The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 136-147
© The Association for Computational Linguistics and Chinese Language Processing

類神經網路訓練結合環境群集及專家混合系統於強健性語音辨識

徐家鏞 Chia-Yung Hsu[1]、王家慶 Jia-Ching Wang[1]、曹昱 Yu Tsao[2]

[1] 國立中央大學資訊工程學系
Department of Computer Science and Information Engineering, National Central University
[2] 中央研究院資訊科技創新研究中心
Research Center for Information Technology Innovation, Academia Sinica

摘要

近年來，類神經網路 (Neural Network) 在語音辨識上的研究有著豐碩的成果，有效地減少環境以及語者變異對語音訊號造成的影響，大幅提升辨識率，但系統的語音辨識能力仍有改善空間。本論文即提出新的自動語音辨識系統架構，結合 Environment Clustering (EC)、Mixture of Experts 與類神經網路以進一步提升系統效能。我們將辨識系統分為 Offline 與 Online 兩階段：Offline 階段依據聲學特性將整個訓練資料集分割成多個子訓練資料集，並建立各子訓練資料集的類神經網路(以類神經子網路稱之)。Online 階段則使用 GMM-gate 來控制類神經子網路的輸出。新提出的系統架構保留子訓練資料集的聲學特性，強健語音辨識系統。實驗上，我們使用 Aurora 2 連續數字語音資料庫，依據字錯誤率(word error rate, WER)比較我們提出的語音辨識系統架構與傳統以類神經網路建立的辨識系統，平均字錯誤率進步 5.9% ，由 5.25%降低至 4.94%。

Abstract

Recently, automatic speech recognition (ASR) using neural network (NN) based acoustic model (AM) has achieved significant improvements. However, the mismatch (including speaker and speaking environment) of training and testing conditions still confines the applicability of ASR. This paper proposes a novel approach that combines the environment clustering (EC) and mixture of experts (MOE) algorithms (thus the proposed approach is termed EC-MOE) to enhance the robustness of ASR against mismatches. In the offline phase, we split the entire training set into several subsets, with each subset characterizing a specific speaker and speaking environment. Then, we use each subset of training data to prepare an NN-based AM. In the online phase, we use a Gaussian mixture model (GMM)-gate to determine the optimal output from the multiple NN-based AMs to render the final recognition results. We evaluated the proposed EC-MOE approach on the Aurora 2 continuous digital speech recognition task. Comparing to the baseline system, where only a single NN-based AM is used for recognition, the proposed approach achieves a clear word error rate (WER) reduction of 5.9 % (5.25% to 4.94%).

關鍵詞：Neural Network，強健性語音辨識，環境群集，專家混合系統

Keywords: Neural Network, Robust Speech Recognition, Environment Clustering, and Mixture of Experts.

一、簡介

雖然語音辨識系統在安靜環境下可以達到不錯的辨識率，但是在實際應用上，由於環境噪音(environment noise)產生的加成性雜訊(additive noise)及通道失真(channel distortion)產生的卷積性雜訊(convolutive noise)等情況，造成訓練及測試語料的環境不匹配問題，限制語音辨識系統的效能。

欲解決上述的不匹配問題，在模型空間(model space)的處理中有許多模型調適(model adaptation)的方式，例如最大後驗機率估計(maximum a posteriori estimation)[1]、最大似然線性迴歸(maximum likelihood linear regression)[2]、最小分類錯誤線性回歸(minimum classification error linear regression)[3]等等。

在強健性語音辨識上已經有許多使用研究使用類神經網路，例如，在環境不匹配的情況下使用線性轉換強健模型[4][5]；結合 GMM-HMM 與 DNN-HMM 進行輸出[6]；類神經網路產生分別為目標訊號與干擾訊號的兩個輸出，使用分離的結果進行辨識[7]等等許多方式。在這些相關研究中，都是使用同一個類神經網路來處理所有環境的情況。在整體學習(ensemble learning)的相關研究中，有使用 bagging[8]或是 boosting[9]等等方式，這裡我們使用基於 Environment Clustering (EC)[10]及 Mixture of Experts[11]的架構來訓練多個類神經網路，並在最後選擇一個適當的類神經網路進行輸出。

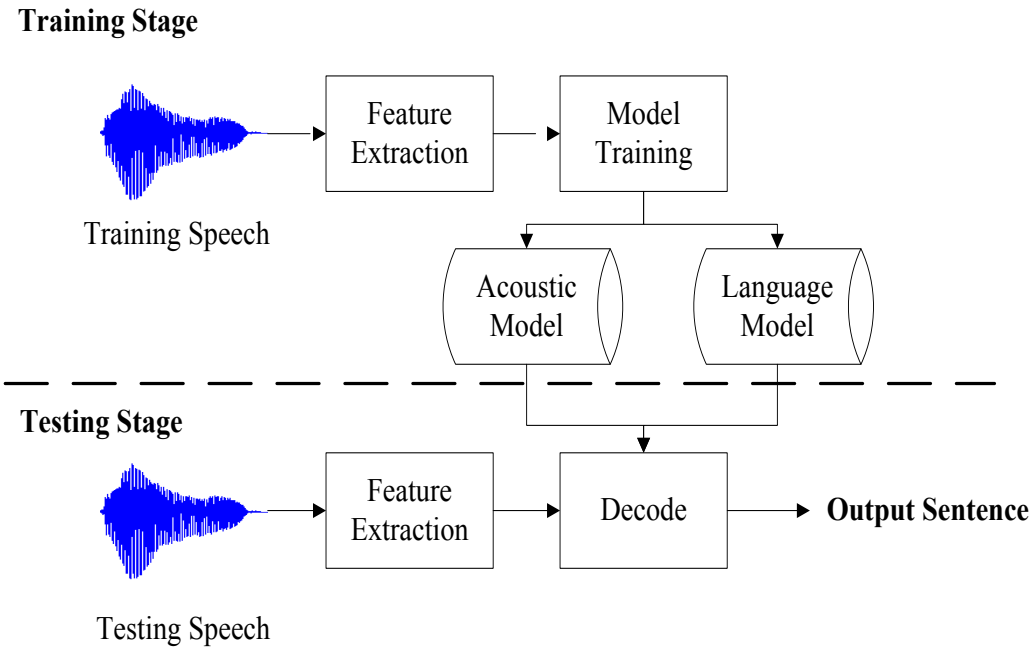
在接下來的內容，第二章將介紹整個語音辨識的主要流程以及一些相關的研究方法。第三章將介紹本篇論文的系統架構。第四章為實驗的部分，此章節包含介紹實驗語料與實驗設定、baseline 系統以及本論文系統的 Word Error Rate (WER)。第五章為此研究的結論。

二、語音辨識流程及相關研究方法介紹

在此章節中我們將簡單介紹基本的語音辨識流程，及辨識中所使用的高斯混合模型(Gaussian Mixture Model，GMM) 與類神經網路(Neural Network)。

(一)、語音辨識流程

圖一為一個基本的語音辨識流程，可分為訓練及測試階段。首先擷取語音訊號的特徵(feature extraction)，如梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC)；接著利用擷取的語音特徵在訓練階段訓練模型(model training)，或在測試階段解碼(decode)為文字。訓練階段將產生聲學模型(acoustic model)及語言模型(language model)，並供給測試階段解碼使用。此外，目前訓練聲學模型的方式主要為 GMM 與類神經網路，將於下一節介紹。



圖一、語者辨識流程圖

(二)、高斯混合模型(Gaussian Mixture Model, GMM)

高斯混和模型是用來模擬複雜資料分布的機率模型。一個高斯混合模型為 K 個高斯機率密度函數的加權總合，如式(1)。

$$p(x \mid \phi) = \sum_{k=1}^{K} \lambda_k N(x \mid \mu_k, \Sigma_k) \qquad (1)$$

其中

$$N(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{\frac{-1}{2}(x-\mu_k)^T (\Sigma_k)^{-1}(x-\mu_k)} \qquad (2)$$

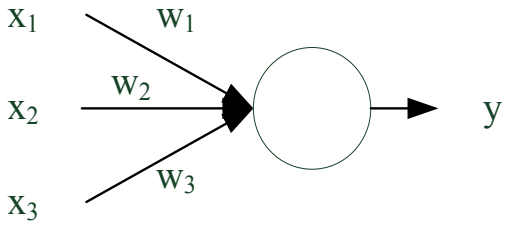
另外，$\phi = (u_k, \Sigma_k)_{k=1}^{K}$ 為各個高斯的參數，u_k 及 Σ_k 分別為第 k 個高斯成分(Component)的平均(mean)及共變異矩陣(covariance matrix)。λ_k 為第 k 個高斯成分的先驗機率，並且滿足：

$$\sum_{k=1}^{K} \lambda_k = 1 \quad and \quad \lambda_k \geq 0 \qquad (3)$$

高斯混合模型的參數，可以使用 EM 演算法(Expectation-Maximization algorithm)，經過 Expectation 步驟及 Maximization 步驟的疊代來進行模型參數的估計。

(三)、類神經網路

類神經網路 (Neural Network，NN)是一種模擬生物大腦的機器學習(machine learning)模型。構成一個類神經網路的基本元素為神經元(neuron)，如圖二所示。一個神經元的結構，是由多個輸入經過線性組合，並經過激發函數(activation function)後產生輸出 y 。



圖二、神經元示意圖

可由式(4)表示：

$$y = f(\sum_i x_i w_i + b) \tag{4}$$

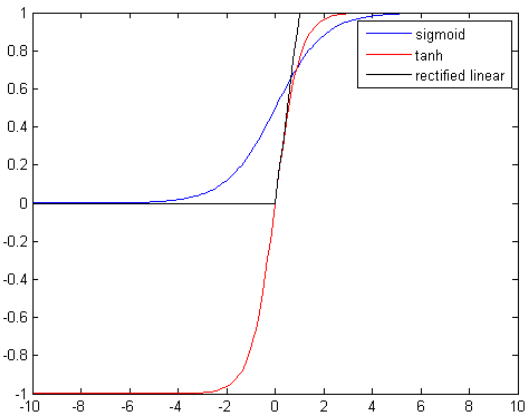
其中 $\{x_i \mid i = 1, 2, \ldots, n\}$ $\{x_i | i = 1,2,\cdots N\}$ 為輸入資料、$\{w_i \mid i = 1, 2, \ldots, n\}$ 為權重值(weight)，代表由資料 x_i 進入神經元的權重；b 為偏移量(bias)，最後，$f(\bullet)$ 為激發函數。

常見的激發函數有：

$$\text{Sigmoid：}\ f(x) = \frac{1}{1+e^{-x}} \tag{5}$$

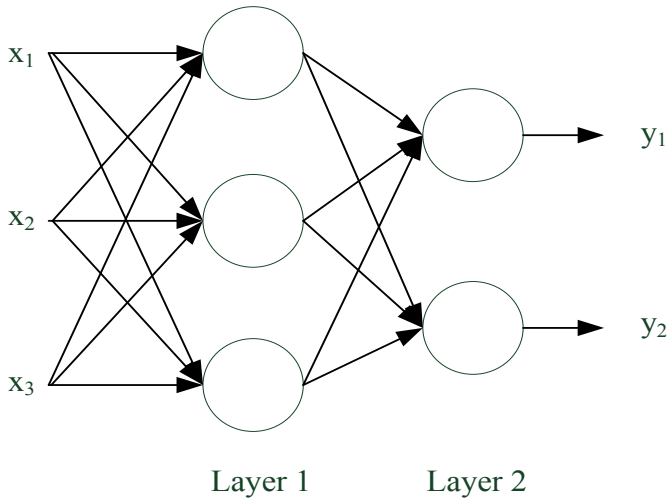
$$\text{Tanh：}\ f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{6}$$

$$\text{Rectified Linear：}\ f(x) = \max(0, x) \tag{7}$$



圖三、Sigmoid、Tanh 及 Rectified Linear

此外，一個完整的類神經網路為多個神經元架構而成，如圖五為雙隱藏層（hidden layer）的類神經網路，總共由五個神經元組成(第一層有三個神經元節點，第二層則為兩個神經元節點)。資料輸入至第一層的神經元的，而第二層的輸入則為第一層的輸出。其中的參數 $\{w_i \mid i = 1, 2, \ldots, n\}$ 與 b 可由倒傳遞(back propagation)訓練而得；詳細的網路訓練流程可參考[12]。



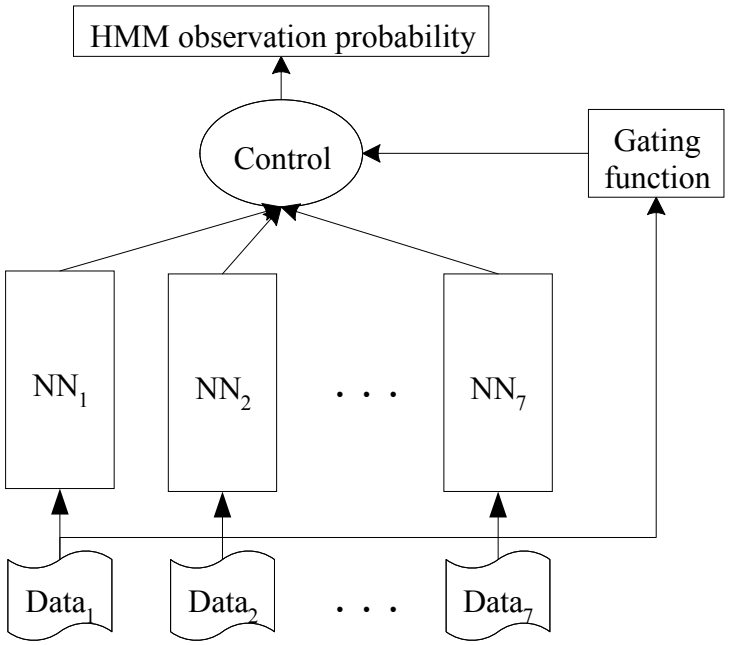
圖四、雙隱藏層類神經網路示意圖

三、本論文系統架構

同一句語音訊號在不同的語者、環境等等情況表現的聲學特性不盡相同，因此可依據不同的聲學分類方式，例如性別、訊噪比(signal-to-noise ratio，SNR)等等，將一份訓練語料庫分割成數種不同的子集，並以類神經網路與 GMM 模型化每一個子集所代表的聲學特性。測試時，首先以 GMM 模型決定測試語料的類別，再依據其結果，選擇相對應的類神經網路模型，最後得到較具代表性的語音特性輸出，進而增進辨識效果。

我們將系統分成 online 與 offline 階段，圖五則為一 online 的流程圖。offline 階段依據不同聲學特性的資料集，各別訓練出對應的類神經網路 $\{\mathrm{NN}_1, \mathrm{NN}_2, \ldots, \mathrm{NN}_n\}$（以類神經子網路稱之），並供給 online 階段使用。另外，在 online 階段使用一個 gating function 來選擇類神經子網路的輸出，並得到最後的辨識結果。最後，我們選擇 GMM 做為 gating function，以 GMM-gate 稱之。

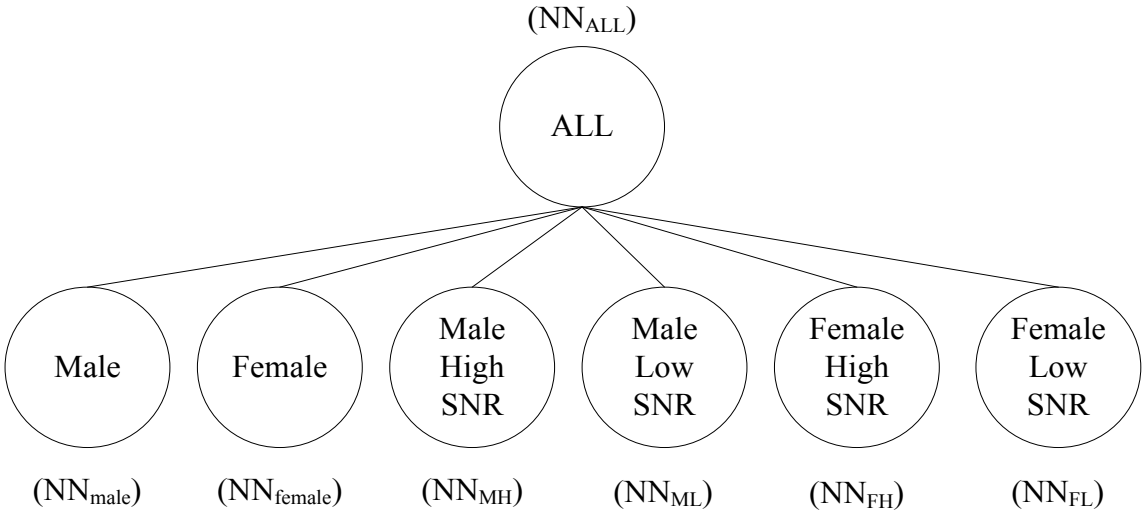
我們提出基於環境群集(Environment Clustering，EC)[10]以及 mixture of local experts[11]的多類神經子網路訓練及結合各子網路輸出之架構，下一節將介紹 offline 的系統建構流程及 online 的測試流程。



圖五、本論文系統架構示意圖

(一)、Offline 系統建構

在 offline 系統中，我們將訓練資料集依據性別以及訊噪比分成六個子訓練資料集：男性、女性、男性高 SNR、男性低 SNR、女性高 SNR 以及女性低 SNR；如圖六所示：



圖六、EC 樹架構

其中，類神經子網路的訓練，首先以訓練資料集訓練出 global 的 NN_{ALL}，接著依據不同

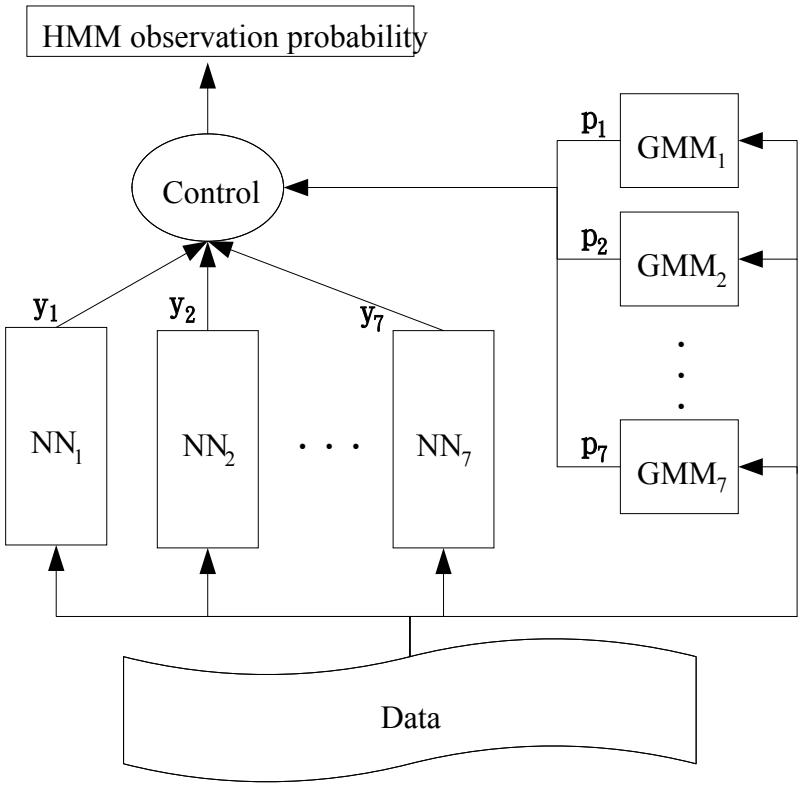
的聲學特性分類訓練資料夾為六個子訓練資料集，並分別對 NN_{ALL} 進行倒傳遞調整網路參數，得到六個類神經子網路 NN_{male} 、 NN_{female} 、 NN_{FH} 、 NN_{FL} 、 NN_{MH} 以及 NN_{ML} 。另外，GMM 模型的訓練，首先以訓練資料集依據式(1)，使用 EM 演算法訓練 UBM[13] 模型， GMM_{ALL} ，接著對每一種子訓練資料集以 MAP(Maximum a Posteriori) estimation 調適(adaptation)出六種子集 GMM 模型： GMM_{male} 、 GMM_{female} 、 GMM_{FH} 、 GMM_{FL} 、 GMM_{MH} 與 GMM_{ML} 。

(二)、Online 系統建構

前一小節得到的整體模型及六個子集模型，將提供給 online 階段使用。如圖七所示，在 online 階段時，我們將整句測試資料利用式(1)，分別計算各子集 GMM 模型的七個平均後驗機率，得到七個平均後驗機率 $p_1, p_2, \ldots, p_7$ ，並決定其中的最大值與相對應的第 i 個子集，其中 i 為：

$$i = \arg\max_{k=1,2,\ldots,7} p_k \qquad (8)$$

最後，再由第 i 個子集對應的類神經網路的輸出作為 HMM 的觀測機率。



圖七、Online 階段架構圖

四、實驗與結果

在本節，我們將介紹實驗的設定、並分析比較傳統利用類神經網路模型的辨識系統以及本文提出的強健性語音辨識系統的結果。

(一)、實驗語音資料與實驗設定

語音辨識的實驗，我們使用 Kaldi 這套用於語音辨識的開放原始碼工具[14] ，並做為我們的 baseline NN-HMM 系統；並以 Aurora 2 資料庫[15] 做為本實驗的語料庫。Aurora 2 為一個英文連續數字語音的資料庫，包含八種不同的加成性雜訊環境(Subway, Babble, Car, Exhibition, Airport, Street, Train Station, Restaurant)、兩種不同的通道雜訊(G712 and MIRS) 與七種不同的 SNR (clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB)。語料庫中，含有雜訊的語音為人工添加不同的雜訊環境與 SNR 到乾淨語音上。另外，Aurora 2 語料庫包含訓練與測試的語料集：訓練語料庫包含 clean-與 multi-condition 兩種訓練語料庫，本實驗使用 multi-condition 訓練語料庫。該語料庫包含四種噪音類型 (Subway, Babble, Car, Exhibition) 與五種 SNR (clean, 20 dB, 15 dB, 10 dB, 5 dB)，一共有 8440 句，總長度約為四個小時；測試語料集則分成三個子集 Set A、Set B 及 Set C，各測試子集中皆有不同的 SNR 環境，從 20 dB 至-5 dB 與 clean。Set A 包含與訓練語料相同的四種噪音，Set B 則為包含 Restaurant, Street, Airport 與 Train Station 的環境雜訊，Set C 為兩種噪音 (Subway, Street) 加上通道失真。

我們使用歐洲電信標準化協會 (European Telecommunications Standards Institute, ETSI) 所提出用於進行分散式語音辨識的 AFE (Advanced Front-End)，做為實驗用的特徵。音框長度為 25 毫秒，音框移動長度為 10 毫秒。神經網路的訓練使用 13 維 AFE 加上其一階及二階動態特徵，並前後串接 5 個音框，輸入向量共 429 維。HMM 我們定義靜音為 3 個狀態，數字的聲音為 16 個狀態，共有 179 個狀態。

在實驗中，類神經子網路我們使用 1 層隱藏層，一層有 2560 個神經元。訓練使用 dropout[16]以避免 overfitting。此外，dropout rate 為 0.8；詳細的實驗設定可參考[17]。

(二)、評估方法

實驗結果的評估方面，我們使用字錯誤率(Word Error Rate, WER)來評估實驗結果，其計算方式如下式：

$$WER = \frac{S + D + I}{N} \times 100\% \quad (9)$$

在字串比對中，兩個字串可能會發生插入(Insertion)、刪除(Deletion)以及替換(Substitution)。在(9)式中，S 為替代字數、D 為刪除字數、I 為替換字數、N 為總字數。由式(9)，給定辨識結果字串，我們可以計算出相對應的字錯誤率。

(三)、辨識結果

表一及表二為 baseline 系統在不同層數下的結果。表一列出三個子測試集的平均詞錯誤率，與整體的平均詞錯誤率的結果；表二則列出不同 SNR 下，平均詞錯誤率的實驗結果。從表一及表二的實驗結果可以看出，使用三層類神經網路在 sets B 與 C 測試集、平均的詞錯誤率與不同的 SNR 環境中，有最差的辨識結果；而使用一層類神經網路，則在各種測試集合中有最佳的辨識效能，這可能是因為 Aurora 2 的訓練語料不足以訓練多層的類神經網路。在經過 512、1024、1536、2048、2560 及 3072 個神經元的實驗後，2560 個神經元獲得最好的辨識結果，我們因此將一層神經網路的設定當作我們的 baseline 系統。

表一、Baseline 類神經網路各層之辨識結果

	Set A	Set B	Set C	Avg.
1	**4.65**	**5.83**	**5.28**	**5.25**
2	4.78	6.95	5.76	5.85
3	4.90	7.26	6.08	6.08
4	4.98	6.61	5.80	5.79

表二、Baseline 類神經網路各層於各種 SNR 下之辨識結果

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
1	**0.72**	**0.69**	**1.03**	**2.16**	**5.50**	**16.87**	**47.28**
2	1.24	0.81	1.25	2.55	6.37	18.24	48.53
3	1.48	0.87	1.39	2.66	6.57	18.90	49.67
4	1.18	0.79	1.19	2.40	6.10	18.49	49.53

我們一開始進行了將整體模型及子集模型使用線性組合的方式組合類神經網路輸出，我們求取一組組合係數 $\mathbf{w}$，使得類神經網路的輸出乘上這組係數會與期望輸出的差距越小越好，其目標式如下：

$$\arg\min_{w} \sum_{i} (\mathbf{t}_i - X_i\mathbf{w})^T (\mathbf{t}_i - X_i\mathbf{w}) \quad (10)$$

其中 X_i 為第 i 筆資料經過全部 7 個類神經網路的輸出、$\mathbf{t}_i$ 為第 i 筆資料經過正確類神經網路的輸出、$\mathbf{w}$ 為欲求得的組合係數。對 $\mathbf{w}$ 求解並加入一般化項可寫為：

$$\mathbf{w} = (\sum_{i} X_i^T X_i + \delta I)^{-1} (\sum_{i} X_i^T \mathbf{t}_i) \quad (11)$$

則我們可以使用 $\mathbf{w}$ 線性組合整體模型及子集模型的輸出，其辨識結果如表三。從結果可以看出其效果明顯低於 baseline 系統，我們推測原因為對於每筆測試資料都使用同一組加權值進行組合，沒有考慮到每筆測試資料的獨特性，整個系統只會得到對於各類型資

料平均的效果。

表三、線性組合法與 baseline 比較

	Set A	Set B	Set C	Avg.
Baseline	**4.65**	5.83	**5.28**	**5.25**
Linear Combination	4.78	**5.81**	5.48	5.33

在進行語音辨識的實驗前，我們首先測試使用 GMM 來進行模型選擇的能力。在表四的實驗中，分別為 GMM 分別有 64 個與 128 個高斯成分的性別辨識錯誤率。由結果可以看出使用 128 個高斯成分的錯誤率較低，而且也有著不錯的辨識率，因此在後面的實驗中，我們使用 128 個高斯成分的 GMM 來進行類神經網路的選擇。

表四、GMM 性別辨識之結果

	GMM components	Test Error Rate
GMM	64	7.8
GMM	128	7.3

表五比較本文提出的強健性語音辨識系統與 baseline 辨識系統的系統辨識效能，在三個測試子集中。可以看出在三個測試子集的部分，本文提出的辨識系統，詞錯誤率相較於 baseline 都有明顯的下降，平均的詞錯誤率則降低了 5.9% (從 5.25 到 4.94)，我們相信此辨識結果支持依據聲學結性切割訓練語料庫，並在測試中選擇較佳的聲學模型做為輸出，即能適當的提升語音辨識系統的效能並強健語音辨識系統。

表五、本文方法與 baseline 比較

	Set A	Set B	Set C	Avg.
Baseline	4.65	5.83	5.28	5.25
Proposed method	**4.39**	**5.41**	**5.10**	**4.94**

五、結論

在此篇論文中，我們提出基於 EC 及 Mixture of Experts 的架構來訓練神經網路；依據訓練語料不同的聲學特性，切割並以類神經網路與 GMM 模型化不同的聲學模型；在測試時，將測試語料經由 GMM-gate 得到對每個聲學模型的後驗機率，選擇最佳的聲學模型做為辨識系統的基礎。實驗上，我們以 Aurora 2 做為實驗的語料庫，將訓練語料依據性別以及 SNR 的方式切割訓練語料，並比較了傳統使用 DNN-HMM 架構與本文提出的強健性語音辨識系統。我們提出的語音辨識系統能提升傳統的語音辨識系統達 5.9%。未來我們將探討不同的聲學特性模型與不同的 gate function，並嘗試在大詞彙語料庫中。

參考文獻

[1] J. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, Apr. 1994.

[2] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol. 12, no. 2, pp. 75-98, Apr. 1998.

[3] X. He and C. Wu, “Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs,” International Conference on Multimedia and Expo, vol. 1, pp. 397-400, July 2003.

[4] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals and T. Robinson, “Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system,” In Proceedings of Eurospeech, pp. 18-21, Sep. 1995.

[5] B. Li and K. C. Sim, “Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems,” In Proceedings of Interspeech, pp. 526-529, 2010.

[6] B. Li and K. C. Sim, “On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 195-199, May 2014.

[7] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, “Deep neural network based speech separation for robust speech recognition,” in International Symposium on Chinese Spoken Language Processing, pp.532-536, Oct. 2014.

[8] L. Breiman, “Bagging predictors,” Journal of Machine Learning, vol. 24, no. 2, pp. 123-140, Aug. 1996.

[9] R. E. Schapire, “The strength of weak learnability,” Journal of Machine Learning, vol. 5, no. 2, pp. 197-227, Jun. 1990.

[10] Y. Tsao, X. Lu, P. Dixon, T.-y. Hu, S. Matsuda, and C. Hori, "Incorporating local information of the acoustic environments to MAP-based feature compensation and acoustic model adaptation," Computer Speech and Language, vol. 28, no. 3, pp. 709-726, May 2014.

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, “Adaptive mixtures of local experts,” Neural Computation, vol. 3, no. 1, pp. 79-87, Spring 1991.

[12] Y. Bengio, “Learning deep architectures for AI,” Foundation and Trends in Machine Learning, vol. 2, pp. 1-127, 2009.

[13] D. Povey, S. M. Chu, B. Varadarajan, “Universal background model based speech recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4561-4564, Mar. 2008.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech

recognition toolkit," IEEE Workshop on Automatic Speech Recognition and Understanding, Dec. 2011.

[15] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in ASR2000 Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop, Sep. 2000.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, Jan. 2014.

[17] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," IEEE Transactions on Audio, Speech and Language Processing, vol. 22, pp. 1296-1305, Aug. 2014.