Computational Linguistics and Chinese Language Processing
Vol. 25, No.2, December 2020, pp. 55-68
© The Association for Computational Linguistics and Chinese Language Processing

NSYSU+CHT 團隊於 2020 遠場語者驗證比賽之語者驗證系統

NSYSU+CHT Speaker Verification System for Far-Field Speaker Verification Challenge 2020

張育嘉*、陳嘉平*、蕭善文+、詹博丞+、呂仲理+

Yu-Jia Zhang, Chia-Ping Chen, Shan-Wen Hsiao,

Bo-Cheng Chan, and Chung-li Lu

摘要

在本論文中，我們描述了 NSYSU+CHT 團隊在 2020 遠場語者驗證比賽 (2020 Far-field Speaker Verification Challenge, FFSVC 2020) 中所實作的系統。單一系統採用基於嵌入的語者識別系統。該系統的前端特徵提取器是結合了時延神經網路，與卷積神經網路模組兩者的優點，稱為時延殘差神經網路的架構。在池化層，我們實驗了不同方式：統計池化層和 GhostVLAD。而後端的評分器則採用機率線性判別分析，我們訓練跟調適機率線性判別分析用以不同系統的融合。我們分別參加了 FFSVC 2020 採單一麥克風陣列資料的文本相關（任務一）與文本無關（任務二）的語者驗證任務。我們提出的系統在任務一上取得 minDCF 0.7703，EER 9.94%，在任務二上則是 minDCF 0.8762，EER 10.31%。

Abstract

In this paper, we describe the system Team NSYSU+CHT has implemented for the 2020 Far-field Speaker Verification Challenge (FFSVC 2020). The single systems

*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: M083040025@student.nsysu.edu.tw; cpchen@mail.cse.nsysu.edu.tw
+中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
E-mail: {whsiao; cbc; chungli@cht.com.tw

are embedding-based neural speaker recognition systems. The front-end feature extractor is a neural network architecture based on TDNN and CNN modules, called TDResNet, which combines the advantages of both TDNN and CNN. In the pooling layer, we experimented with different methods such as statistics pooling and GhostVLAD. The back-end is a PLDA scorer. Here we evaluate PLDA training/adaptation and use it for system fusion. We participate in the text-dependent(Task 1) and text-independent(Task 2) speaker verification tasks on single microphone array data of FFSVC 2020. The best performance we have achieved with the proposed methods are minDCF 0.7703, EER 9.94% on Task 1, and minDCF 0.8762, EER 10.31% on Task 2.

關鍵詞：遠場語者驗證、時延神經網路、卷積神經網路、時延殘差神經網路、GhostVLAD

Keywords：Speaker Verification, TDNN, CNN, TDResNet, GhostVLAD

1. 緒論 (Introduction)

自動語者驗證(Automatic Speaker Verification, ASV)系統隨著深度學習技術的發展，有著顯著的提昇，每一年舉行的相關競賽更是不勝枚舉，不管是 NIST Speaker Recognition Evalution (SRE) (NIST, 2019)，抑或是防止欺騙語音攻擊的 ASVspoof (Todisco *et al.*, 2019)，這些競賽都促使自動語者驗證系統日趨成熟。目前最被廣泛採用的自動語者驗證系統是基於嵌入(Embedding)的架構，該架構由前端的特徵提取器(Feature Extractor)，以及後端的評分器(Scorer)組合而成。前端在音框層(Frame level layer)將原始輸入提取成高階的表徵，並經由池化層整合音框層資訊成為音段層(Segment level layer)，而後透過全連接層提取嵌入並 softmax 計算機率以分類語者。前端架構從傳統的深度神經網路(Deep Neural Network, DNN) / i 向量(i-vector) (McLaren, Lei, & Ferrer, 2015)，一直到近年在語者驗證比賽中，大放異彩的時延神經網路(Time Delay Neural Network, TDNN) / x 向量 (x-vector) (Snyder, Garcia-Romero, Sell, Povey & Khudanpur, 2018) 與以其為延伸的架構：擴展時延神經網路 (Extend-TDNN) (Snyder *et al.*, 2019)；而原先基於影像辨識所建立的卷積神經網路，被認為在語者辨識任務中也能取得不錯的表現，像是殘差神經網路(Residual Neural Network, ResNet) (Nagrani, Chung, Xie & Zisserman, 2020; Xie, Nagrani, Chung & Zisserman, 2019; Qin, Bu & Li, 2019)；另一方面，也有許多研究是針對池化層與損失函數來作改進，池化層除了在時延神經網路中最常使用的統計池化層(Statistic Pooling)之外，自注意池化層(Self-attentive Pooling) (Okabe, Koshinaka & Shinoda, 2018; Zhu, Ko, Snyder, Mak & Povey, 2018)，NetVLAD (Chen *et al.*, 2018)都能使系統在整合音框層資訊的效果更好；損失函數則是從人臉辨識領域借鑒而來，嘗試了許多 softmax 不同的變體：L-softmax (W. Liu, Wen, Yu & Yang, 2016)、A-softmax (W. Liu *et al.*, 2017)、AM-softmax (Wang, Cheng, Liu & Liu, 2018)、AAM-softmax (Deng, Guo, Xue & Zafeiriou, 2019)。而後端的評分器，除了餘弦相似度(Cosine Similarity)，機率線性判別分析(Probabilistic Linear

Discriminant Analysis, PLDA) (Kenny, 2010)成為了最常使用的方法之一。

近年來隨著物聯網設備與智慧家居產品的普及，短語音指令的處理，以及在遠場噪音的真實使用場景下，成為了自動語者驗證系統的新挑戰，而錄音設備的不匹配，更再加深了識別的難度，為了推動該情景下的自動語者驗證系統研究，FFSVC 2020 (Qin *et al.*, 2020)因應而生。

因此，本論文旨在參加 FFSVC 2020，並針對任務一：單一麥克風陣列的遠場文本相關語者驗證(Far-Field Text-Dependent Speaker Verification from single microphone array)與任務二：單一麥克風陣列的遠場文本無關語者驗證(Far-Field Text-Independent Speaker Verification from single microphone array)採取不同的解決方法。我們以基於時延神經網路的擴展時延神經網路，與基於卷積神經網路的殘差神經網路，建立了前端的特徵提取器。聲學特徵採用 FBank (Filter Bank)配上音調(Pitch)。而後也針對殘差神經網路架構進行修改，將其與擴展時延神經網路結合，成為一個新的網路架構稱為時延殘差神經網路(Time Delay Residual Neural Network, TDResNet)，並使用機率線性判別分析作為後端評分器，分別實驗上述模型架構在各任務上的表現。此外，我們也針對池化層做改變，將原先的統計池化層替換成 NetVLAD 的改進：GhostVLAD (Zhong, Arandjelovic & Zisserman, 2018)。更多的實作細節將會在後續的章節詳細說明。

2. 網路架構 (Network Architecture)

2.1 音框層 (Frame Level Layers)

在此章節，我們總共實作了三種不同的架構，一種是擴展時延神經網路，另一種是殘差神經網路，而最後一種則是我們發現前兩者在網路架構上有互補之處，因此我們將其結合，成為一個全新的架構，稱為時延殘差神經網路。

2.1.1 基於時延神經網路 (TDNN-based)

我們參照(Snyder *et al.*, 2019)建立了擴展時延神經網路作為我們的基準(Baseline)，其架構使用了十層來提取音框層的特徵，並且在第 3、5、7 層中使用到了擴張(dilation)的概念，這也是時延神經網路的精髓所在，以擴張來擴大音框層資訊的感知範圍，擴張數分別為 2、3、4，當擴張數為 2 時，我們會取相鄰的 5 個音框進行運算，擴張數為 3 時，則取 7 個音框，以此類推，透過層層堆疊，因此最終可以感知 23 個音框的資訊。接著將音框層輸出經過統計池化層，將音框層資訊整合成音段層資訊，而後由兩層全連接層組成音段層，最後輸出經 softmax 計算機率進行分類；在推論階段，我們從音段層的第一層取出 512 維代表語者的嵌入。每一層皆經過批量標準化(Batch Normalization)與 Rectified Linear Unit (ReLU)激活函數。

2.1.2 基於卷積神經網路 (CNN-based)

根據 (Nagrani *et al.*, 2020; Xie *et al.*, 2019; Qin, Bu & Li, 2019) 的研究，都表明殘差神經網路架構，在有噪音與迴響 (reverberate) 的遠場環境下，對於特徵的擷取是相當出色的，因此採用殘差神經網路作為我們基於卷積神經網路的音框層架構，並參考 (Nagrani *et al.*, 2020) 所提到的 thin-ResNet 架構，實作了參數量較少的殘差神經網路，接著同樣將音框層的輸出經過統計池化層整合，而後的音段層與擴展時延神經網路不同，只採用一層全連接層，並將 softmax 替換成 AM-softmax 計算機率進行分類。每一層皆經過批量標準化與 ReLU 激活函數。每一個殘差區塊 (Residual Block) 皆使用殘差連結 (Residual connect) 連接，最終架構圖如表 1。

表 1. 殘差神經網路架構
[Table 1. Network architecture of ResNet]

#	Module	Structure	Size
0	-	Input 43 Fbank-pitch($43 \times T$)	43
1	Conv	Conv1d, $1 \times 43, 64$	64
2	Conv	$\begin{bmatrix} 1,48 \\ 3,48 \\ 1,96 \end{bmatrix} \times 2$	96
3	Conv	$\begin{bmatrix} 1,64 \\ 3,64 \\ 1,128 \end{bmatrix} \times 3$	128
4	Conv	$\begin{bmatrix} 1,128 \\ 3,128 \\ 1,256 \end{bmatrix} \times 3$	256
5	Conv	$\begin{bmatrix} 1,256 \\ 3,256 \\ 1,512 \end{bmatrix} \times 3$	512
6	Statistic Pooling	Full-seq	2×512
7	Segment	FC	512
8	AM-Softmax		# of speakers

2.1.3 結合時延神經網路與卷積神經網路 (Combination of TDNN and CNN)

擴展時延神經網路與殘差神經網路的差別在於，擴展時延神經網路的前幾層是由擁有擴張的卷積層所組合而成，透過擴張來獲取較大範圍的音框層資訊，隨後緊接著數層無擴張且相同卷積核大小的卷積層來提取並整合音框層資訊，而殘差神經網路的層數雖然較擴展時延神經網路深，但其最終感知的音框層範圍卻不如擴展時延神經網路來得廣闊，所以我們認為兩者有互補之處，擴展時延神經網路後幾層無擴張的部份，如果採用殘差神經網路的機制，就能讓後續的層數越深越大，使得特徵萃取能力更好，因此按照該想法，設計了一個混合的架構，架構如表 2，前五層採用了原先擴展時延神經網路的擴張設計，擴張數分別為 2、3、4、5，後幾層則使用了不同通道大小的殘差區塊各 3 個，該架構保留了擴展時延神經網路獲取較大範圍音框層資訊的能力，同時也擁有殘差神經網

路良好萃取特徵的能力。

表 2. 時延殘差神經網路架構
[Table 2. Network architecture of TDResNet]

#	Module	Structure	Size
0	-	Input 43 Fbank-pitch($43 \times T$)	43
1	TDNN	$[t-2, t+2]$	512
2	TDNN	$\{t-2, t, t+2\}$	512
3	TDNN	$\{t-3, t, t+3\}$	512
4	TDNN	$\{t-4, t, t+4\}$	512
5	TDNN	$\{t-5, t, t+5\}$	512
6	ResNet	$\begin{bmatrix} 1,512 \\ 3,512 \\ 1,1024 \end{bmatrix} \times 3$	1024
7	ResNet	$\begin{bmatrix} 1,1024 \\ 3,1024 \\ 1,2048 \end{bmatrix} \times 3$	2048
8	Statistic Pooling	Full-seq	2×2048
9	Segment	FC	512
10	AM-Softmax		# of speakers

而在 (Li *et al.*, 2018) 論文中，作者提出了與我們想法相近的時延殘差區塊(Time Delay Residual Block, TDResBlock)架構，但他們選擇將時延神經網路模組加入到殘差區塊中。這樣的不同之處在於，我們透過前幾層的擴張得到了固定的感知範圍，才接著使用殘差區塊來萃取與整合特徵，但他們的作法則是讓每個殘差區塊的感知範圍皆不相同，因此每個殘差區塊整合著不同感知範圍的資訊，同時，他們的擴張數最終可達 11，但我們資料的音框數並不足以應付這麼大的擴張，從而導致最終結果可能會受到零填充的影響而變差，因此我們在設計時延殘差神經網路架構時，擴張數只到 5。

2.2 池化層 (Pooling)

除了原先透過計算平均值跟標準差的統計池化層之外，為了解決遠場噪音對於分類的影響，我們嘗試使用 GhostVLAD (Xie *et al.*, 2019)方法，該方法是由 NetVLAD 為基礎改進而來的，NetVLAD 是一種可訓練的分群法，主要做法是將每個音框層的特徵分配到不同的群，接著計算該特徵到群中心的殘差並編碼成最後的輸出，產生 $K \times D$ 大小的矩陣 V。以下為 NetVLAD 計算公式：

$$V(k,j) = \sum_{t=1}^{T} \frac{e^{a_k x_t + b_k}}{\sum_{k'=1}^{K} e^{a_{k'} x_t + b_{k'}}} (x_t(j) - c_k(j)) \quad (1)$$

其中 K 表示群總數，是一個自訂的超參數，D 表示每一個群的維度，與音框層的輸出

通道數相同，a_k， b_k， c_k 是由網路訓練得到的參數，該公式的前半部為 softmax，表輸入 x_t 屬於群 k 的機率，後半部為計算 x_t 與群中心的距離，並以前半部計算出來的 softmax 值作為該距離的權重，再將所有結果相加，最終把所有的群串連成最後的輸出向量 V，而 GhostVLAD 的改進在於向後傳遞時，有些群並不會被包含在最後的輸出當中，如此一來，再訓練網路時，能讓網路自主學習哪些特徵作用較低，應該被分類到需要被排除的群中，而因為被排除的群不會參與到整個網路權重的更新，因此在訓練中似有非有，所以又被稱為 Ghost 群，這也是這個方法的由來，而 Ghost 群也是事先設定好的超參數，它將原先的 K 個群額外增加 G 個 Ghost 群，最後再將 $(K+G)\times D$ 的輸出，只採用 $K\times D$，將代表噪音的 Ghost 群排除掉。我們按照原始論文中的設定，$K=8$，$G=2$。

2.3 損失函數 (Loss Function)

近年來，基於 AM-Softmax 損失函數訓練的語者驗證系統，比起傳統的 softmax 效果有著很大的提昇(Y. Liu, He & Liu, 2019)，因此比起原先的 softmax，我們更偏向採用 AM-softmax，該損失函數將角度間隔的概念引入 softmax 。AM-softmax 損失函數公式如下：

$$L=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s\cdot(cos\theta_j)}} \tag{2}$$

$cos\theta_{y_i}$ 代表第 i 個輸入的特徵向量與權重向量的角度餘弦值，m 則代表角度邊界，s是尺度係數，用於調整角度餘弦值的大小，m 和 s 皆是超參數，這個損失函數的目標是要最大化$cos\theta_{y_i}-m$ 來讓特徵向量與權重向量的夾角最小。我們參考原論文設定$s=30$，$m=0.2$。

2.4 後端評分 (Back-end Scoring)

2.4.1 高斯機率線性判別分析 (Gaussian PLDA)

後端評分器是基於高斯機率線性判別分析，我們先針對擷取出來的語者嵌入作平均正規化，來降低語者嵌入數值的變異性，接著經由線性判別分析 (Linear discriminant analysis, LDA) 來將嵌入的維度降維到 250 維，並用降維過後的嵌入訓練機率線性判別分析，以及用於機率線性判別分析調適 (Adaptation) 的調整，最後以訓練好的機率線性判別分析模型，計算經轉換過後的語者嵌入間的分數。

2.4.2 分數融合 (Score Fusion)

每個系統都有其不同的嵌入提取器的架構，以及機率線性判別分析和機率線性判別分析調適的評分器，而為了能得到最佳的系統表現，我們結合了多個系統所計算出來的分數，

結合方式如圖 1，我們依照機率線性判別分析與機率線性判別分析調適評分器，將一個模型拆分成兩個子系統，並使用 BOSARIS toolkit (Brummer & De Villiers, 2013) 來校正我們系統分數之間的權重，校正資料集採用 FFSVC 2020 開發集。

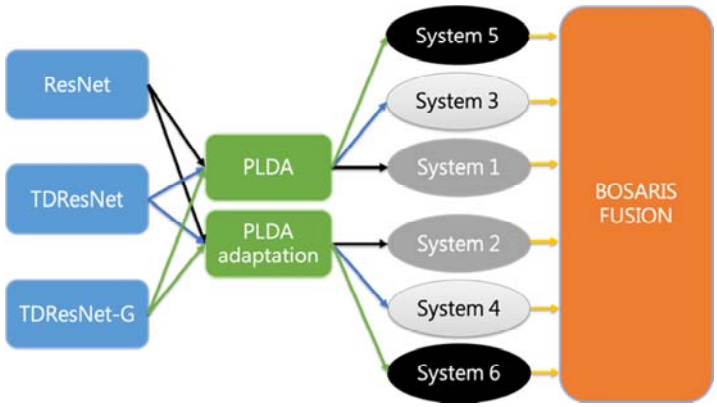


圖 1. 融合策略的示意圖
[Figure 1. Schematic illustration of fusion strategy]

2.5 模型微調 (Model Fine-tuning)

我們採用大量文本無關資料來訓練我們的模型，用以吻合任務二的條件，但如果直接套用在任務一的情境下，所得到的效果會很差，而最直接的解決方法是使用文本相關資料重新訓練嵌入提取器，或者以文本無關資料訓練好的模型作為預訓練(Pre-train)模型，採遷移學習(Transfer Learning)的方式，使用文本相關資料調適模型，但這兩種方法勢必得花費大量的時間，因此我們的作法是選擇將訓練機率線性判別分析與機率線性判別分析調適的資料更換成文本相關資料，能夠節省時間且能達到一定的效果。

3. 實驗設置 (Experimental Setup)

3.1 訓練資料 (Training Data)

競賽方提供的 FFSVC 2020 訓練集，採用麥克風陣列及手機，在不同空間距離、不同雜訊以及不同語速下錄製而成，錄音內容與智慧家居產品的使用情境有關，而除了使用該資料之外，依據競賽要求，我們也從 OpenSLR (OpenSLR, 2020)上挑選開放資料集，其中包含競賽評估計畫中有提及的 SLR-85 (HI-MIA)，而為了要符合本競賽的測試情境，我們也選擇了中文且錄音內容與智慧家居有關的資料集，分別是 SLR-33 (AISHELL)，SLR-38 (FreeST Chinese)，以及 SLR-68 (MAGICDATA)，同時為了增加訓練資料的語者多樣性，我們也加入了在語音任務上經常使用 SLR-49 (VoxCeleb)與 SLR-12 (LibriSpeech)資料集，因此最後總共採用了 7 個不同的資料集用以訓練模型。至於訓練的超參數，我們設定批量大小(Batch Size)為 32，起始學習率為 0.001，並隨著訓練迭代數遞減至 0.0001，模型使用 Nvidia GeForce GTX 1080 Ti GPU 訓練 6 個 epoch。

3.2 資料增強 (Data Augmentation)

訓練資料採用資料增強，一直以來都是被使用於增強語者嵌入模型的強健性(Robustness)，而該篇論文(Qin, Cai & Li, 2019)中有提及，在遠場的環境下，訓練資料與測試資料存在著不匹配的現象，因此為了要模擬遠場的環境，我們針對幾個較大的資料集，使用KALDI toolkit (Povey *et al.*, 2011)以迴響的方式增強我們的訓練資料，最終採用經增強過後的資料來訓練模型，表 3 為我們訓練過程的資料數量與使用方式。

表 3. 訓練過程的資料數量與使用方式
[Table 3. Data usage in the training process]

資料集	語者數	音檔數	語言	資料增強	嵌入提取器訓練	PLDA/PLDA Adaptation
FFSVC 2020 訓練集	120	1,403,383	中	√	√	√
HI-MIA	296	1,157,723	中		√	√
AISHELL	2,331	1,129,626	中	√	√	
FreeST	443	102,600	中		√	
MAGICDATA	1,080	609,550	中		√	
VoxCeleb	7,363	1,281,762	英	√	√	
LibriSpeech	5,831	292,367	英	√	√	

3.3 聲學特徵 (Acoustic Feature)

我們的聲學特徵，採用 KALDI 40 維的 FBank 配 3 維的音調，並且統一取樣頻率為16kHz，音框長度為 25-ms，音框偏移為 10-ms，而特徵擷取完後，使用基於能量的語音活性偵測(Energy-based Voice Activation Detection)來除去沒有聲音的語音片段，許多實驗表明，有無採用基於能量的語音活性偵測對於結果的影響是很大的，接著針對特徵做倒頻譜平均值與變異數正規化(Cepstral Mean And Variance Normalization, CMVN)，降低離群特徵的影響，使模型的訓練效能提昇。

3.4 開發集與驗證集 (Development and Evaluation Data)

使用競賽方所提供的 FFSVC 2020 開發集與驗證集，錄製方式與內容同 FFSVC 2020 訓練集，但彼此語者不重疊。依照競賽要求，測試方式需以手機錄製的音檔作為註冊，麥克風陣列錄製的音檔作為測試。所有實驗皆經由開發集來測試結果，並以其結果來預估在驗證集上的表現，因此開發集並無參與到任何形式的訓練中，僅用來評估模型訓練結果的好壞。

4. 結果 (Result)

我們總共實驗了五種不同的模型在開發集上的表現，逐一對應表 4 的五個系統，分別是擴展時延神經網路、殘差神經網路、時延殘差神經網路、時延殘差神經網路-G（採用GhostVLAD 的時延殘差神經網路）和融合模型，並且僅有時延殘差神經網路、時延殘差

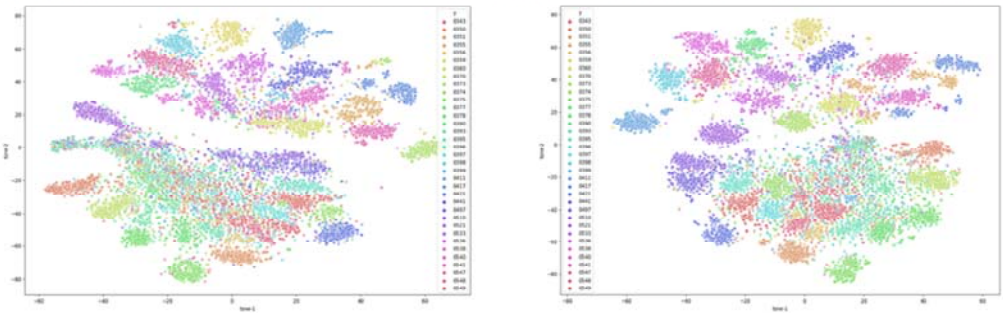
神經網路-G 和融合模型在驗證集上測試並上傳成績，並以有無經過分數融合來區分上傳的系統，經過分數融合的融合模型作為我們競賽的 Primary System 1，而這也是我們在該競賽的最佳系統，另外沒有經過分數融合的時延殘差神經網路與時延殘差神經網路-G 則作為 Single System 1、2， 其中又以 Single System 2 表現較佳。所有結果如表 4 所示。

表 4. FFSVC 2020 開發集與驗證集的最小 DCF 和 EER
[Table 4. Minimum DCF and EER of the FFSVC 2020 development data and evaluation data]

		Development Set								Evaluation Set			
		Task 1				Task 2				Task 1		Task 2	
		PLDA		PLDA Adapt		PLDA		PLDA Adapt					
ID	Model	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER	minDCF	EER
1	E-TDNN	0.9286	11.94%	0.9231	11.82%	0.9503	12.86%	0.9479	12.43%	-	-	-	-
2	ResNet	0.8755	11.21%	0.8851	10.41%	0.9131	12.47%	0.9191	12.02%	-	-	-	-
3	TDResNet	0.8694	11.02%	0.8740	10.23%	0.9220	11.6%	0.93	10.88%	0.8566	11.45%	0.9132	11.36%
4	TDResNet-G	0.8374	11.59%	0.8295	10.33%	0.8650	12.11%	0.87	11.39%	0.8197	12.19%	0.8994	12.11%
5	Fusion(2+3+4)	-	-	-	-	-	-	-	-	0.7703	9.94%	0.8762	10.31%

4.1 單一系統 (Single System)

Single System 1 的架構如時延殘差神經網路章節中所描述，由 5 層的時延神經網路與 6 個殘差區塊組合而成，池化層採用統計池化層，損失函數是 AM-softmax，嵌入提取器的訓練資料使用訓練資料章節提到的 7 種不同資料集。而因為任務一與任務二的差別在於，任務一為文本相關，註冊與測試音檔的內容皆為“你好，米雅”。任務二則為文本無關，內容與智慧家居設備指令與日常用語，因此機率線性判別分析與機率線性判別分析調適依照不同任務使用不同資料，在任務一我們使用 FFSVC 2020 訓練集(只使用編號 1-30) 加上 SLR-85，其錄音內容皆是“你好，米雅”，這樣做的目的是為了與文本相關的任務一測試情境相同，而文本無關的任務二則使用全部的 FFSVC 2020 訓練集。



(a) TDResNet ***(b) TDResNet-G***

圖 2. FFSVC2020 開發集在不同模型所擷取出來的嵌入經 t-SNE 視覺化
[Figure 2. The t-SNE visualization of the embeddings extracted from the different model embedding layer on FFSVC 2020 development data]

Single System 2 則是採取與 Single System 1 相同的模型架構，唯一的不同在於池化層從原來的統計池化層替換成GhostVLAD，訓練資料與超參數並無做任何的更動。此外，我們也使用 t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008) 來分別對 Single System 1、2 的高維度嵌入視覺化，以此評估不同池化層對於最終嵌入學習的影響，結果展示於圖 2，我們可以發現，採用 GhostVLAD 所擷取出來的嵌入經 t-SNE，分群表現較統計池化層佳，尤其是在圖片上半部鮮有重疊者。而從 minDCF 的評估標準來看，分群結果較佳的 Single System 2 也確實表現較佳，這也就表示在 false alarm 與 miss 相同權重的條件下，Single System 2 的驗證效果比 Single System 1 好。

4.2 主系統 (Primary System)

使用經 BOSARIS toolkit 融合過後的系統作為 Primary System 1，選用 ID 2、3、4 作為前端的模型，並且每一個模型分別對應後端的機率線性判別分析和機率線性判別分析調適評分器，因此最終的融合結果由 6 個不同的子系統參與融合後產生，而這個融合系統是我們所有系統中最佳的，於任務一上 EER 9.94%，minDCF 0.7703，在 22 隊參賽隊伍中排名 14 名；於任務二上 EER 10.31%，minDCF 0.8762，在 19 隊參賽隊伍中排名 11 名。

4.3 開發集分析 (Development Data Analysis)

針對 FFSVC 2020 開發集，我們以任務一的時延殘差神經網路測試了空間、雜訊與語速對於結果的影響，為了公平性，我們直接採用競賽方提供的開發集 trials 共 53,996 筆進行測試，並依據我們的測試情形，從 53,996 筆測試中挑選指定的配對，因此每個測試情形的 trials 數量會有些微的差異。

首先，我們實驗註冊與測試在不同空間距離下對於結果的影響，如表 5 所示，0.25m、1m、-1.5m、3m 及 5m 分別表示錄音裝置不同的收音距離，0.25m 為錄音裝置面對說話人 0.25 公尺遠，以此類推，而負號則表示收音距離相反，即是背對說話人來收音。我們發現在 1m 的距離下效果最好，而 -1.5m 的距離效果最差，因為其收音方向與其他距離相反，因此推測效果差與收音方向有關。

表 5. 不同空間距離影響的結果
[Table 5. Effects of different spatial distances]

註冊 / 測試	1m	-1.5m	3m	5m
0.25m	8.519	10.99	10.75	9.312

接著我們實驗雜訊的影響，表 6 為各雜訊表示情境與影響的結果，從結果可以看出，註冊與測試在相同的噪音環境下，結果都表現的較好，即對角線的部分，而當採用無噪音的音檔來註冊時，效果比起其他有噪音的來得好。

表 6. 各雜訊表示情境與影響的結果
[Table 6. Each noise condition and result]

F - 電視 / 辦公室 + 電風扇
T - 電風扇
S - 無噪音

註冊 / 測試	F	T	S
F	3.922	13.06	12.15
T	10.65	6.25	7.543
S	9.735	8.561	7.143

最後測試語速的影響，表 7 顯示各語速平均秒速與影響的結果，結果顯示慢語速的效果是最好的，其次是快語速，最差的是正常語速，我們也發現，當註冊與測試秒速差距越大，效果也越差。

表 7. 各語速平均秒速與影響的結果
[Table 7. The average speed of each speech and the result]

註冊 / 測試	Slow	Normal	Fast
Slow	8.318	9.457	11.69
Normal	9.683	9.584	11.46
Fast	11.75	11.5	9.324
平均秒數(s)	2.39	1.96	1.76

5. 結論 (Conclusions)

在這篇論文中，我們參加了 FFSVC 2020， 並基於時延神經網路與卷積神經網路實作前端的系統，同時也將兩個不同基礎的系統結合，設計出一個新的被稱為時延殘差神經網路的系統，後端實作機率線性判別分析與機率線性判別分析調適並用來融合系統。各系統分別在 FFSVC 2020 的開發集與驗證集上評估，從結果看出時延殘差神經網路勝過原先的兩個系統， 此外，我們也實驗了 GhostVLAD 並與原先的統計池化層做比較。最終，我們的最佳融合系統能在任務一上達到 minDCF 0.7703，EER 9.94%，在任務二上則是 minDCF 0.8762， EER 10.31%。

參考文獻 (References)

Brümmer, N., & De Villiers, E. (2013). The bosaris toolkit: Theory, algorithms and code for surviving the new dcf. arXiv preprint arXiv:1304.2865.

Chen, J., Cai, W., Cai, D., Cai, Z., Zhong, H., & Li, M. (2018). End-to-end language identification using netfv and netvlad. In *Proceedings of 11th International Symposium on Chinese Spoken Language Processing (ISCSLP 2018)*, 319-323. doi: 10.1109/ISCSLP.2018.8706687

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690-4699. doi: 10.1109/CVPR.2019.00482

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Proceedings of Odyssey 2010*, 14.

Li, S., Lu, X., Takashima, R., Shen, P., Kawahara, T., & Kawai, H. (2018). Improving very deep time-delay neural network with vertical-attention for effectively training ctc-based asr systems. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 77-83. doi: 10.1109/SLT.2018.8639675

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 212-220. doi: 10.1109/CVPR.2017.713

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *Proceedings of ICML 2016*, *48*, 507-516.

Liu, Y., He, L., & Liu, J. (2019). Large margin softmax loss for speaker verification. In arXiv preprint arXiv:1904.03479.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*, 2579-2605.

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 4814-4818. doi: 10.1109/ICASSP.2015.7178885

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027. doi: 10.1016/j.csl.2019.101027

NIST. (2019). NIST speaker recognition evaluation. Retrieved from https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. In arXiv preprint arXiv:1803.10963.

OpenSLR. (2020). Open Speech and Language Resources. Retrieved from https://openslr.org/

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The kaldi speech recognition toolkit. In *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Qin, X., Bu, H., & Li, M. (2019). Hi-mia: A far-field text-dependent speaker verification database and the baselines. In arXiv preprint arXiv:1912.01231.

Qin, X., Cai, D., & Li, M. (2019). Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation. In *Proceedings of Interspeech 2019*, 4045-4049. doi: 10.21437/Interspeech.2019-1542

Qin, X., Li, M., Bu, H., Das, R. K., Rao, W., Narayanan, S., & Li, H. (2020). The ffsvc 2020 evaluation plan. In arXiv preprint arXiv:2002.00387.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, 5329-5333. doi: 10.1109/ICASSP.2018.8461375

Snyder, D., Villalba, J., Chen, N., Povey, D., Sell, G., Dehak, N., & Khudanpur, S. (2019). The jhu speaker recognition system for the voices 2019 challenge. In *Proeedings of Interspeech 2019*, 2468-2472. doi: 10.21437/Interspeech.2019-2979

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado,H., Nautsch, A., . . . Lee, K. A. (2019). Asvspoof2019: Future horizons in spoofed and fake audio detection. In arXiv preprint arXiv:1904.05441.

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, *25*(7), 926-930. doi: 10.1109/LSP.2018.2822810

Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, 5791-5795. doi: 10.1109/ICASSP.2019.8683120

Zhong, Y., Arandjelović, R., & Zisserman, A. (2018). Ghostvlad for set-based face recognition. In *Proceedings of Asian Conference on Computer Vision 2018*, 35-50. doi: 10.1007/978-3-030-20890-5_3

Zhu, Y., Ko, T., Snyder, D., Mak, B., & Povey, D. (2018). Self-attentive speaker embeddings for text-independent speaker verification. In *Proceedings of Interspeech 2018*, 3573-3577. doi: 10.21437/Interspeech.2018-1158