The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 5-20

運用類神經網路方法之語音端點偵測研究

A Study on Voice Activation Detection by Using Neural Networks

鄧有志　Yu-Chih Deng
國立臺北大學通訊系
Department of Communication Engineering
National Taipei University
ted790.ycd@gmail.com

江振宇　Chen-Yu Chiang
國立臺北大學通訊系
Department of Communication Engineering
National Taipei University
cychiang@mail.ntpu.edu.tw

潘振銘　Chen-Ming Pan
中華電信研究院
Chunghwa Telecom Laboratories
chenming@cht.com.tw

摘要

本研究以深層類神經網路 (Deep Neural Network, DNN) 進行語音端點偵測，討論了以下影響語音端點偵測表現的幾個變量：(1) 特徵參數抽取時考量的分析視窗大小、(2) DNN 層數、(3) 訊躁比以及(4) 背景環境類型。實驗是使用台北大學雜訊語料庫 (NTPU Noise Corpus)，此資料庫是由智慧型手機錄製的各種背景雜訊以及 TCC300 語料庫混音而成，背景環境包含：(1) 公車站、(2) 捷運站、(3) 火車站、(4) 餐廳，而混音的訊躁比有：10dB、5dB、0dB 以及乾淨語音。系統評量的標準為音框正確率 (frame accuracy) 以及 equal error rate (EER)。實驗結果指出特徵參數分析視窗越大而在訓練與發展集合的表現有明顯變好的趨勢，但在測試集合則進步幅度較小。DNN 層數在 2 layer 時的 multi-condition 其表現較好，訊躁比越高則進步也比較顯著，尤其是在背景環境為餐廳的情況下。最後 multi-condition 訓練法中的各個 condition，在測試集合的表現皆優於 matched-condition，證實了 multi-condition 中的各個 condition，在 hidden layer 中能夠互相的學習。

Abstract

This study used DNN (Deep Neural Network) to process Voice Activation Detection, and discussed the following variable which affect the performance of VAD: (1) The analyzed window size of MFCC feature extraction, (2) Layer number of DNN, (3) Signal to Noise Ratio, and (4) The type of background condition. This experiment used NTPU Noise Corpus, which is mixed by many kinds of background noise recorded by smart phone and TCC300 Corpus. The background noise includes: (1) Bus Stop, (2) MRT, (3) Train Station, (4) Restaurant, and the SNR is 10 dB, 5 dB, 0 dB and clean speech. Evaluated standards of system are frame accuracy and equal error rate (EER). The experiment result indicated that when the feature parameter analyzed window is bigger, the performances of training and validation set obviously become better, but the improved range of outside test is smaller. When layers number of DNN in 2 layer, the performance of multi-condition is better, and when the SNR is higher, the improvement is obviously, in particularly, the background condition is restaurant. In conclusion, in every conditions of the multi-condition training, the performances of outside test are all better than in matched-condition, and it proved that every conditions in multi-condition can learn each other in the hidden layer.

關鍵詞：語音端點偵測，MLP，DNN，台北大學雜訊語料庫

Keywords: VAD, MLP, DNN, NTPU Additive Noise Corpus, layer #, feature frames, multi-condition, matched-condition, frame accuracy, EER.

一、緒論

（一）、研究動機

隨著時代劇進，科技的進步猶如一眨眼一瞬間。各式創新的技術及想法使得生活更加趨於便利及高效率，智慧型掌上裝置及人機互動裝置的普及已成現代人生活中不可或缺的部份。其中語音處理技術亦被廣泛的應用於智慧型掌上裝置中，提升了其使用頻率及便捷性，例如智慧型裝置上的通話雜訊消除及免持電話等，使用者可以直接地改善通話時的品質。

以類神經網路為基礎(NN-based)的語音端點偵測技術(Voice Activation Detection, VAD)逐漸成熟，在語音辨認的品質已有不錯的表現。在語音辨認系統中必須要有語音

端點偵測，這項技術是重要的關鍵。以通話雜訊消除系統為例，對於智慧型裝置使用者提供良好的語音端點偵測處理技術，可以節省行動裝置的負擔並提升其續航力。因此，為滿足應用在不同環境的實際需求，則必須對於語音端點偵測的性能做進一步的探討，以利開發較完善的語音辨認系統。

（二）、文獻回顧

語音端點偵測(Voice Activation Detection, VAD)目的是在於檢測語音訊號中，語音片段的開始與結束，對於 ASR 系統是很重要的前處理(Front-end)工作之一。因為語音端點偵測的效能，會直接影響 ASR 系統的辨識率，所以此類方法被應用於語音喚醒(Voice Trigger, VT)、語音會議(Audio conference)、語音編碼(Speech codding)、免持通話(Hands-free)、語音降噪(Speech enhancement)、聲音定位(Sound positioning)、語者辨識(Speaker Recognition)及語音辨識(Speech Recognition)裡。我們可以將眾多 VAD 的演算法分成以下四種方法，以 Energy-based[1][2]、Statistical-based[3][4]、GMM-based[5][6][7][8]與 NN-based[9][10][11][12]之檢測法。

在 Energy-based 檢測法中語音部分的 energy 明顯比雜訊 energy 大。所以我們可以在時域上定義一個簡單的 Threshold 來對於 Energy、Zero Crossing Rate (ZCR)和 Pitch 做判斷，並使用 VAD state machine 來描述語音的開始與結束。

Statistical-based 檢測法，此方法在過去的研究是觀察時域上某頻帶上語音訊號長期穩定的變化以及在頻域上觀察其平坦度，利用這些方法來判別出各該片段為語音還是非語音的變化。近期的 Statistical-based VAD 研究則是試圖去最佳化檢測雜訊的存在，例如是使用 low-variance spectrum 估計法並且配合統計檢測機制來確定最佳的 Threshold，並且搭配 Hangover state machine 來避免語音快結束時的語句，在 low-energy 的語音片段中出現誤判(false reject)。

GMM-based 檢測法，此方法主要是依據語音內容為基礎的非監督式訓練法，需要利用 Threshold 來對語音及非語音建立模型後做判斷。我們將此檢測法利用 TCC300 乾淨的語料庫實驗後可以發現，在實驗結果的 ROC (Receiver Operating Characteristics) curve 上其 EER (Equal Error Rate)的表現結果並不如預期。故此我們則使用 NN-based 檢測法於實驗中，希望能得到更佳的結果。

在 NN-based 檢測法中傳統是使用 MLP 的架構，但近年來許多學者對於 NN 有突破性的研究成果，所以逐漸有 DNN、RNN 甚至是 LSTM、GRU 等架構出現。DNN 改善了傳統 MLP 只有三層之架構(input layer、hidden layer 與 output layer)，增加了 MLP 在 hidden layer 的數目、hidden layer 裡面 node 的數目，使得整個 network 變得又寬且深。並且 DNN 加入了 dropout 及 mini-batch 在傳統的 MLP 訓練過程中，對於 neural network 的 unit 依照一定比例暫時性隨機的丟棄，其優點：是在於訓練數據較少時，則可以用於避免 over-fit；但是缺點：則會使訓練時間加長，但不影響其測試的時間，且每一個 mini-batch 都在訓練不同的 network。

（三）、研究方向

本研究考慮語音端點偵測對於 ASR 系統的影響，提出了以 NN-based 的檢測法應用於研究中，並自行建立台北大學雜訊語料庫。此訓練法可用來對訓練資料做模型的建立及分類，相關說明如下: 利用 MLP、DNN 類神經網路，對台北大學雜訊語料庫做訓練及測試，各別建立語音以及非語音模型，將個別類神經網路輸出之結果，用來探討語音端點偵測的表現。從 ASR 系統與科學的角度出發，對應用於 ASR 系統之語音端點偵測分析，並討論哪種神經網路之架構或方法更適合運用至 ASR 系統。

在過去的研究中，發現以 Energy-based 與 GMM-based 之語音端點偵測，對於 ASR 系統的表現其效果有限。本論文以 NN-based 的方法，提出一語音端點偵測之技術並探討不同種類的輸入資料對於語音端點偵測之影響。

二、語料庫簡介

（一）、TCC300 語料庫

本論文使用了 TCC300 語料庫[13]。TCC300 語料庫是由國立台灣大學、國立交通大學、國立成功大學各自擁有之語料庫集合而成，各校錄製之語料庫皆屬於麥克風是朗讀語音。其中台大語料庫主要包含詞語短句，文本經過仔細設計並考慮了音節及其相連出現機率，由 100 人錄製而成。交大及成大語料庫主要包含長文語料，文章由中研院提供之 500 萬詞詞類標示語料庫所選取，每篇文章包含數百字在切割成 3-4 段，每段包含至少 231 字由 200 人朗讀錄製，每人朗讀文章皆不相同。

表一：TCC300 語料庫資訊統計表

語料庫	文章屬性	語者總數		總音節數		檔案總數	
TCC300	長短句	男	152	男	193,167	男	4,614
		女	151	女	197,296	女	4,265
		總計	303	總計	390,463	總計	8,879

（二）、智慧型手機雜訊資料庫

隨著科技的進步，錄音裝置已經不僅限於傳統的外接麥克風，像是筆記型電腦、錄音筆、平板、智慧型手機等都具有麥克風錄音裝置。但在這些裝置中並無法掌控其錄音的品質，錄音裝置可能因為年久使用下而造成器材的損耗，然而這類的損耗以不影響人耳能夠識別的條件下並不會被更換，為了能夠保有錄音的便捷性及其真實性，則使用現代人都具備的智慧型手機來錄製其語料庫，以進行在多環境下的語音端點測試。

2、錄音計畫及內容

台北大學雜訊語料庫所錄製的內容為多種環境下的雜訊語料庫，所有語料檔案均以 Sampling Rate: 16kHz、Sound Encoding: Lin16 及 Channel: 1 的 PCM 格式設定進行錄製，並將音檔儲存成*.wav 檔案格式。錄音裝置使用 HTC Desire 並利用實驗室的 Android 錄音程式進行錄音。

雜訊資料庫共分成 4 個類別為台北大學學校餐廳、板橋火車站、板橋捷運站、板橋公車等候亭來進行錄製。每個類別皆會錄製近 60 分鐘長度的音檔，語料庫錄製者為 ycdeng。

表二：台北大學雜訊語料庫資訊統計表

雜訊種類	地點	日期	時間	錄製者	裝置	雜訊長度
學校餐廳	台北大學	12/27/2016	11:54~12:44	ycdeng	HTC Desire	50:48.96
火車站	板橋火車站	12/26/2016	19:02~20:02	ycdeng	HTC Desire	1:00:22.40
捷運站	捷運板南線	12/26/2016	20:11~21:03	ycdeng	HTC Desire	52:12.54
公車站牌	板橋公車站	12/26/2016	17:47~18:48	ycdeng	HTC Desire	1:01:08.35

3、語料庫使用

將錄製完的雜訊資料庫與 TCC300 語料庫做結合並形成加成性雜訊。將其隨機分成 7:2:1 的訓練集、測試集與發展集三部分進行實驗，且每一部分的加成性雜訊環境比例皆相同。

表三：台北大學雜訊語料庫

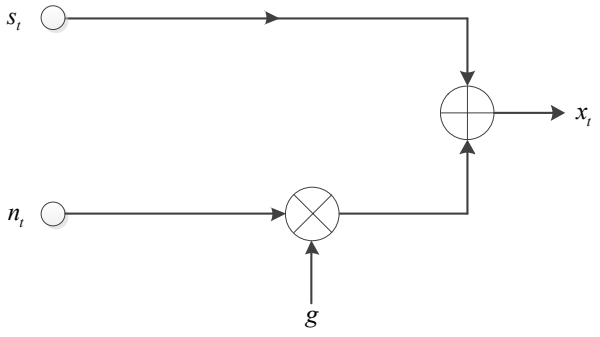
語料庫	NTPU Additive Noise Corpus	
取樣頻率	16kHz	
取樣編碼	Lin 16	
聲道	1	
語音內容	中研院 500 萬詞詞類標記語料庫	
語音長度	長句+短句(TCC300)	
模式種類	**Clean**	**Multiple Additive Noise**
Training Mode	無加成性噪音	加成性噪音: 餐廳(台北大學)、火車站(板橋)、捷運站(板橋)、公車站牌(板橋) SNR: 0、5、10 dB
Testing Mode	無加成性噪音	加成性噪音: 餐廳(台北大學)、火車站(板橋)、捷運站(板橋)、公車站牌(板橋) SNR: 0、5、10 dB
Development Mode	無加成性噪音	加成性噪音: 餐廳(台北大學)、火車站(板橋)、捷運站(板橋)、公車站牌(板橋) SNR: 0、5、10 dB

表四：台北大學雜訊語料庫細項說明

雜訊種類	SNR 種類	語者總數	男女比	Utterance	Utterance length	檔案總數
Restaurant	0、5、10 dB	131 人	65 : 66	131 句	50:48.96	393 筆
Train Station		157 人	78 : 79	157 句	1:00:22.40	471 筆
MRT		135 人	67 : 68	135 句	52:12.54	405 筆
Bus Stop		160 人	80 : 80	160 句	1:01:08.35	480 筆
Clean	∞ dB	160 人	98 : 62	160 句	52:45.04	160 筆

4、Noise Speech 之建立方法

Noisy speech 資料庫的建立方法如圖一所示，首先 TCC300 語料庫為一個在安靜環境下錄製的麥克風語料，所以我們可以假設 TCC300 語料為 clean speech，也就是 TCC300 的語音部分就可以視為 clean speech，可以利用語音的切割資訊 (label 檔案)來紀錄 TCC300 語料庫音檔的語音及非語音之段落，以便於用來計算使用於混音所需要的 SNR 資訊。



圖一：Noisy speech 資料庫建立方法圖

Noise speech x_t建立之數學式如下式所示

$$x_t = s_t + g \cdot n_t \tag{1-1}$$

其中s_t為 TCC300 的語音部分(clean speech) n_t、則是雜訊的部分(noisy data)，然而 g 為雜訊部分n_t欲合成出資料 Noise speech x_t所需要乘上之倍率並加上語音部分s_t。一般來說標準已知定義的語句之 Global SNR 算法[14]如式(1-2)所示

$$GSNR = 10\log_{10}\left(\frac{\sigma_s^2}{\sigma_n^2}\right) \tag{1-2}$$

其中σ_s^2以及σ_n^2可分別為語音訊號的功率以及雜訊的功率，σ_s^2可由式(1-3)計算：

$$\sigma_s^2 = \sum_{t=0}^{T-1} s_t^2 \cdot \delta_{speech}(t) \Big/ \sum_{t=0}^{T-1} \delta_{speech}(t) \tag{1-3}$$

其中s_t代表某語句 (TCC300 語料之語句) 的第 *t* 個 sample 的 sample value，*T* 代表某語句以 sample 數為單位的長度，而$\delta_{speech}(t)$代表第 *t* 個 sample 是否為語音信號，也就是

$$\delta_{speech}(t) = \begin{cases} 1, & \text{if sample } t \text{ is a speech sample} \\ 0, & \text{if sample } t \text{ is a non-speech sample} \end{cases} \tag{1-4}$$

而雜訊的功率σ_n^2可由式(1-5)計算得到

$$\sigma_n^2 = \frac{1}{L}\sum_{t=0}^{L-1} g^2 \cdot n_t^2 = g^2 \cdot \hat{\sigma}_n^2 \tag{1-5}$$

其中n_t代表某雜訊信號段落的第 *t* 的 sample 之 sample value； *L* 代表此雜訊段落的 sample 數； *g* 代表雜訊信號的放大倍率(magnification)； $\hat{\sigma}_n^2 = \frac{1}{L}\sum_{t=0}^{L-1} n_t^2$ 代表原始雜訊段落的 noise power。為了要使混音的 noisy speech 之 GSNR 符合實驗的要求值，我們必須調整 g 的值如式(1-6)所示

$$g = 10^{\left(\frac{GSNR}{20}\right)} \times \frac{\sigma_s}{\hat{\sigma}_n} \tag{1-6}$$

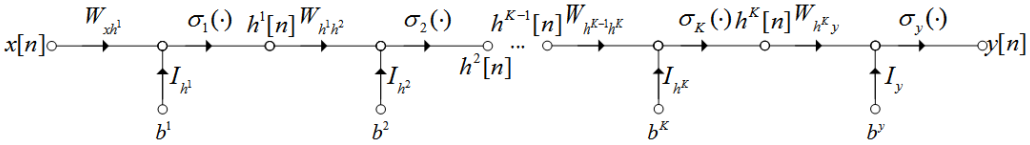
圖 2 之混音計算法是使用了改良式的混音計算法[15]。相較於以往的混音演算法，則是將多個輸入數據做線性疊加的方式，可以很明顯地聽到背景雜訊、波形會突變失真並且出現比較輕微得爆音，造成少數的語音並無法辨識並且會產生溢出的問題。改善方法；對於以往的混音計算法來說是使用更多的位元數來表示其音檔的一個 sample，在混音後降低其振幅並使其分布在 16bit 所能表式的範圍內，此種方法為 Normalize 做法[16]，但缺點則是混音後的聲音非常小且其效果不見理想。但改良式混音算法解決溢出的方法則是箝位 (clamping)，箝位以上的值為所能表式的最大值，當發生下溢位時則箝位平移後為所能表式的最小值如下式

$$x_t \leftarrow \begin{cases} MAX, & x_t > MAX \\ MIN, & x_t < MIN \\ x_t, & otherwise \end{cases} \tag{1-7}$$

三、NN-based VAD 方法

（一）、Deep Neural Network (DNN)

Deep Learning 的概念是可以讓各個模組函數經過線性或是非線性的組合後能有具有 end-to-end global optimization 特性，其中最具代表性的 Deep Learning 是其推導出的 Deep Neural Network(DNN)[9][17][18]。若從其架構來看，DNN 與傳統的 Multilayer Perceptron(MLP)是相同的，但是傳統的 MLP 大多就只有使用到三層的架構來進行，這三層分別是一個輸入層(input layer x)、一個隱藏層(hidden layer)以及一個輸出層(output layer y)。然而 DNN 則是將其 hidden layer 的數目增加，hidden layer 內的 node 數目也增加，目的是要讓整個 Neural network 很深且很寬。在此處則不是以一般介紹 DNN 的示意圖來做表示，如圖二所示則是使用 Signal flow graph 來描述這樣的系統。



圖二：DNN signal flow graph

其中 x 和 y 分別代表是系統的輸入及輸出向量、h 代表是 hidden layer 的輸出、**W** 代表轉置矩陣、**b** 代表偏壓(bias)向量、**I** 代表是單位(identity)矩陣、$\sigma(\cdot)$代表為激發函數

(Activation function)以及最後 n 代表輸入或輸出參數的時間 index，下式為其輸入和輸出的數學關係式

$$\begin{cases} \mathbf{y}[n] = F(\mathbf{x}[n]) = \sigma_y(\mathbf{W}_{h^K y}\mathbf{h}^K[n] + \mathbf{I}_y\mathbf{b}^y) \\ \mathbf{h}^k[n] = \sigma_k(\mathbf{z}^k[n]) = \sigma_k(\mathbf{W}_{h^{k-1}h^k}\mathbf{h}^{k-1}[n] + \mathbf{I}_k\mathbf{b}^k),\ k = 2 \sim K \\ \mathbf{h}^1[n] = \sigma_1(\mathbf{z}^1[n]) = \sigma_1(\mathbf{W}_{xh^1}\mathbf{x}[n] + \mathbf{I}_1\mathbf{b}^1) \end{cases} \tag{4-1}$$

(4-1)式中的激發函數 $\sigma(\cdot)$ 可以是 element-wise 的 Sigmoid、Hyperbolic、Linear、Rectified linear functions，而訓練整個 DNN 的 criterion 可以是 Minimum mean squared error(MMSE)或是 Maximum likelihood(ML)；其中 ML 的 criterion 在預估目標是以 category 的情況下就等同於 Minimum cross entropy(MCE)的條件，根據以上的條件，可以利用以下的數學式來表示 DNN 的訓練過程

$$\begin{aligned} &\mathbf{W}^*, \mathbf{b}^* = \arg\min_{\mathbf{W},\mathbf{b}} J(\mathbf{W}, \mathbf{b}) \\ &J(\mathbf{W}, \mathbf{b}) = \begin{cases} \sum_{n=0}^{N-1} \left\| \hat{\mathbf{y}}[n] - F(\mathbf{x}[n]) \right\|_2^2 & \text{for MMSE} \\ -\sum_{n=0}^{N-1} (\hat{\mathbf{y}}[n])^T \log(F(\mathbf{x}[n])) & \text{for MCE} \end{cases} \end{aligned} \tag{4-2}$$

其中 $\hat{y}[n]$ 代表是第 n 個的 input sample $x[n]$ 所對應到的正確答案(Reference)，而這樣的訓練過程是利用 Gradient decent 的方法 Integrative 來得到最佳解，因此每一層的 $\mathbf{W}$ 和 $\mathbf{b}$ 都是其他層 $\mathbf{W}$ 和 $\mathbf{b}$ 之函數，所以有 chain rule 的特性並可以歸納出著名的 Back propagation 演算法，此算法可以使 DNN 各層的參數估計是一個以統一的演算方法進行，並且便於訓練模型與各 layer 的串接模組化；在理論上如果當 K 層數越大則整個 DNN 的預估能力越強。在(4-2)式中是 DNN 常見的訓練準則，對於 MMSE 準則來說

$$\begin{aligned} J_{MMSE}(\mathbf{W}, \mathbf{b}; \mathsf{S}) &= \frac{1}{M}\sum_{m=1}^{M} J_{MMSE}(\mathbf{W}, \mathbf{b} \mid \mathbf{o}^m, \mathbf{y}^m) \\ J_{MMSE}(\mathbf{W}, \mathbf{b} \mid o, y) &= \frac{1}{2}\left\| \mathbf{h}^K - \mathbf{y} \right\|^2 = \frac{1}{2}(\mathbf{h}^K - \mathbf{y})^T(\mathbf{h}^K - \mathbf{y}) \end{aligned} \tag{4-3}$$

M 為訓練資料的總數而 m 則代表其 index。對於 category 的情況來說，假設 $\mathbf{y}$ 是一個機率分布、C 表示類別數量而 i 為其 index，則 ML 準則為

$$J_{CE}(\mathbf{W},\mathbf{b}\,|\,\mathsf{S}) = \frac{1}{M}\sum_{m=1}^{M} J_{CE}(\mathbf{W},\mathbf{b}\,|\,\mathbf{o}^m,\mathbf{y}^m)$$
$$J_{CE}(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = -\sum_{i=1}^{C} y_i \log h_i^K \tag{4-4}$$

給定了訓練準則、模型參數**{W,b}**，可以利用上述提到的 back propagation 演算法來做學習；模型參數可以使用以下公式來做優化

$$\mathbf{W}_{t+1}^k \leftarrow \mathbf{W}_t^k - \varepsilon\Delta\mathbf{W}_t^k$$
$$\mathbf{b}_{t+1}^k \leftarrow \mathbf{b}_t^k - \varepsilon\Delta\mathbf{b}_t^k \tag{4-5}$$

(4-5)式中W_t^k及b_t^k分別是在第 t 次迭代更新後第 k 層的權重矩陣和偏壓向量。

$$\Delta\mathbf{W}_t^k = \frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{w}_t^k} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o}^m,\mathbf{y}^m)$$
$$\Delta\mathbf{b}_t^k = \frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{b}_t^k} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o}^m,\mathbf{y}^m) \tag{4-6}$$

在上式中分別是在第 t 次迭代時的平均權重矩陣梯度和平均偏壓向量梯度。這之中M_b表示訓練 samples、ε 為 Learning rate。輸出層權重矩陣相對於訓練準則的梯度取決於其訓練準則，在 Category 的情況下則使用 CE 訓練準則(4-4)式和 softmax 輸出層

$$\nabla_{\mathbf{W}_t^K} J_{CE}(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = \mathbf{e}_t^K(\mathbf{h}_t^{K-1})^T = (\mathbf{h}_t^K - \mathbf{y})(\mathbf{h}_t^{K-1})^T$$
$$\nabla_{\mathbf{b}_t^K} J_{CE}(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = \mathbf{e}_t^K = (\mathbf{h}_t^K - \mathbf{y}) \tag{4-7}$$

對於隱藏層$(k = 2 \sim K-1)$則有

$$\nabla_{\mathbf{W}_t^k} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = [\sigma_k'(\mathbf{z}_t^k) \bullet \mathbf{e}_t^k](\mathbf{h}_t^{k-1})^T$$
$$\nabla_{\mathbf{b}_t^k} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = \sigma_k'(\mathbf{z}_t^k) \bullet \mathbf{e}_t^K \tag{4-8}$$

(4-8)中，$\mathbf{e}_t^k \quad \nabla_{\mathbf{h}_t^k} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y})$是在第 k 層的錯誤信號、$\bullet$表示元素相乘、$\sigma_k'(\mathbf{z}_t^k)$則是激活函數的元素導數。錯誤訊號$\mathbf{e}_t^k$的表示如下

$$\mathbf{e}_t^{K-1} = \nabla_{\mathbf{h}_t^{K-1}} J(\mathbf{W},\mathbf{b}\,|\,\mathbf{o},\mathbf{y}) = (\mathbf{W}_t^K)^T\mathbf{e}_t^K$$
$$\mathbf{e}_t^{k-1} = (\mathbf{W}_t^k)^{\mathrm{T}}[\sigma_k'(\mathbf{z}_t^k) \bullet \mathbf{e}_t^k] \tag{4-9}$$

在此對於 DNN 的 back propagation 演算法關鍵步驟做說明。

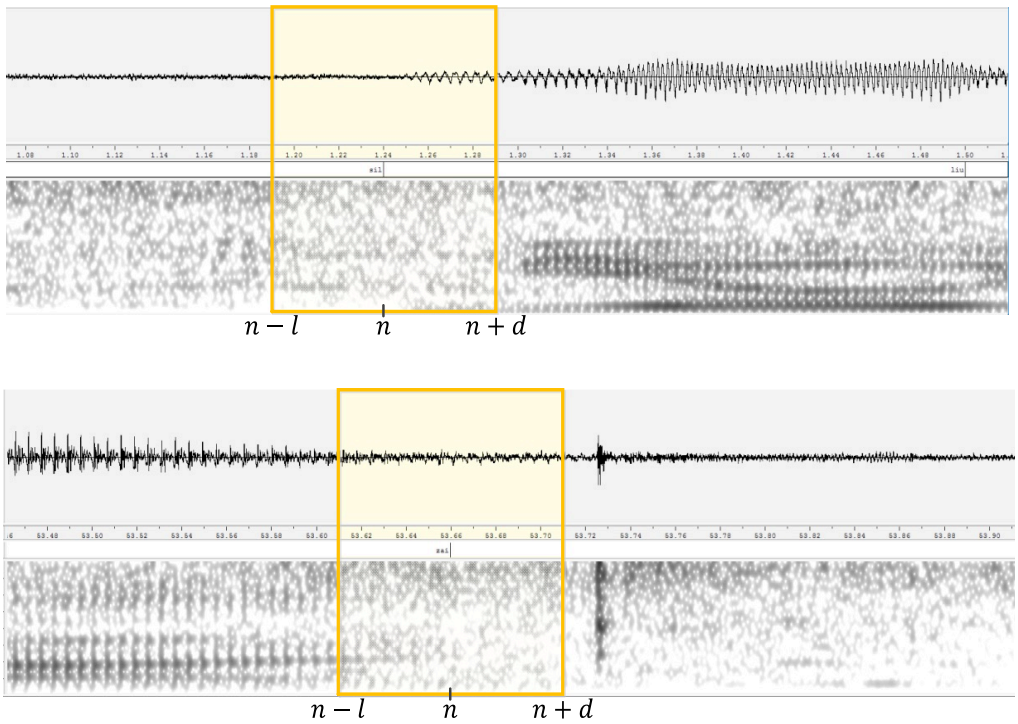
（二）、NN-based VAD 實驗設定

本實驗所使用的語料庫為表 1.3 所示，自行錄製雜訊及混音的台北大學雜訊語料庫。其特徵參數使用了 12 維度的 MFCC 再加上 1 維度之 energy，MFCC 設定裡則是使用了 24 個 filter bank 並且在 cepstrum 裡取前 12 個 cosine 來描述其波峰特性，MFCC 特徵參數抽取之參數設定如表五所示。音檔中語音及非語音之 label 是使用 HTK (Hidden Markov Model Toolkit)來做標記。

表五：MFCC 特徵參數抽取設定

Config of MFCC Feature Extraction	
SOURCEFORMAT	Alien
HEADERSIZE	0
SOURCERATE	625.0
TARGETKIND	MFCC_E
TARGETRATE	100000.0
WINDOWSIZE	320000.0
USEHAMMING	T
PREEMCOEF	0.97
NUMCHANS	24
CEPLIFTER	22
NUMCEPS	12
ENORMALISE	F
ZMEANSOURCE	T

為了要模擬實際系統的語音端點偵測，所以本論文使用將特徵參數 delay 的方式。如圖三所示，是因為在某個時間點下的語音能量(Energy)上升或是下降並無法當下就決定出是否為語音還是雜訊，需要往後或是往前多看幾個 frame 來判斷當下的 frame 為語音還是雜訊；換句話說，則是由 frame n-l 到 frame n+d 的 (d-l+1)個 frame (也就是 window size of feature frame) 的語音特徵參數來預測 frame n 的 VAD 狀態，n 代表目前預估 VAD 狀態的 frame index，d 表示為從目前時間點 n 所 delay 的 frame 數目，而 l 表示為從目前時間點 n 所提前的 frame 數目。NN-based 實驗參數設定如表六所示。



圖三：Windows size of feature frames 示意圖，(上) 語音開始、(下) 語音結束

表六：NN-based VAD 實驗設定

NN 種類	**DNN**
輸入資料	NTPU Additive Noise Corpus
Windows size of feature frames (*d*+*l*+1)	1、3、5、7、9、11 -frame
實驗設定	
Optimizer	Adam
Batch_size	64
Nb_epoch	1500
Data set	Train (7) : Validation (2) : Test (1)
Earlystopping patience	50
Activation function	ReLU
Loss function	categorical crossentropy
Node size	256
Dropout	0.3
Output layer function	Softmax

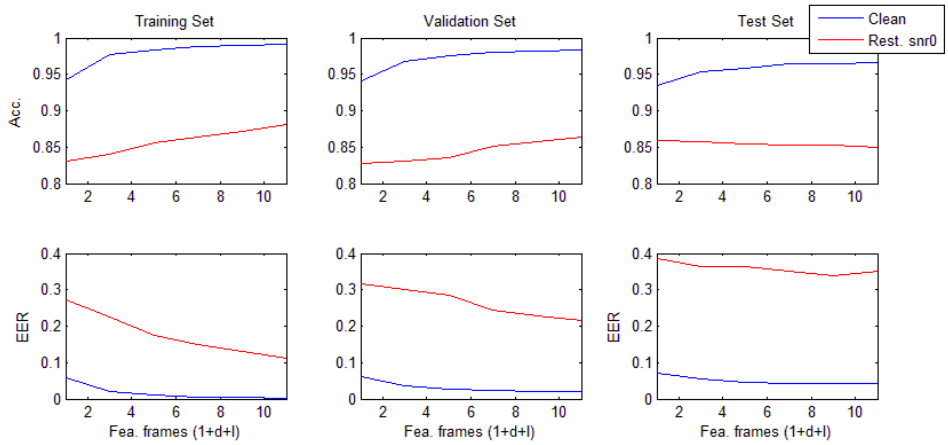
（三）、NN-based VAD 實驗結果與分析

本小節將不同種類 NN-based VAD 之 5 種問題，如: feature frames、layer 數目、matched-condition 與 multi-condition 和 delay decision 的問題進行主觀討論。NN-based 的 VAD 研究方法實做於 Tensorflow 平台上，在給定輸入以及輸出之答案後，依照不同的 NN 架構來決定每個 frame 是語音還是非語音。為了要與不同 NN 方法做比較，本研究挑選出較

具代表性之情況來探討。

1、DNN 下 feature frames 的討論

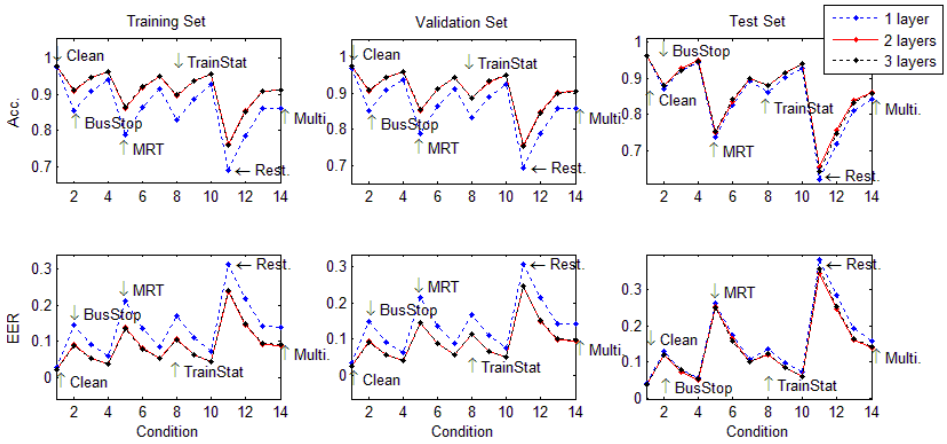
圖四表示兩極端情況下，DNN 的 feature frames 結果。橫軸為 feature frames 數目、縱軸分別是 accuracy (Acc.)以及 EER。圖四，隨著 feature frames 數目的上升其 Acc.以及 EER 在 training set 與 validation set 中的表現有明顯變好的趨勢。但是在 outside test 中其 EER 進步幅度較小。故此推論說，當 feature frames=8 時，就已 over-trained。



圖四：較具代表性情況下 DNN 之 feature frames 結果

2、DNN 下 layer 數目的討論

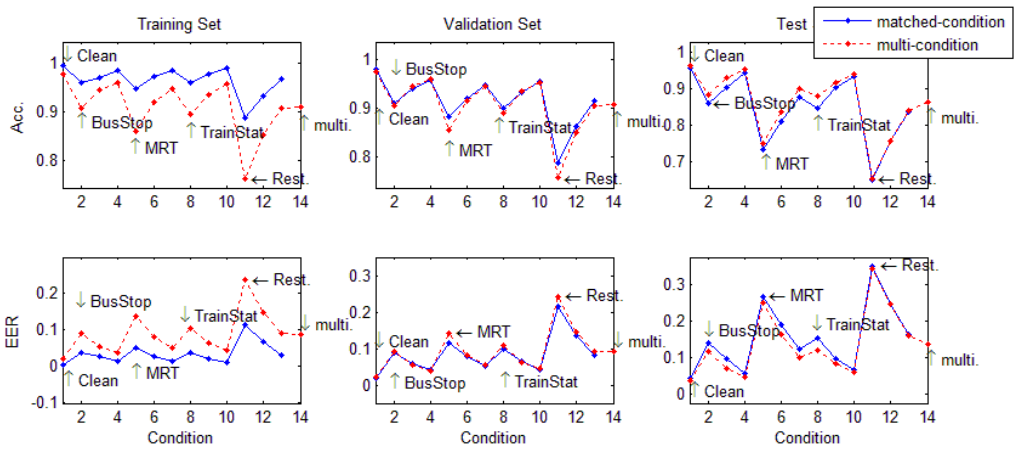
圖五表示在 multi-condition 中各個情況下，DNN 的 layer 數目結果。橫軸為每個 condition、縱軸分別是 Acc.以及 EER，其中橫軸的每個 condition 依序分別是：Clean 為乾淨、BusStop 為公車站、MRT 為捷運站、TrainStat 為火車站、Rest.為餐廳、Multi.為 multi-condition 的情況，而在 BusStop、MRT、TrainStat 與 Rest.的情況當中又包含有 snr=0, 5, 10 (dB)之結果。分析完圖五後可以得到的結論是，將每個 condition 的資料合併成 multi-condition 後，解決了資料量不足的情況。我們先從 DNN 的 layer 數目來觀察每個 layer 間彼此的關係，可以得到的結果是隨個 layer 數目的上升，其 Acc.與 EER 在 training set、validation set 與 outside test 中的表現有變好之趨勢，尤其是在各個 condition 中 snr=0 (dB)的時候。故此推論說在 hidden layer 越深時，每個 condition 可以互相學習各個 condition 間共同的特性。但是在 2 layers 與 3 layers 時的 Acc.與 EER 進步幅度較小，其原因是已 over-trained。



圖五：multi-condition 與 multi-condition 中各個情況之 DNN layer 數目結果，其中 BusStop 為公車站、MRT 為捷運站、TrainStat 為火車站、Rest.為餐廳、Multi.為 multi-condition，而在 BusStop、MRT、TrainStat、Rest.情況當中又包含有 snr=0, 5, 10 (dB)的結果

3、matched-condition 與 multi-condition 的討論

圖六表示 matched-condition 與 multi-condition 的結果，是在每個情況下挑選出最好的 layer 數目來做討論。橫軸為每個 condition、縱軸分別是 Acc.以及 EER，其中橫軸的每個 condition 依序分別是：Clean 為乾淨、BusStop 為公車站、MRT 為捷運站、TrainStat 為火車站、Rest.為餐廳、Multi.為 multi-condition 的情況，而在 BusStop、MRT、TrainStat 與 Rest.的情況當中又包含有 snr=0, 5, 10 (dB)之結果。由圖六可以得到以下之結論，在 training set 與 validation set 來觀察 Acc.與 EER 時，可以發現在最好設定下 overall 的 multi-condition 結果，是優於在 matched-condition 下最糟情況的 Rest.結果。但是從 outside test 來觀察 multi-condition 裡各個 condition 的結果後，可以發現其結果是優於 matched-condition 之結果，尤其是在 TrainStat snr=0 (dB)時進步幅度最為明顯。故此可以推論說，因為在 multi-condition 裡的 hidden layer 能夠學習到不同 condition 的特性，所以對於不同的環境跟情況下能夠更加強健(Robustness)。



圖六：matched-condition 與 multi-condition 的結果

四、結論

本論文探討將類神經網路應用於語音端點偵測中，使用智慧型手機來錄製不同種類的雜訊，並且自行混音出特定之 SNR 種類，在由不同架構的類神經網路來做學習。

經過研究與分析，本論文在不同實驗下的類神經網路結果，可以得到以下之結論是：(1)DNN 的 frames 問題會隨著 layer 數目的增加，而使 Acc.與 EER 的表現有變好、(2)DNN 的 layer 數目問題，在 matched-condition 的結果並未隨著 layer 數目的上升而使 Acc.及 EER 有變好之趨勢，故推論其原因是在於訓練的資料量不足所造成，或是有些語音中重要之特性在轉換成 MFCC 參數的過程中，就被忽略了。在 multi-condition 的結果中可以發現隨著 layer 數目的上升，其 Acc.與 EER 的表現在各個 set 中有變好的趨勢，所以可以推論其原因是隨著 hidden layer 數目越深時，每個 condition 可以互相學習各個 condition 間共同的特性、(3)matched-condition 與 multi-condition 的問題，在 multi-condition 之 performance 優於 matched-condition (MRT 與 Rest. condition)，所以由此可推論出在 multi-condition 中的 hidden layer 能夠學習到不同 condition 之特性，明顯展現了深度學習的優勢。

五、參考資料

[1] Deng, C. Z. (2007, September). *Voice Activity Detection and Keyword Spotting System on Embedded Platform,* National Chiao Tung University, Hsinchu.

[2] Schafer, R., & Rabiner, L. (1975, April). Digital Representations of Speech Signals. *IEEE, 63*(4), 662-667. doi:10.1109/PROC.1975.9799.

[3] Davis, A., Nordholm, S., & Togneri, R. (2006, February). Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold. *IEEE Signal Processing Society, 14*(2), 412-424. doi:10.1109/TSA.2005.855842.
[4] Li, X., Horaud, R., & Girin, L. (2016, October). Voice Activity Detection Based on Statistical Likelihood Ratio with Adaptive Thresholding. *IWAENC*, China.
[5] Makhoul, J., Roucos, S., & Gish, H. (1985, November). Vector Quantization in Speech Coding. *Proceedings of the IEEE, 73*(11), 1551-1558. doi:10.1109/PROC. 1985.13340.
[6] Reynolds, D. (2008). *Gaussian Mixture Models.* Springer US.
[7] Dlamini, N. S. (2015, November). *Acoustic Model Training for Speech Recognition System,* National Taipei University, New Taipei City.
[8] Shen, Z., Wei, J., & Dang J. (2016, October). Voice Activity Detection Based on Sequential Gaussian Mixture Model and with Maximum Likelihood Criterion. *ISCSLP,* China.
[9] Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* Springer, Berlin, Heidelberg.
[10] Elamn, J., L. (1990, April). Finding structure in time. *Cognitive Science,14*(2), 179-211. doi:10.1207/s15516709cog1402_1.
[11] Jordan, M., I. (1997, September). Serial order: A parallel distributed processing approach. *Advances in Psychology, 121*(1), 471-495. doi:10.1016/S0166-4115(97)80111-2.
[12] Dong, Y., & Deng, L. (2016). **解析深度學習語音識別實踐**. 電子工業出版社.
[13] 麥克風語料庫 TCC-300Edu.
[14] Vondasek, M., & Pollak, P. (2005). Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency. *Radioengineering, 14(1)*, 6-11.
[15] Wei, Z., & Lou, P. (2009, September). *A new self-adaptive audio-mix algorithm's research and realization based on voice energy*, Beijing University of Posts and Telecommunications, Beijing.
[16] Hawwa, S. (2002, August). Audio mixing for centralized conferences in a SIP environment. *ICME, 2*(1), 269-272. doi:10.1109/ICME,2002.1035572.
[17] Geoffrey, E. H., Osindero, S., & Teh, Y. W. (2006, May). A fast learning algorithm for deep belief nets. *Neural computation 18(7)*, 1527-1554. doi:10.1162/neco.2006.18.7.1527.
[18] Bengio, Y. (2006, January). Greedy layer-wise training of deep networks. *Advances in neural information processing systems, 19(NIPS'06),* 153-160.
[26] HTK Book 3.4.1.
[27] Garcia, A. L. (2009). *Probability, Statistics, and Random Process for Electrical Engineering (3rd Edition),* New Jersey: Pearson Prentice Hall.
[28] Vaseghi, S. V. (2007). *Multimedia Signal Processing, Theory and Applications in Speech, Music and Communications.* UK: John Wiley & Sons Ltd.
[29] 王小川（2012）。*語音訊號處理*。台北：全華。
[30] 謝秀琴（1996）。*數位語音訊號基本處理*。台北：全華。