 筆資料作為訓練資料以及 8 筆資料作為測試資料，外層迴圈使用學習率 0.001 的適應性矩估計演算法(Adam)進行優化，內層迴圈使用學習率 0.1 之隨機梯度下降法(Stochastic Gradient Descent, SGD)進行優化，之後則將語碼轉換的資料作為元測試$\mathcal{D}_{meta-test}$中的訓練資料並訓練 5 個世代，讓模型快速適應於語碼轉換的目標任務，最後在測試集上進行測試，最後的字錯誤率約為 49.7%、詞錯誤率為 58.9%。最終實驗結果如表二所示。

四、結論

本論文提出以元學習方式改善語碼轉換語音辨識的方法，並使用近期元學習中的主流方法與模型無關之元學習方法進行語音辨識器的訓練，由於目前語碼轉換的訓練資料稀少，以元學習的方式進行學習，可以在相同的訓練資料集之下，訓練出更好的語音辨識器，於實驗中我們所提出的方法，可以在字錯誤率與詞錯誤率上獲得改善，在這樣的研究結果下，我們為語碼轉換語音辨識提供了一種新的解決方法，未來，我們將繼續這個研究方向，期望可以結合其他單語言資料集或是對資料進行資料擴增(Data Augmentation)，以及發展出一套專屬於語碼轉換或語音辨識器的元學習演算法，為語音辨識任務提供一個新的方向與效能的提升！

致謝

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 109-2636-E-011-007 (Young Scholar Fellowship Program), and by the Project K367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

參考文獻

[1] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[2] A. Graves, “Sequence transduction with recurrent neural networks,” in *ICML Representation Learning Worksop*, 2012.

[3] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in *Neural Information Processing Systems*, 2015.

[4] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

[5] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4835–4839.

[6] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.

[7] T. Munkhdalai and H. Yu, “Meta networks,” in *Proc. ICML*, 2017, pp. 2554–2563.

[8] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” *arXiv preprint arXiv:1703.03400*, 2017.

[9] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.

[10] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *ICML deep learning workshop*, 2015, vol. 2: Lille.

[11] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," *arXiv preprint arXiv:1811.00241*, 2018.

[13] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.

[14] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 7844-7848.

[15] E. Yılmaz, H. v. d. Heuvel, and D. A. van Leeuwen, "Acoustic and textual data augmentation for improved ASR of code-switching speech," *arXiv preprint arXiv:1807.10945*, 2018.

[16] P. Guo, H. Xu, L. Xie, and E. S. Chng, "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition," *arXiv preprint arXiv:1806.06200*,

2018.

[17] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[19] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.