The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 310-324

以三元組損失微調時延神經網路語者嵌入函數之語者辨識系統

Time Delay Neural Network-based Speaker Embedding Function Fine-tuned with Triplet Loss for Distance-based Speaker Recognition

葉致廷 Chih-Ting Yehn, 王伯晉 Po-Chin Wang,
張蘇瑜 Su-Yu Zhang, 陳嘉平 Chia-Ping Che
國立中山大學資訊工程學系
Department of Computer Science and Engineering
National Sun Yat-sen University
M063040008@student.nsysu.edu.tw, M063040058@student.nsysu.edu.tw,
M073040069@student.nsysu.edu.tw, cpchen@mail.cse.nsysu.edu.tw,

蕭善文 Shan-Wen Hsiao, 詹博丞 Bo-Cheng Chan, 呂仲理 Chung-li Lu
中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
swhsiao@cht.com.tw, cbc@cht.com.tw, chungli@cht.com.tw

摘要

本研究工作提出以語者驗證的 x 向量（x-vector）架構為基礎，建立一套語者辨識系統。系統訓練時，我們提出利用三元組損失（triplet loss）來拉開不同語者語句嵌入向量之間的距離。系統辨識時，則是直接使用歐氏距離做為註冊語者與測試音檔之間相似度的量測，並以最小距離的註冊語者為辨識結果。我們以名人聲音（VoxCeleb）語者辨識資料集評估所提出的系統，其中測試資料集包含 1,251 位名人的語音資料。我們所提出的系統單一輸出（top-1）的辨識正確率為 59.57 %，前五個輸出（top-5）的辨識正確率則可以達到 80.32%。

Abstract

In this research work, we build a speaker recognition system based on the x-vector framework for speaker verification. During training, we propose to use the triplet loss to increase the distance between the embedding vectors from different speakers in high-dimensional space. During recognition, we use the European distance between test-utterance embedding vector and enrolled-speaker embedding vector for similarity measure, thus predicting the enrolled speaker with the minimum distance. The proposed system is evaluated with VoxCeleb speaker recognition dataset. The test set consists of utterances from 1,251 test speakers. The proposed

model achieves the top-1 recognition accuracy of 59.57% and the top-5 accuracy of 80.32%.

關鍵詞：時延神經網路、語者辨識、三元組損失

Keywords: TDNN, Speaker Recognition, Triplet Loss

一、緒論

隨著大數據的時代到來，深度學習成為時下的熱門議題之一，並慢慢走入我們的生活之中，而身分驗證與識別就是一個主要的應用範疇，以前只出現在電影裡的生物特徵辨識系統也逐漸可以在我們的日常生活中發現蹤跡，如聲紋這類生物特徵不似傳統的鑰匙或是遙控器，可能會有遺失的風險存在，基於聲紋的語者辨識不需攜帶額外的物品，僅需聲音便可以進行辨識，達到更便利、安全的效果。

其中，語者驗證是當前熱門的研究項目，說話人須先宣稱其身份，再由語者驗證系統進行比對，做出是否通過驗證的決斷。然而在一些實際的應用上，人們開始追求指令簡明與便捷性，若能去除宣稱身份的步驟，便可讓使用者的體驗更佳。如智慧家庭產品許多都採取聲控的方式來下達指令，除了辨識使用者所下達的指令之外，更希望能對下達指令的人進行辨識，來達到一些簡單客製化回應的效果，例如我們若能分辨下達指令的是家中哪位成員，便能提供適合且客製化的回應給使用者，就像是一個簡單的播音樂指令，對家中長者可以播放台語經典歌曲，而年少者則可播放時下偶像團體的歌曲。

然而，在實際使用時需要辨識的語者或是說註冊者，大多時候會與訓練時所使用的訓練資料中的語者不同，所以無法簡單地使用基於 Softmax 的分類器來處理，為了解決這個問題，在本論文中，我們以 x 向量 [1] 架構為基礎，並透過表徵學習 (Representation Learning) [2] 的方式，從時延神經網路 (Time Delay Neural Network, TDNN) [3] 中取得嵌入向量 (Embedding Vector) ，並藉由對註冊語音之嵌入向量進行註冊的動作，為註冊者建立語者模型，在測試語音進入系統時，會對測試語音之嵌入向量與所有註冊者之語者模型進行相似度比對，並選出最相似者做為系統判定測試語音所屬之語者。為了使準確率提升，我們不僅使用交叉熵損失 (Cross Entropy Loss) 來訓練模型，也採用三元組損失 (Triplet Loss) 來對模型進行調適，使不同語者的嵌入向量在高維空間有更好的判別性。

本文主要分為五個部份：第一部份為緒論；第二部份為相關研究的回顧與探討；第三部份為研究方法與流程，介紹使用資料集、資料前處理、模型架構、訓練流程以及註冊與相似度比對等流程；第四部份則為實驗結果與分析，說明實驗環境與設定，並根據實驗結果進行分析；第五部份為結論。

二、相關研究

在語者辨識與驗證技術發展之中，高斯混和模型 (Gaussian Mixture Model) [4] 扮演十分重要的角色，而高斯混和模型-通用背景模型 (Gaussian Mixture Model-Universal Background Model, GMM-UBM) [5] 更是一個重要的應用，通用背景模型是一種大型的高斯混和模型，並非像傳統高斯混和模型針對每個語者訓練一個高斯混和模型來表示語者的特徵分佈，而是先使用所有語者的資料去訓練一個通用的背景模型，表示出語者無關 (Speaker-independent) 的特徵分佈，然後再使用指定語者的資料去調適背景模型產生語者模型。之後，為了解決在不同錄音裝置上，同語者的錄音聽起來會不一樣的問題，聯合因素分析 (Joint Factor Analysis, JFA) [6] 提出將 GMM-UBM 所得出超級向量 (Supervector) 進行因素分析，可分為通道子空間 (Channel Subspace) 與語者子空間 (Speaker Subspace)，JFA 相信若能去除通道因素的影響，那我們就能去除不同錄音條件的影響，使系統更強健。然而在[7] 中卻發現，通道部份仍然包含語者資訊，為了解決這個問題，提出了一種將語者空間與通道空間整合為單一的全局差異空間 (Total Variability Space)，而對應的全局因子則被稱為 i-vector (Identity Vector) [8]。在 2010 年至 2016 年之間的語者驗證系統幾乎都採用 i-vector 或以此為基礎進行改進。

另一方面，由於深度學習 (Deep Learning) 在圖像辨識的成功，人們也嘗試將深度類神經網路應用在語者辨識的任務上。在 2014 年 d-vector [9] 使用四層 256 維隱藏層的多層感知器進行表徵學習，並從最後一層隱藏層擷取出嵌入向量，這也啟發了其他使用卷積神經網路 (Convolutional Neural Network, CNN) [10] 或是遞迴神經網路 (Recurrent Neural Network, RNN) [11] 開發語者驗證系統的想法。到了 2016 年，使用時延神經網路 [3] 的 x-vector [1] 被提出，它最重要的特色在於對訓練音檔進行如：加入噪音、迴響、變速等數據增強 (Data Augmentation) [12]，使訓練資料呈倍數成長並獲得超越當時其他系統的強健性。如今 x-vector 已經是目前最主流的語者識別與驗證系統之一，本論

文的語者辨識系統也以 x-vector 為基礎進行改進。

三、研究方法與流程

（一）、VoxCeleb 資料集

VoxCeleb 資料集可分為 VoxCeleb1 [13] 與 VoxCeleb2 [14]，兩者皆為文本無關 (Text-Independent) 語音資料集，內容源自於 Youtube 中名人的影片，因此內容可能會有背景雜音甚至是其他人說話的聲音，資料集除了提供語者身分之外，也提供該語者國籍以及性別，兩資料集的資料分佈狀況如表一。VoxCeleb1 官方提供兩種資料集分割方式，語者驗證分割 (Verification Split) 與語者識別分割(Identification Split) ，兩種分割包含之音檔相同，差別僅在於依任務目的不同而把資料集進行不同的切割分法，其中語者驗證分割之驗證集與測試集中的語者不重複，而語者識別分割之訓練集、驗證集、測試集中的語者相同。

表一、VoxCeleb 資料分佈表

	VoxCeleb1					VoxCeleb2	
	驗證分割		辨識分割				
	dev	test	train	dev	test	dev	test
語者數	1,211	40	1,251	1,251	1,251	5,994	118
音檔數	148,642	4,874	138,361	6,904	8,251	1,092,009	36,237

在資料使用上，我們使用 VoxCeleb2 驗證集來訓練模型、使用 VoxCeleb1 識別分割的測試集來進行語者辨識的效能評估、使用 VoxCeleb1 驗證分割的測試集來進行語者辨識，更進一步的來研究是否語者驗證的準確率是否與語者辨識正相關。

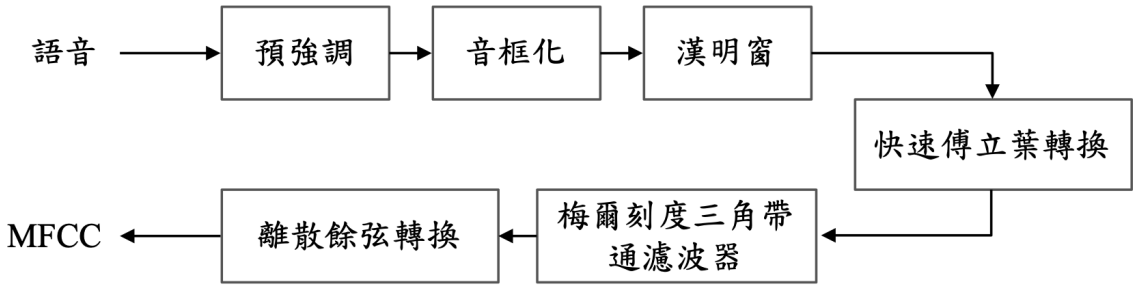
（二）、數據增強

我們採用數據增強 (Data Augmentation) 的技術，對資料加入噪音、或是對音檔加入迴

響，不僅僅增加資料的數量與多樣性，也更能使系統更加強健。我們使用 MUSAN 語料庫 [12] 加入噪音與利用房間脈衝響應 (Room Impulse Response) 加入迴響來進行數據增強。而 MUSAN 語料庫內容包含三個部分，分別為演說 (Speech) 、音樂 (Music) 與噪音 (Noise) ，演說部分的內容為朗讀書本某章節內容的語音或是美國聽證會或辯論會的公開演說；音樂部分則涵蓋古典與現代流行樂；噪音部分包含各類常見噪音，但不包括明顯可辨識說話內容的人聲。在產生數據增強後的音檔後，因為運算資源的考量，我們並不會全部使用，而是隨機取 1,000,000 個數據增強的音檔與原始音檔進行模型的訓練。

（三）、聲學特徵與正規化

我們所使用的聲學特徵為梅爾倒頻譜係數 (Mel-Frequency Cepstral Coefficient, MFCC) ，是一種針對人耳聽覺而設計的一種聲學特徵，由於人耳對不同頻率的聲音有不同的敏銳程度，故我們在頻率座標軸依梅爾刻度 (Mel Scale) 配置在低頻較密集、高頻較稀疏的三角帶通濾波器 (Triangular Bandpass Filters)，表示人耳對低頻聲音感受較為敏銳但面對高頻聲音便相較為遲鈍。梅爾倒頻譜係數聲學特徵處理流程如圖一，語音訊號經過預強調(Pre-emphasis)，來提升高頻的部份，始信號的頻譜變得平坦。之後將多個取樣點集合成一個觀測單位，稱為音框(Frame)，並對每一個音框乘上漢名窗(Hamming window)來增加音框左端與右端的連續性。由於訊號在時域上的變化很難看出訊號的特性，因此會先經由快速傅立葉轉換(Fast Fourier Transfrom, FFT) 將訊號從時域信號轉換到頻域信號上，以能量分佈來觀察。再將得到的頻譜乘上多組三角帶通濾波器，得到每一個濾波器輸出的對數能量(Log energy) 後，將對數能量經離散餘弦轉換(Discrete Cosine Transform, DCT)後即可得梅爾倒頻係數。



圖一、梅爾倒頻譜係數聲學特徵處理流。

由於實際應用時收音容易受環境以及錄音裝置等因素干擾，測試時的環境可能與訓練音檔的錄製環境大不相同，進而影響實際使用時的準確率。除了使用數據增強盡可能的模仿各種不同的環境之外，我們也加入特徵正規化的方法幫助我們增加系統的強健性，其中倒頻譜平均值與變異數正規化法 (Cepstral Mean and Variance Normalization, CMVN) 是常見特徵正規化的方法之一，藉由整段語音特徵的平均與標準差對特徵進行標準化，假設一音檔的音框長度為T，每一音框之特徵維度為N，$1 \leq t \leq T$且$1 \leq i \leq N$，$x_t(i)$表示為第t個音框中第i維的特徵，而將 $x_t(i)$正規化至 $\hat{x}_t(i)$的公式如下：

$$\hat{x}_t(i) = \frac{x_t(i) - \mu(i)}{\sigma(i)} \tag{1}$$

其中

$$\mu(i) = \frac{1}{T}\sum_{t=1}^{T} x_t(i) \tag{2}$$

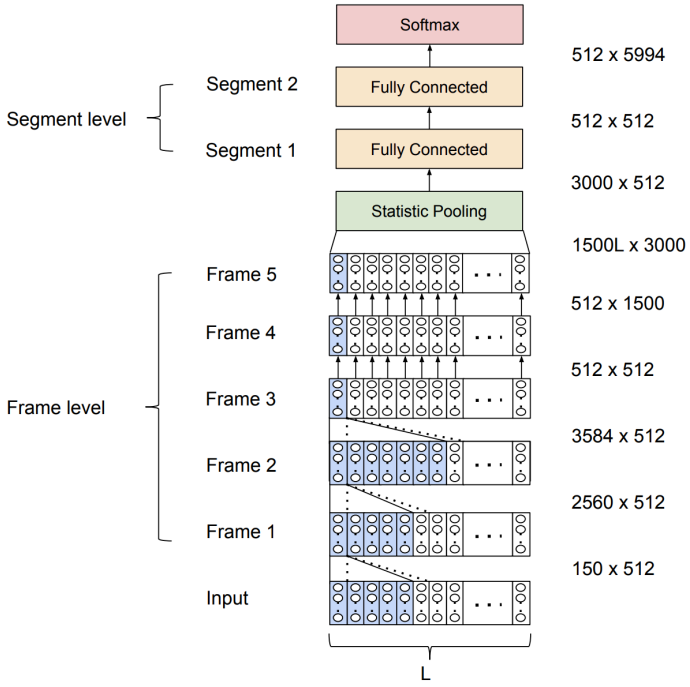
$$\sigma(i) = \frac{1}{T}\sum_{t=1}^{T} (x_t(i) - \mu(i))^2 \tag{3}$$

（四）時延神經網路

我們使用的時延神經網路模型如圖二，時延神經網路可以分為兩個部份，分別為音框層級 (Frame Level) 與音段層級 (Segment Level) ，中間由統計池化層 (Statistic Pooling

Layer) 來將音框層級的資訊轉為音段層級。而模型最後一層輸出層為 5,994 維 softmax 機率。

在我們的模型中音框層級為五層架構，第一層與第二層取相鄰的 5 個音框為輸入進行運算，到第三層則是取相鄰的 7 個音框，第四層與第五層則僅取 1 個。時延神經網路便是透過隱藏層的堆疊來提煉連續音框的特徵，且隨著隱藏層的增加，神經網路可以收集到更大範圍音框的資訊；在音框層級結束時，會在統計池化層對所有音框計算平均與變異數，整合成音段層級的資訊。此外，在這個模型中，每層隱藏層皆經過批量標準化(Batch Normalization) 與 Rectified Linear Unit (ReLU)激活函數。當模型訓練完成後，我們從 Segment 1 輸出取得 512 維的嵌入向量。



圖二、Softmax 訓練階段之時延神經網路架構圖：右側數字各層的輸入與輸出維度；L 為音檔音框數。

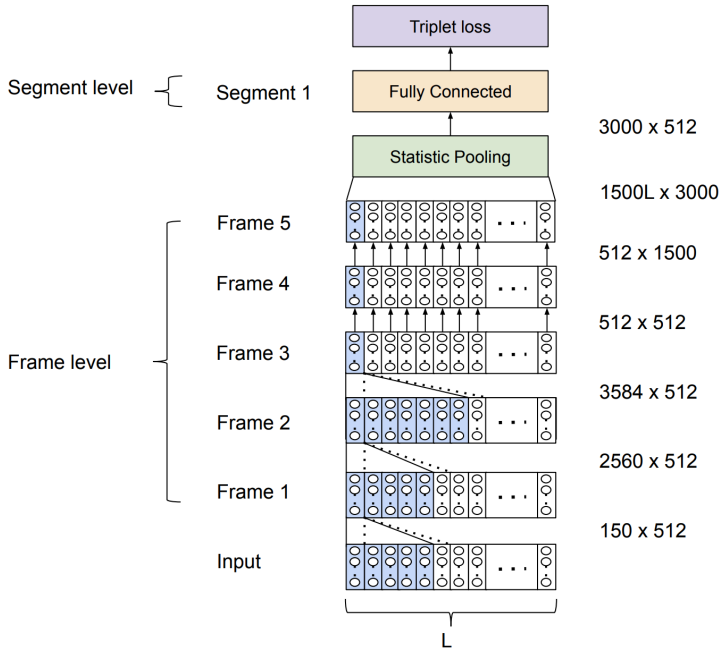
（五）損失函數

在初期訓練時延神經網路時，我們使用交叉熵損失為損失函式來更新模型，單筆資料時

的公式如下：

$$Loss_{CE} = -\sum_{i=1}^{C} y_i \, log(p(i)) \tag{4}$$

C為最後一層分類的類別數；y_i為表示該筆資料的真實類別，僅在該筆資料真實類別屬於第i類時為 1，其餘時候為 0；$p(i)$為神經網路預測第i類的機率。



圖三、三元組訓練階段之時延神經網路架構圖：右側數字各層的輸入與輸出維度；L 為音檔音框數。

在訓練完成後，為了更清楚得分辨出同語者的語音與其他語者的語音之間的差異，我們捨棄了用來提取嵌入向量的那層以後的每一層網路，如圖三所示，並改為使用三元組損失 [15] 來調整模型，三元組的定義如下：

錨點 (Anchor)： 訓練集中一語者 A 的真音檔。

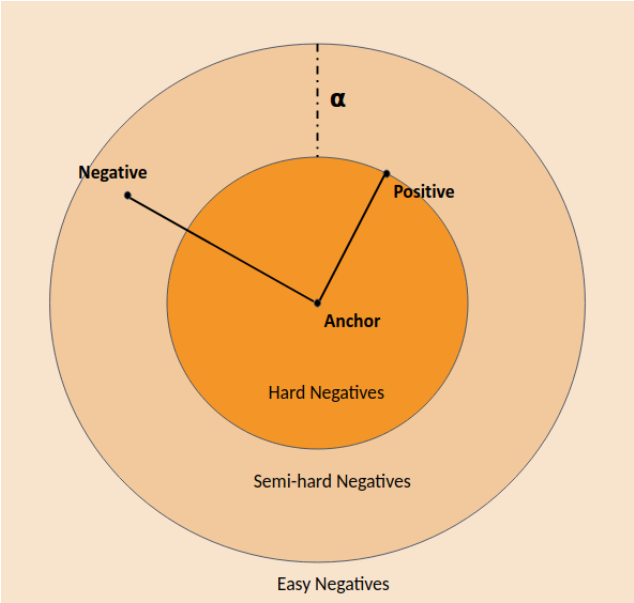
正樣本 (Positive)： 語者 A 與錨點不同的另一音檔。

負樣本 (Negative)：非語者 A 的另一語者之音檔。

在生成三元組的過程中，每筆音檔皆會成為錨點，與所有可能的正樣本去找尋一個符合條件的負樣本組成三元組，其中錨點與正樣本的組合不得重複，也就是說，任一錨點不會在它的正樣本被選為錨點時作為正樣本，而對負樣本的選擇條件為：

$$||E^a - E^p||_2^2 < ||E^a - E^n|| < ||E^a - E^p||_2^2 + \alpha \quad (5)$$

E^a為錨點的嵌入向量、E^p為正樣本的嵌入向量、E^n為負樣本的嵌入向量，α為定義負樣本的邊界值。在選定E^a與E^p的情況下，負樣本與錨點的距離必須小於正樣本與錨點的距離加上α，且大於正樣本與錨點的距離，這樣這筆其他語者的音檔才能被選為負樣本，目的在於選出半難負樣本 (Semi-hard Negatives) 來訓練，如圖四所示，如此一來，可以避免模型收斂在局部最小值 (Local Minima)。



圖四、三元組損失中各類負樣本的定義圖：圖中之負樣本為半難負樣本。

如此一來，三元組損失的損失函數定義如下：

$$Loss_{triplet} = [||E^a - E^p||_2^2 - ||E^a - E^n||_2^2 + \alpha]_+ \quad (6)$$

（六）註冊與相似度比較

不同於傳統語者識別中訓練集語者與測試語者重複，我們預設大多時候訓練集中的語者

會與實際應用時有所不同，所以我們必須對想要註冊的使用者，擷取其語音的嵌入向量，以提供註冊的功能，而我們採用對所有註冊語音得出的嵌入向量取平均做為註冊者的語者模型，當測試語音輸入時，則與所有註冊語者模型比較相似度，回傳相似度最高且高於閾值的註冊者為最終系統做出的辨識結果。

在相似度的比較上，由於三元組損失函式是計算錨點與正樣本、錨點與負樣本歐氏空間下的距離差，所以相似度評估使用歐氏距離來計算，得出的值愈小則表示兩者相似度愈高，歐氏距離的公式如下：

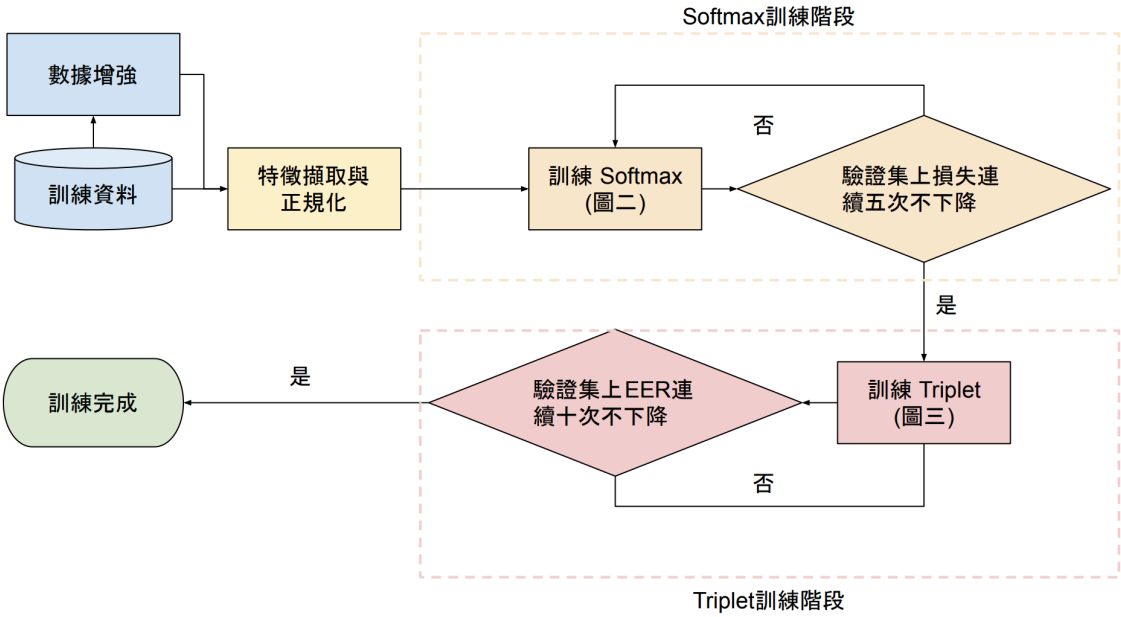
$$||X - Y||_2^2 = \sum_{i=1}^{n} (X(i) - Y(i))^2 \tag{7}$$

X與Y為要計算相似度的兩向量，皆為n維。這個方法與我們在計算三元組損失時計算嵌入向量之間距離的方法相同，也希望能藉此更符合訓練時的訴求，達到更好的效果。

四、實驗結果與分析

（一）實驗設定

在語者辨識的實驗中，我們使用 VoxCeleb2 驗證集共 5,994 位語者來訓練時延神經網路，我們從 VoxCeleb1 識別分割之驗證集每位語者取 5 句音檔進行註冊，再由測試集來測試模型，測試集共包含 1,251 位語者與 8,251 句音檔，模型的訓練流程如圖五。



圖五、模型訓練流程圖。

在此系統中，使用 30 維的 MFCC 特徵為輸入，音框 (Frame) 長度為 25 毫秒、每次移動 10 毫秒。包含特徵提取、數據增強等前處理採用 Kaldi 作為實作工具，模型訓練與相似度比對則是使用 TensorFlow 作為實作工具。

在 softmax 訓練階段過程中，我們取訓練集中每位語者 10 筆音檔做為驗證集，且使用早停法 (Early Stopping) 的機制，每訓練完一輪訓練資料去計算一次在驗證集上的損失，若該損失連續不下降 5 次則停止訓練。

在三元組訓練階段時，使用三元組損失訓練模型，我們並不從所有的訓練資料中組成三元組，而是每次取 90 位語者，每位語者取 20 句音檔，共 1,800 句音檔，我們由這些音檔構成三元組，來訓練與更新模型，這樣對語者進行採樣的方式相對於事先找出所有語者的三元組，能更快速地反應模型的現況，並找出對改善當前模型有幫助的三元組。在三元組訓練階段時，使用 VoxCeleb1 驗證分割的測試資料為驗證集，並進行語者驗證，以語者驗證的 EER (Equal Error Rate) 為判斷是否停止訓練的標準。每次訓練完一次採樣的資料後，便會進行一次語者驗證，由於每次採樣的資料僅 1,800 筆，所以連續 10 次的參數更新 EER 皆沒有下降才停止訓練。

在超參數方面，實驗中學習率皆設為 0.001，優化方法使用 Adam 演算法，三元組損失

的選擇邊界α為 0.2。

（二）實驗結果

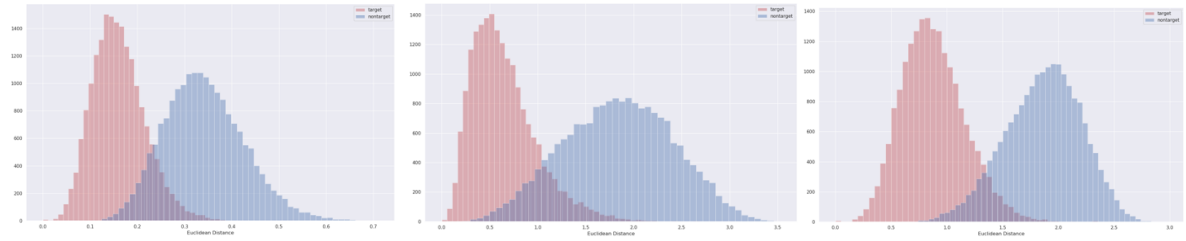
我們在本節比較訓練流程中是否使用三元組損失之間的差別 (是否有圖五中的三元組訓練階段) ，也觀察是否使用 softmax 預訓練模型對三元組損失訓練方法 (是否有圖五中的 softmax 訓練階段) 的影響，探討是否三元組損失在少了 softmax 的條件下，是否使同類資料的嵌入向量在高維空間中聚集。在三元組損失的損失函式中我們可以看出，訓練的核心概念在於拉開不同類別之間的距離，但缺少使同類別內的資料內聚的能力，例如在三元組的選擇中並未對正樣本的選擇做出限制，缺少拉近正樣本與錨點之間的距離的功能，所以在少了 softmax 時，也缺少了使同類資料內聚的能力，僅憑三元組損失把不同類別的距離拉開，是否能對訓練資料未曾出現過的語音得出判別性佳的嵌入向量是值得探討的問題。

首先，表二為在 VoxCeleb1 驗證分割的測試資料中利用官方提供的 trials 進行驗證的實驗結果，也是在使用三元組損失時使用的驗證資料的結果，包含 40 位語者共 37,720 筆 trials，target : nontarget 為 1:1。我們除了 EER 之外，也使用 minDCF (Minimum Decision Cost Function)來評估系統，參數設定比照 [14]，可以發現僅使用 softmax 時 EER 為 9.64%，而使用三元組損失而未使用 softmax 預訓練模型時，EER 則略高來到 10.39%，不過使用三元組損失且加上 softmax 預訓練模型做為起始權重的話，EER 有效降低至 6.84%，minDCF 也是最低，為 0.6278。

表二、在 VoxCeleb1 語者驗證之實驗結果

訓練方法	*EER*	*minDCF*
Softmax	9.64 %	0.7174
Triplet	10.39 %	0.8919
Softmax + Triplet	6.84 %	0.6278

此外，圖六為各訓練方法在語者驗證上分數的分佈圖，橫軸為 trials 中比對目標間的歐氏距離，而縱軸代表預測結果為該距離時資料的數目，紅色為 target trials 的分數分佈情況，藍色為 nontarget trials 的分數分佈情況。我們觀察的重點有二：一是 target 與 nontarget 分佈重疊的部份，代表模型可能分類錯誤的部份，重疊的面積愈小表示愈能將不同語者分類清楚，系統的效能也愈佳，從圖中可以發現 Softmax+Triplet 面積最小，同時也在實驗中有最好的效果；第二是 target 與 nontarget 分佈中心，在 Softmax 我們看到分布狀況與常態分佈相似，但有經三元組調適後，target 與 nontarget 的分佈中心皆有拉開彼此之間距離的情形發生，展現三元組拉開不同語者之間距離的效果。



(a) Softmax　　(b) Triplet　　(c) Softmax+Triplet

圖六、在 VoxCeleb1 語者驗證之分數分佈圖

在語者辨識上，我們以 Top-1 準確率與 Top-5 準確率為評估標準，Top-1 準確率表示僅系統判斷相似度最高者為測試語音所屬的語者才算正確，Top-5 準確率則是系統判定前五相似者中有測試語音所屬語者即算正確。

實驗結果如表三，我們發現有 softmax 預訓練且經三元組損失調適之後，在語者辨識上也有所進步，相較於僅使用 softmax 訓練 Top-1 準確率提升了約 5%，Top-5 準確率更上升了約 6.3%。但從實驗結果中我們也發現，沒有使用 softmax 預訓練模型的話效能會大大的降低，這點在語者辨識時尤其明顯。

表三、在 VoxCeleb1 語者辨識之實驗結果

訓練方法	*Top-1 準確率*	*Top-5 準確率*
Softmax	54.59 %	73.67 %

Triplet	23.68 %	45.58 %
Softmax + Triplet	59.57 %	80.32 %

五、結論

在本論文中，我們以 x 向量架構為基礎，開發語者辨識系統，透過改變損失函式的方法，將原本後端繁瑣的分類流程簡化至計算歐氏距離來比較測試語音與註冊者語音的相似度。此外，不同於常見的語者識別訓練集與測試集中的語者相同，為了使如智慧家庭產品等應用上更加方便，在註冊新增或刪除用戶上不受限制，我們在訓練集語者與測試集語者不同的條件下進行辨識，利用嵌入向量對註冊者建立語者模型，並在測試時找出與測試語音最相似的註冊者做為系統判定的結果。在實驗中，我們比較是否使用三元組損失的差異，發現無論在語者驗證上或是語者辨識上皆有所幫助，在 VoxCeleb1 識別分割測試集單一輸出（top-1）的辨識正確率為 59.57 %，前五個輸出（top-5）的辨識正確率則可以達到 80.32%，但另一方面我們也建議在使用三元組損失時，使用 softmax 預訓練模型，可以使模型更穩定且有更好的效果。

參考文獻

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in Proc. ICASSP, 2018.

[2] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture forefficient modeling of long temporal contexts,” in Proc. Interspeech, 2015.

[4] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” IEEE transactions on speech and audio processing, vol. 3, no. 1, pp. 72–83, 1995.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted

gaussian mixture models," Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1435–1447, 2007.

[7] N. Dehak, Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification. PhD thesis, ´ Ecole de technologie sup´erieure, 2009.

[8] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[9] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056, IEEE, 2014.

[10] Y.-h. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez, and C. Parada , "Locally-connected and convolutional neural networks for small footprint speaker recognition," in Proc. Interspeech, 2015.

[11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in Proc. ICASSP, 2016.

[12] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in Proc. Interspeech, 2017.

[14] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in Proc. Interspeech, 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823, 2015.