Computational Linguistics and Chinese Language Processing
Vol. 27, No. 2, December 2022, pp. 31-46

探討語者驗證系統中特徵處理模組與注意力機制

Investigation of Feature Processing Modules and Attention Mechanisms in Speaker Verification System

陳廷威 *、林威廷*、陳嘉平*、呂仲理 +

詹博丞+、鄭羽涵+、莊向峰+、陳威妤+

Ting-Wei Chen, Wei-Ting Lin, Chia-Ping Chen, Chung-Li Lu,

Bo-Cheng Chan, Yu-Han Cheng, Hsiang-Feng Chuang, Wei-Yu Chen

摘要

本論文建構並替換不同的音訊特徵前處理模組與注意力機制來改進語者驗證系統。我們使用了基於 ECAPA-TDNN 所改進的模型作為基準模型，並透過替換與組合不同的前處理模組與注意力機制來進行比較，以選出最佳的組合作為論文提出的最終模型。訓練上我們使用了 VoxCeleb 2 資料集進行訓練，並使用多個測試集來測試模型的表現。最終模型在 VoxSRC2022 驗證集中對比基準模型有 16% 的進步幅度，成功在語者驗證系統上取得了更好的成效。

Abstract

In this paper, we use several combinations of feature front-end modules and attention mechanisms to improve the performance of our speaker verification system. An updated version of ECAPA-TDNN is chosen as a baseline. We replace and integrate different feature front-end and attention mechanism modules to compare and find

* 國立中山大學資訊工程學系
Department of Computer Science and Engineering, National Sun Yat-sen University
E-mail: {m103040017, m093040020}@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw
+ 中華電信研究院
Chunghwa Telecom Laboratories
E-mail: {chungli, cbc, henacheng, gotop, weiweichen}@cht.com.tw

the most effective model design, and this model would be our final system. We use VoxCeleb 2 dataset as our training set, and test the performance of our models on several test sets. With our final proposed model, we improved performance by 16% over baseline on VoxSRC2022 valudation set, achieving better results for our speaker verification system.

關鍵詞：語者驗證、前處理模組、注意力機制、時延神經網路
Keywords: Speaker Verification, Frontend Module, Attention Mechanism, Time Delay Neural Network

1. 緒論 (Introduction)

隨著資訊科技的日新月異，大量的數位化資訊充斥在我們的生活當中，透過各式各樣新穎的設備，任何事物、資料都可以被電子化的儲存，並隨時傳送到地球上的任何地方，這使得人們得以跳脫固有的時間與空間上限制，能以更為宏觀的視角來探索這個世界。然而，當每個人對這些通訊設備的依賴度越來越高時，個人資訊被不合法的洩漏、利用的情形也日漸增加，如何保護自身的資訊安全是一個非常迫切的議題。

語者辨識技術便是其中一項在近年來越來越受到重視的資訊防護方法，藉由這項技術，我們可以將語者的聲紋特徵轉換成具有語者特徵的嵌入向量，透過比對這個嵌入向量來對當前語者的身分進行確認，以防止個人資訊被偽造及竊取。

近年來，有許多過去在圖像領域發光發熱的模型結構被帶入到聲學領域當中，並為語者驗證技術帶來了極大的突破，像是以時延神經網路（Time Delay Neural Network, TDNN）作為主幹，並在其中引入了 Res2Net (Gao *et al.,* 2021) 多分支卷積結構與 SENet (Hu *et al.,* 2018) 注意力機制的 ECAPATDNN (Desplanques *et al.,* 2020) 與基於傳統二維卷積神經網路建構的 ResNet (He *et al.,* 2016a)，兩者都在近年的語者驗證競賽中取得亮眼的表現。而鑒於兩種截然不同架構都在競賽上取得優秀的成果，希望能夠集合兩種架構優點的新型架構被研究出來，也就是 ECAPA CNNTDNN (Thienpondt *et al.,* 2021)。在該模型中，ResNet 結構被設計為 ECAPA-TDNN 的前處理模組，用於降低輸入音檔特徵頻譜圖在頻率軸上的偏移，透過卷積操作重組與保留較重要之特徵訊息。該結構在實驗上進一步的提高模型的表現，並為語者驗證模型的變化性增加了更多的可能性。

在本篇論文中，我們使用基於 ECAPATDNN 架構進行改進的 Improving ECAPATDNN (Zhang *et al.,* 2021) 做為基底，透過修改部份結構以提出 IM ECAPA-TDNN 做為本次的基準模型，並將其依照 ECAPA CNNTDNN 的架構設計進行擴增。我們的實驗與分析集中在不同的前處理模組以及注意力機制上。首先，我們會將前處理模組替換為不同的結構進行訓練，除了原始的 CNN 結構外，我們另外實驗了預激活的 CNN 結構以及導入兩個維度注意力的 MFA 模組(Liu *et al.,* 2022)。之後我們會取這三組模型中表現較好的模型替換其中使用的注意力機制，將原有的 SE 模組分別替換成 CBAM 模組(Woo *et al.,* 2018) 以及 GC 模組(Cao *et al.,* 2019)。在我們的最終模型中，使用了預激活的 2D CNN 模組作為前處理模組以及 CBAM 模組作為模型的注意力機制，在

Voxceleb 1-O、Voxceleb 1-E、Voxceleb 1-H 及 VoxSRC2022 測試集上都實現了比起基準模型更好的表現。

本文主要分為五個部份，第一部份為緒論；第二部份為研究方法，會介紹使用到的資料前處理方法、模型架構、特徵前處理模組以及注意力機制；第三部份為實驗設置，說明實驗所使用到的資料集、參數設置以及評估準則；第四部份為實驗結果，會比較不同前處理模組與注意力機制的實驗數據，並根據實驗結果進行分析與討論；第五部份為結論。

2. 研究方法 (Research Methods)

在這個章節我們將會詳細的講解本次實驗所使用到的各種方法，包含對輸入音檔進行的處理、主幹模型架構的細節、不同前處理模組以及不同注意力機制的介紹。實驗上我們使用了 VoxSRC 官方所提供的訓練工具(Chung *et al.,* 2020) 進行訓練，並以 IM ECAPA-TDNN 做為基準模型，透過結合不同的前處理模組以及注意力機制觀察這些改動對模型效能所造成的影響。

2.1 資料前處理 (Data Preprocessing)

為了提高模型的強健性以及避免產生過度擬和（overfitting）的狀況，我們利用了資料增強的方法增加訓練資料的多樣性。透過對訓練音檔加入噪音跟迴響，能夠有效的提昇模型的泛化能力，使其在推論階段的表現更加優秀。而在將音檔轉換為特徵向量方面，在參考了近年競賽中各隊伍的作法後，我們選用梅爾頻譜作為主要聲學特徵。

2.1.1 資料增強 (Data Augmentation)

我們使用了兩種用於資料增強的資料集來對我們的訓練資料進行強化。首先是透過 MUSAN 資料集(Snyder *et al.,* 2015) 來為輸入音檔加入噪音，在 MUSAN 資料集中共分成了三個部份，分別為語音（speech）、音樂（music），以及噪音（noise），語音部份的內容全都是來自公共場合中的背景說話聲，包含朗讀書本章節以及美國政府部門聽證會等等，語音部份總共由 12 種語言組成，其中以英語的比例為最多；音樂部份的內容包含了多種不同時期、流派的音樂，比如有傳統流派的巴洛克、浪漫、古典音樂，也有流行流派的爵士、藍調、嘻哈音樂等等；噪音部份的內容則包含了科技性噪音（如撥號音、傳真機噪音等）以及環境聲音（如雷聲、雨聲、動物噪音等），有些檔案也會有包含模糊的人群噪音。另一個則是利用 RIR（Room Impulse Response，空間脈衝響應）資料集(Ko *et al.,* 2017) 將音檔加入迴響（Reverberate），在 RIR 資料集中有真實與模擬的聲音資料，我們只會使用模擬的空間音進行資料增強。

2.1.2 聲紋特徵擷取 (Acoustic Feature Extraction)

我們使用 80 維的梅爾頻譜（Mel-filter bankfeatures，FBank features）作為我們的主要聲學特徵，理由是相較於梅爾倒頻譜係數（Mel-Frequency Cepstral Coefficients，MFCC）來說，梅爾頻譜因為沒有經過 DCT 變換，使得其保留了更多的聲音訊號資訊，能夠在分析語者特徵上取得更好的結果。

2.2 模型架構 (Model Architecture)

在 ECAPA-TDNN 推出之後，得益於優秀的多層聚合策略以及多尺度特徵卷積，該模型在各個語者驗證競賽中都取得優秀的表現，許多人也以其架構作為基底進行不同程度的改良。本篇論文我們以基於 ECAPA-TDNN 改進的 Improving ECAPA-TDNN 作為基底，配合後續實驗進行調整，降低了模型計算量並維持相近之模型表現。我們把修改後的模型命名為 IM ECAPA-TDNN，並將其作為本篇論文中的基準模型。

2.2.1 Improving ECAPA-TDNN

Improving ECAPA-TDNN 是基於 ECAPATDNN 所設計的一個改進版本。在該模型中，Zhang et al.使用了帶有 SE 注意力機制的 SCBlock (Liu *et al.,* 2020) 取代了原始架構主幹網路裡的 Res2Block，通過 SC-Block 所帶有的自校準計算及分割卷積來獲得更大的感受野（receptive field）及上下文的空間注意力，以此避免特徵中不必要的資訊，並在 SC-Block 後面接上 SE-Block，透過注意力機制使有效特徵圖（feature map）權重要大於低效的特徵圖。Zhang et al.還在每一層 SE-SC-Block 之間插入聚合（aggregation）層的結構，用來將不同分辨率的特徵串接整合並降採樣為下一層 SE-SC-Block 的輸入大小。這些聚合層會與原始 ECAPA-TDNN 的多層聚合方法結合，使模型成為一個階層式的聚合結構，也就是每一層 SE-SC-Block 的輸出都會作為之後每一層聚合層的輸入，而越接近模型尾端的聚合層就會融合越多不同分辨率的特徵，以提取更具語者資訊的嵌入向量。

2.2.2 IM ECAPA-TDNN

我們以 Improving ECAPA-TDNN 作為基底進行修改，最主要的改動便是我們減去了一層聚合層結構，與此同時也減去了一層的 SESC-Block，並將第一層 TDNN 結構的輸出也作為後面各聚合層的輸入，修改後的模型如圖 1 所示。我們想要透過聚合層來將保留更多特徵資訊的第一層 TDNN 輸出向量一併與後面每一層的 SE-SC-Block 的輸出向量進行特徵重組，以此來獲取更多的語者特徵訊息；而將 SE-SC-Block 及聚合層各減少一層的主要是考量到實驗彈性，由於首層 TDNN 的輸出會加入到每一層聚合層當中進行特徵重組，若是保留原有的四層結構，在替換前處理模組以及注意力機制的實驗上便會出現硬體限制的情況發生。基於以上原因，我們對原始的 Improving ECAPA-TDNN 進行了修改，並將修改後的模型命名為 IM ECAPA-TDNN。

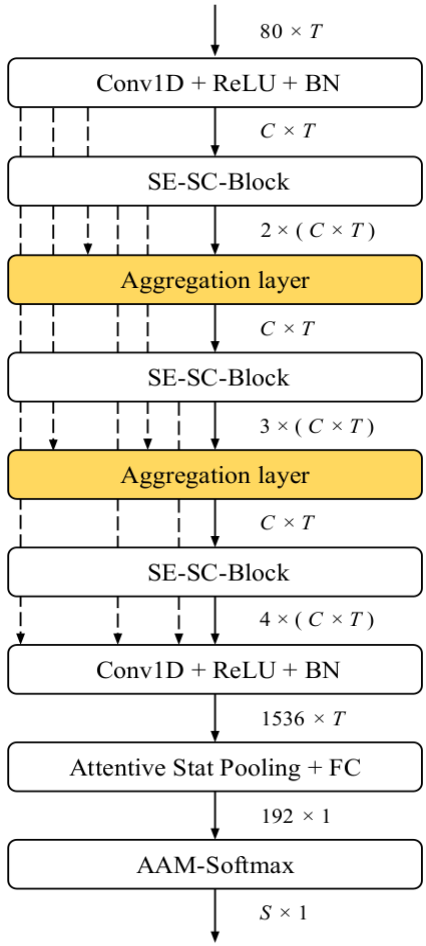


圖 1. 修改提出的 IM ECAPA-TDNN。其中 C 表示通道數，T 表示音框數，S 表示分類與者數量。

[Figure 1. The proposed IM ECAPA-TDNN. C denotes aschannels, T denotes as frames, S denotes as numbersof speaker.]

2.3 特徵前處理模組 (Feature Preprocessing Modules)

在 ECAPA CNN-TDNN 的研究成果中，通過將輸入音檔的特徵頻譜圖先傳入前處理模組中進行特徵重組，再將重組後的特徵圖在通道及頻率維度攤平（flatten），使其作為一般輸入傳入 ECAPA-TDNN 進行訓練能夠有效的提高模型表現，因此我們將這個設計加入基準模型當中。我們在 IM ECAPA-TDNN 前面實作了 3 種不同結構的前處理模組進行實驗，分別為原始論文中的 2D CNN 模組、經過預激活（pre-activation）修改的 2D CNN 模組，以及引入兩維度注意力 MFA 模組。

2.3.1 2D CNN 模組 (2D CNN Module)

為原始在 ECAPA CNN-TDNN 中所使用的前處理模組，通過一般的二維卷積與 ResNet 結構中的 ResBlock 進行組合而成，在實做上我們還有在 ResBlock 中加入 SE 模組，整體結構如圖 2 所示。由於實驗環境以及訓練時間等因素考量，我們將 residual block 的通道數下調為 64 以降低模型大小，同時參照原始模型設定將第一個及最後一個二維卷積的步幅（stride）設置為 2 來增加計算效率。

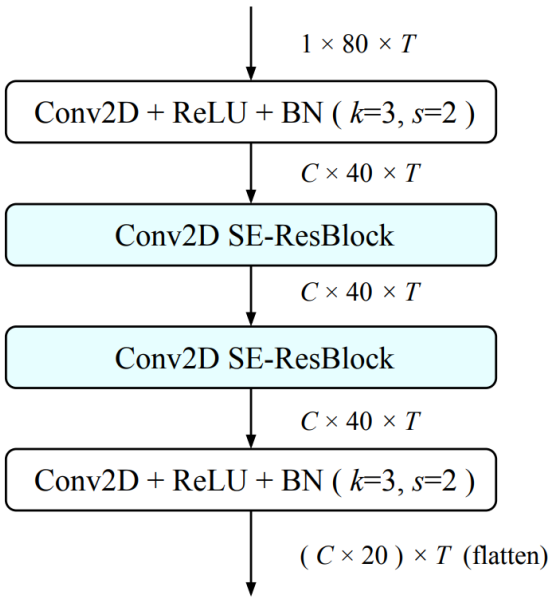
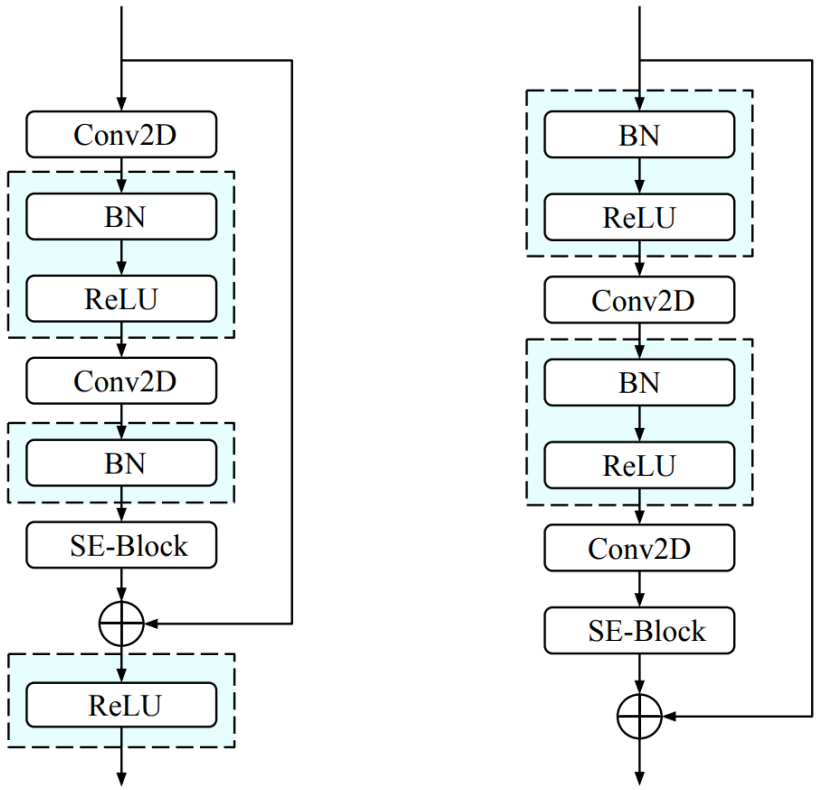


圖 2. 2D CNN 模組。其中 C 表示通道數，T 表示音框數。而卷積中的 k 與 s 表示卷積核大小及步伐長度。

[Figure 2. 2D CNN module. C denotes as channels, T denotes as frames. k and s in convolutions denote kernel size and stride.]

2.3.2 預激活的2D CNN 模組 (Pre-activated 2D CNN Module)

我們參考了(He *et al.,* 2016b) 中對殘差網路的研究結果，在該研究中表明當在 ResBlock 的捷徑連結（shortcut connection）上進行任何操作都會降低模型的表現；同時若是將模型中的激活函數從傳統的後激活（post-activation）改為預激活（pre-activation），能夠使模型更易於訓練，並有效的提高模型的泛化度。基於上述研究結果，我們將 2D CNN 模組中 ResBlock 的結構順序進行調整，新結構與舊結構比較如圖 3 所示。



(a) original (b) pre-activation

圖 3. 原始 SE-ResBlock 與預激活結構之比較。⊕ 表示元素對應相加。
[Figure 3. Comparison between original and pre-activation SE-ResBlock. ⊕ denotes aselement-wise addition.]

2.3.3 MFA 模組 (MFA Module)

MFA 模組是 Liu *et al.*在 MFA-TDNN 中設計用來取代 2D CNN 模組的新結構，其中使用了一個 Res2Block 變體來取代 ResBlock，這個變體是在傳統的 Res2Block 中改進了兩個新結構，也就是雙通道多尺度模組（dualpathway multi-scale module）以頻率及通道注意力模組（frequency-channel attention module），模組結構如圖 4 所示。雙通道多尺度模組的做法是在 Res2Block 中的每個分支卷積後額外再進行一個 TDNN 模組的卷積，並且這個模組的輸出會傳入到另一個分支當中，這就與 Res2Net 原有的卷積輸出形成了雙通道輸入到另一個分支中進行計算。頻率及通道注意力模組則是建構在前面提到的 TDNN 模組當中，結構如圖 5。其整體的概念其實與 SE 模組相似，不同的是特徵向量通過全局平均池化（Global average pooling，GAP）後是會留下頻率以及通道兩個維度的平面向量，接著將此向量攤平進行 SE 模組中激發（excitation）計算，最後再將激發後的向量重塑（reshape）回原來的平面向量並且作為權重值乘回原始的特徵向量。

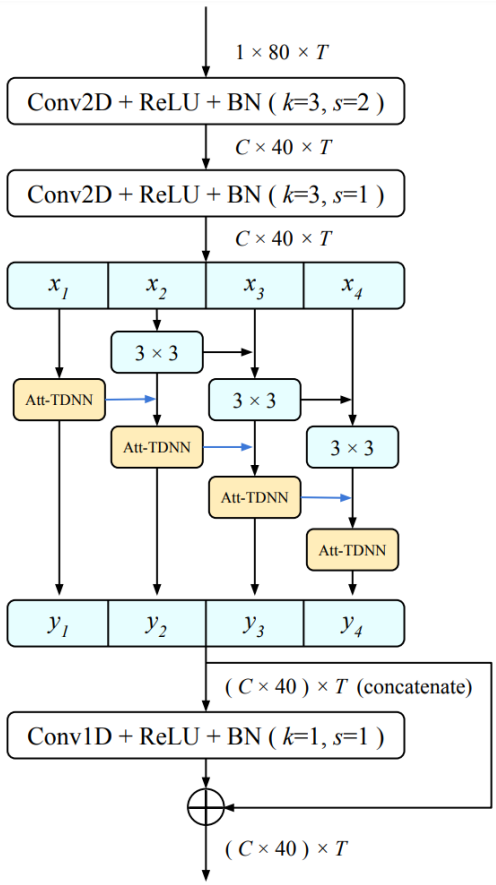


圖 4. MFA 模組。其中 C 表示通道數，T 表示音框數，卷積中的 k 與 s 表示卷積核大小及步伐長度，⊕ 表示元素對應相加。

[Figure 4. MFA module. C denotes as channels, T denotes as frames, k and s in convolutions denote kernel size and stride, ⊕ denotes as element-wise addition.]

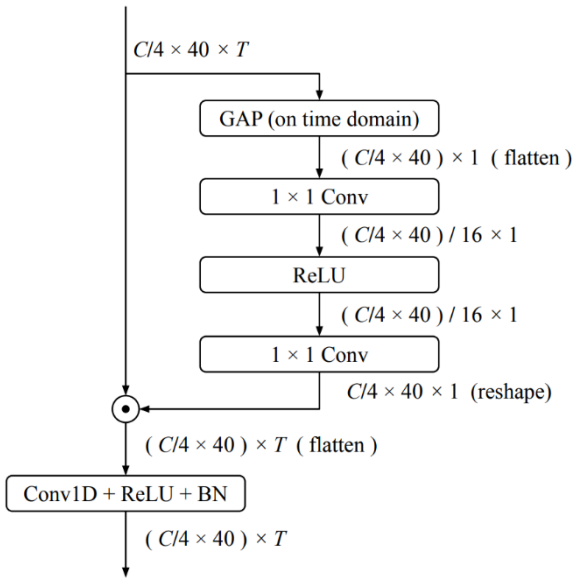


圖 5. MFA 模組中的 Att-TDNN 模組之結構。其中 C 表示通道數，T 表示音框數，⊙ 表示元素對應相乘。

[Figure 5. Att-TDNN module, which inside the MFA module. C denotes as channels, T denotes as frames, ⊙ denotes as element-wise product.]

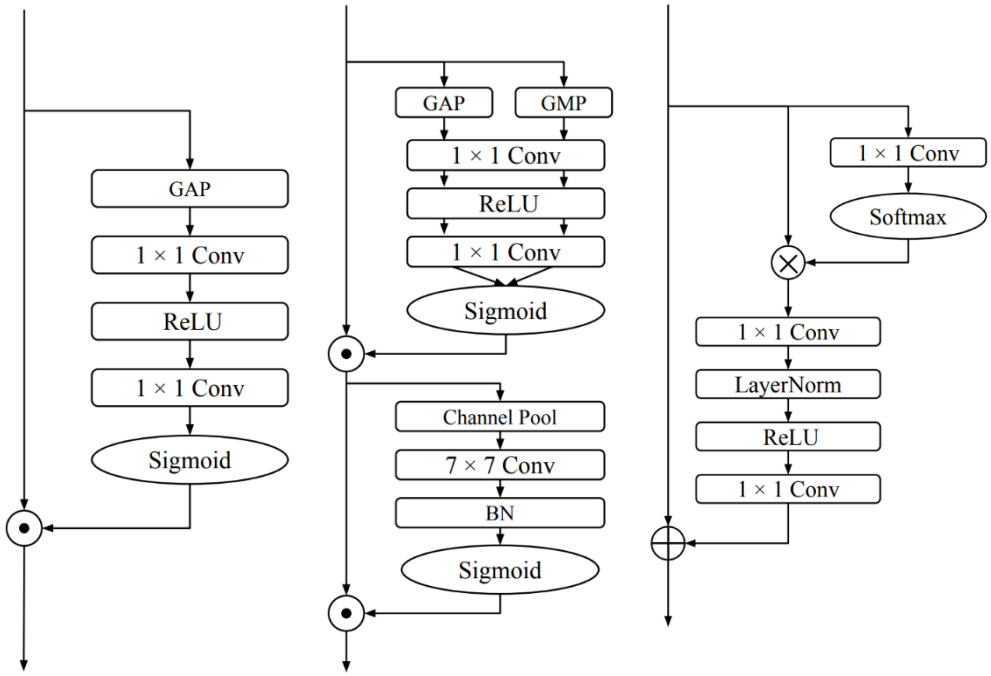
2.4 注意力機制 (Attention Mechanisms)

在原始的 ECAPA-TDNN 及後續的各個改進版本中，不論如何修改、擴增網路結構，其中都會引入注意力機制來提高模型整體的表現。就我們的基準模型以及 2D CNN 模組中使用到的 SE 模組來說，SE 模組會對特徵向量操作後取得特徵向量各通道不同的權重，透過權重，我們可以抑制特徵中不重要的資訊，並有效的將重要的特徵資訊給凸顯出來。而考慮到在 SE 模組問世至今，已有許多後起之秀在各大競賽中脫穎而出，藉由自身獨特的結構設計進一步增強注意力機制在模型上的影響，我們在此替換並比較包含 SE 模組在內，共計 3 種不同結構的注意力機制在本次語者驗證系統上的表現，要替換成的模組分別是 CBAM 模組以及 GC 模組。關於這些注意力模組的詳細結構請見圖 6。而由於 MFA 模組中自身較特殊的注意力設計，我們並不會替換 MFA 模組當中使用的注意力機制。

2.4.1 SE 模組 (SE Module)

SE（Squeeze and Excitation）模組為原始結構中所使用的注意力機制模組，模型結構如圖 6(a) 所示。其透過壓縮（squeeze）與激發（excitation）兩步驟來計算不同通道的權重。首先是壓縮，輸入特徵會對通道以外的維度進行全局平均池化計算，以取得各個通道的記述子（descriptor）；再來是激發，各通道的記述子會輸入兩層卷積層中進行降維升維

的操作，來學習不同通道記述子的重要程度，並透過 sigmoid 函數將其轉換成通道權重乘回原始特徵向量當中。



(a) SE 模組(SE module) (b) CBAM 模組 (CBAM module (c) GC 模組(GC module)

圖 6. 不同注意力機制模組之結構。⊙ 表示元素對應相乘，⊗ 表示矩陣相乘，⊕ 表示元素對應相加。

[Figure 6. Different attention mechanism architectures. ⊙ denotes as element-wise product, ⊗ denotes as matrix multiplication, ⊕ denotes as element-wise addition.]

2.4.2 CBAM 模組 (CBAM Module)

CBAM（Convolutional Block Attention Module）模組是基於 SE 模組的擴展，模型結構如圖 6(b) 所示。其在計算完通道權重之後，會接著計算空間權重以突顯更重要的空間特徵。同時在兩種權重的計算當中除了使用全局平均池化之外，還會使用全局最大池化（Global map pooling，GMP）來取得更多不同的資訊。

2.4.3 GC 模組 (GC Module)

GC（Global Context）模組是將 SE 模組與 Non-local 模組(Wang *et al.*, 2018) 進行結合而成，模型結構如圖 6(c) 所示。鑑於 Non-local 模組優秀的上下文建模（context modeling）能力與 SE 模組輕量的計算結構， Cao et al.通過簡化 Non-local 模組，然後將 Non-local

模組的特徵轉換層修改為類 SE 模組的結構以融合兩模組的優點。透過這樣的設計，GC 模組在各項電腦視覺領域的競賽當中皆有不俗的表現。

3. 實驗設置 (Experiments)

這個章節我們會介紹本論文實驗中所使用到的訓練資料集以及測試資料集，也會詳細描述模型在訓練中所設置的各項超參數，並說明最終用來評估模型表現的準則。

3.1 資料集 (Datasets)

我們使用 VoxCeleb 2 (Chung *et al.,* 2018) 中 dev 的部份作為我們的訓練資料集，並使用以 VoxCeleb 1 (Nagrani *et al.,* 2017) 資料集音檔所組成的 VoxCeleb 1-O/E/H 測試集以及 VoxSRC 2022 的驗證集作為本次模型的測試集。我們並沒有使用語音活性偵測（Voice activity detection，VAD）對實驗音檔進行調整。

3.2 參數設置 (Implementation details)

為了公平比較模型表現，所有模型皆套用了相同的訓練策略進行訓練：使用 Adam 優化器（optimizer）調整神經網路參數，初始學習率為 1e-03，每 10 個 epoch 會減少 25%。使用 AAM-Softmax 作為損失函數，其中 margin 設為 2，scale 設為 30。訓練期間應用權重衰減來防止模型過度擬合，將值設為 2e-05。訓練時的 batch size 設置為 256，並訓練 100 個 epoch 取其中最好的模型參數。主幹網路 IM ECAPA-TDNN 中的通道數量皆設置為 512，語者嵌入的輸出大小設置為 192；在前處理模組方面，2D CNN 模組不論是否為預激活其通道大小都設置為 64，而 MFA 模組基於模型大小則設為 32。

3.3 評估準則 (Evaluation Metrics)

我們以等錯誤率（Equal Error Rate, EER）以及最小檢測成本函數（Minimum Detection Cost Function, MinDCF）作為我們評估系統表現的準則。其中最小檢測成本函數依照 VoxSRC 2022 設定的標準，將參數設置為 C_{miss}=1、C_{fasle}=1、P_{target}=0.05。我們並沒有使用任何分數正規化方法對分數進行調整。

4. 實驗結果 (Experimental Results)

我們首先比對了原始 ECAPA-TDNN 與本次作為基準模型的 IM ECAPA-TDNN 在最簡單的 VoxCeleb1-O 及最困難的 VoxSRC2022 驗證集上的表現，其結果如表 1 所示。可以看到經過修改後的 IM ECAPA-TDNN 雖然在困難資料集上的表現與原始 ECAPA-TDNN 相差無多，但在簡單資料集上明顯是更為優秀的一方。

接著我們會分別討論不同的前處理模組以及不同的注意力機制對模型表現所造成的影響，並將表現最好的組合做為我們的最終模型。所有模型在各個測試集上的詳細結果如表 2 所示。

表 1. IM ECAPA-TDNN 與 ECAPA-TDNN 在最簡單及最困難的資料集上之表現比較

[Table 1. Comparisonthe performance beteen IM ECAPA-TDNNand ECAPA-TDNN on the easiest and the hardiesttest sets.]

Architecture	VoxCeleb1-O		VoxSRC2022 val	
	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN (Re-implemented)	1.3770	0.0931	**3.6735**	0.2479
IM ECAPA-TDNN	**1.2600**	**0.0849**	3.6824	**0.2462**

表 2. 不同模型在各測試集上的表現比較

[Table 2. Comparison the performance between different models on each test sets.]

Architecture	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H		VoxSRC2022 val	
	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
IM ECAPA-TDNN (baseline)	1.2600	0.0849	1.4733	0.0941	2.6891	0.1621	3.6824	0.2462
不同的前處理模組								
IM ECAPA CNN-TDNN	1.1218	0.0886	1.2763	0.0825	**2.3318**	0.1475	**3.2230**	0.2144
IM ECAPA CNN-TDNN (pre-act)	**1.0424**	**0.0739**	1.2646	0.0831	2.3518	**0.1415**	3.4471	0.2198
IM ECAPA MFA-TDNN	**1.0424**	0.0797	**1.2632**	**0.0813**	2.3526	0.1439	3.2535	**0.2118**
不同的注意力機制								
IM ECAPA CNN-TDNN (pre-act) with SE	**1.0424**	**0.0739**	1.2646	0.0831	2.3518	**0.1415**	3.4471	0.2198
IM ECAPA CNN-TDNN (pre-act) with CBAM	1.1484	0.0817	**1.2507**	**0.0821**	**2.3500**	0.1437	**3.1160**	**0.2053**
IM ECAPA CNN-TDNN (pre-act) with GC	1.2552	0.0992	1.3807	0.0926	2.5533	0.1551	3.4990	0.2282

4.1 前處理模組的比較(Comparison between Feature Prepocessing Module)

在加入了前處理模組之後，所有的模型相較於基準模型都有顯著的進步。相比於 2D CNN 模組在各個資料集上都有穩定的發揮，預激活的 2D CNN 模組雖然在相對簡單的 Voxceleb1-O 測試集上明顯優於原始的 2D CNN 模組，但是其在複雜度越高的測試集上表現卻較為差勁，我們認為主要是由於我們使用了輕量的 IM ECAPA-TDNN 作為主幹

網路，而在(He *et al.,* 2016b) 中表明了預激活的 ResBlock 要在深層的網路結構中才能發揮效果，所以才造成預激活 2D CNN 模組在複雜測試集上表現不佳的原因。而 MFA 模組得益於其多尺度多維度注意力的卷積結構，其在簡單的測試集上可以做到與使用預激活 2D CNN 模組一樣優異的表現，並在複雜的測試集上表現相對穩定。

4.2 注意力機制的比較 (Comparison between Attention Mechanisms)

考慮到 MFA 模組本身自帶的注意力機制無法輕易變動，我們在 2D CNN 模組中選擇了預激活的版本替換其注意力模組，來觀察各注意力機制對模型表現造成的影響。SE 模組在相對簡單的 Voxceleb1-O 測試集上依舊有著較佳的表現，但是 CBAM 模組在其他更為複雜資料集對比另外兩個注意力模組都有著更優秀的結果。會有這樣的差異我們認為是因為 CBAM 模型引入空間注意力能夠有效的將更多重要的語者特徵突顯出來，且相比 SE 模組只做了全局平均池化，CBAM 還加入了全局最大池化進行計算以取得不同方面的資訊，這些設計讓模型能夠在複雜的測試集上擷取更細微的特徵進行辨識，進而提高了辨識結果的表現；對比 CBAM 的優異表現，GC 模組反而在所有測試集的表現都不突出，會有這樣的問題我們認為是模組的設計與 TDNN 結構衝突，將模組結構硬是改寫為相容 TDNN 反而造成擷取特徵時產生冗餘的資訊，導致 GC 模組連 SE 模組的表現都達不到。

4.3 最終提出模型 (Final Proposed Model)

根據我們上述的實驗結果，我們將帶有預激活 2D CNN 前處理模組，並替換注意力機制為 CBAM 的 IM ECAPA-TDNN，即表 2 中的 IM ECAPA CNN-TDNN (pre-act) with CBAM 做為我們的最終提出模型。相比與基準模型，我們的最終模型在各測試集上都有明顯的進步，以最複雜的 VoxSRC2022 驗證集來說，最終模型在 EER 值與 minDCF 值上分別有 15.4% 以及 16.6% 的進步幅度。

5. 結論 (Conclusions)

本論文提出了基於 Improving ECAPA-TDNN 修改的 IM ECAPA-TDNN 結構作為我們的基準模型，並透過結合不同的前處理模組以及調整注意力機制來對模型表現進行進一步的強化。我們提出的最終模型通過結合預激活的 2D CNN 前處理模組與替換注意力機制為 CBAM 模組，在各項測試集上的表現對比基準模型都有著大幅提昇。未來我們將會以此為依據來修改其他更加複雜的主幹網路，希望能夠藉此來進一步的提昇我們語者驗證系統的效能。

參考文獻 (References)

Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of 2019 IEEE/CVF International*

Conference on Computer Vision Workshop (ICCVW). https://doi.org/10.1109/ICCVW.2019.00246

Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., Han, I. (2020) In Defence of Metric Learning for Speaker Recognition. In *Proc. Interspeech 2020*, 2977-2981, https://doi.org/10.21437/Interspeech.2020-1064

Chung, J.S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. CoRR, abs/1806.05622.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, *43*(2), 652-662. https://doi.org/10.1109/TPAMI.2019.2938758

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. In *Proceedings of ECCV 2016*, 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

Hu, J., Shen, Li, & Sun, G. (2018). Squeeze-andexcitation networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5220-5224. https://doi.org/10.1109/ICASSP.2017.7953152

Liu, J.-J., Hou, Q., Cheng, M.-M., Wang, C., & Feng, J. (2020). Improving convolutional networks with self-calibrated convolutions. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10093-10102. https://doi.org/10.1109/CVPR42600.2020.01011

Liu, T., Das, R. K., Lee, K. A., & Li, H. (2022). Mfa: Tdnn with multi-scale frequency-channel attention for textindependent speaker verification with short utterances. In *Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7517-7521. https://doi.org/10.1109/ICASSP43922.2022.9747021

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proceedings of Interspeech 2017*, 2616-2620.

Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484

Thienpondt, J., Desplanques, B., & Demuynck, K. (2021). Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification. arXiv preprint arXiv:2104.02370

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794-7803.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I.S. (2018). Cbam: Convolutional block attention module. arXiv preprint arXiv:1807.06521

Zhang, Y.-J., Wang, Y.-W., Chen, C.-P., Lu, C.-L., & Chan, B.-C. (2021). Improving Time Delay Neural Network Based Speaker Recognition with Convolutional Block and Feature Aggregation Methods. In *Proc. Interspeech 2021*, 76-80. https://doi.org10.21437/Interspeech.2021-356

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.