

以軟體為基礎建構語音增強系統使用者介面

Development of a software-based User-Interface of Speech Enhancement System

王韜維 Tao-Wei Wang 曹昱 Yu Tsao

中央研究院資訊創新科技研究中心

Research Center for Information Technology Innovation

Academia Sinica

賴穎暉 Ying-Hui Lai

國立陽明大學生物醫學工程學系

Department of BioMedical Engineering

National Yang-Ming University

吳家隆 Chia-Lung Wu 許祥平 Hsiang-Ping Hsu

法務部調查局

Investigation Bureau, Ministry of Justice

摘要

本研究的目的是在發展以軟體為基準的語音增強系統使用者介面，提供使用者一個快速且便於操作的輔助工具。此使用者平台包含傳統的方法和基於機器學習發展的語音增強演算法，使用者可以針對不同的噪音類型選擇演算法。處理後的語音除了可從介面上取得語音波形圖與聲譜圖的結果，還可播放與儲存處理後的語音。本研究選用 **TMHINT** 混車噪音與嬰兒哭聲做為驗證的測試語料。從結果與 **PESQ** 顯示，傳統的方式可有效的降低車噪音，但無法有效的降低嬰兒哭聲；而機器學習方式除了可降低車噪音，也可以有效的降低嬰兒噪音。

Abstract

The topic is to develop a user interface of speech enhancement in this study. This system includes of typical and based-on machine learning algorithm and provides a convenient and user-friendly interface. User can obtain waveform and spectrogram of enhancement speech

and play and restore the enhancement speech in this interface. TMHINT database with car noise or baby crying is used to test this noise reduction system. The results show that typical methods are only capable to reduce car noise, but the methods based-on machine learning could reduce both of these two noise.

關鍵詞：語音增強、NMF、DDAE、機器學習、使用者介面

Keywords: Speech enhancement , NMF, DDAE, machine learning, user interface.

一、緒論

語音增強技術為各項語音訊號技術之重要前處理單元，針對收集到的聲音訊號抑制環境噪音來增強訊號的品質，進而提升各項應用的效能。然而不同的語音增強技術有不同的應用，傳統的語音增強演算法適合處理穩態的環境噪音，譬如車聲、工廠聲等能量集中在某些頻率的噪音，但非穩態的噪音如鳴笛聲、人聲、風切聲等則是以機器學習所發展的演算法較為有效，因此針對不同環境或使用需求應使用不同的處理方式。本研究預計建立使用者介面，其使用者介面包含傳統與機器學習的語音增強方法；期望一套簡單的使用者介面可以提供使用者以較便利的方式進行多種語音增強方法模擬與比較，輔助使用者快速的選擇適合的語音增強方法。

二、理論

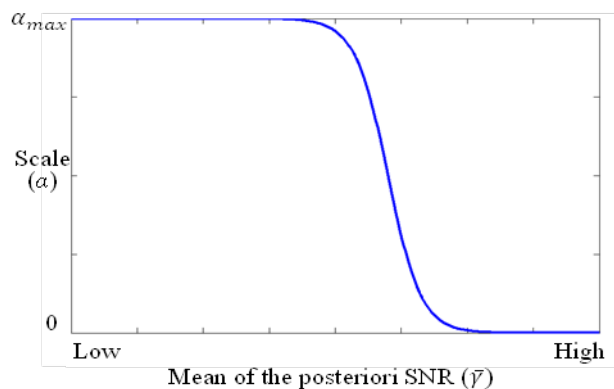
傳統語音增強方法

傳統的語音增強系統主要目的為消除背景噪音及降低增強後語音訊號的失真。大多數的語音增強技術在頻域上強化聲音訊號，通常由兩個子系統結合而成，分別是雜訊與增益值估測系統[1]。首先藉由短時傅利葉轉換將訊號進行分頻的處理，並取得帶噪語音頻譜上的振幅與相位。保留相位成份，雜訊與增益值估測系統強化振幅成份，最後經反短時傅利葉轉換將其重組為時域上較為乾淨的聲音訊號。傳統的方法中，較常見的語音增強方式包了韋納濾波器[2][3]、頻譜刪減法[4]、最小化均方誤差估測 [5]和最大事後頻譜振幅預估器的語音增強演算法（Generalized Maximum A Posteriori spectral Amplitude, GMAPA）[6]等。

韋納濾波器、頻譜刪減法等典型的語音增強方法已廣被應用，過去有許多研究改

良這些典型的語音增強方法[3]，提升降噪的能力。而最大事後頻譜振幅預估器的語音增強演算法（GMAPA）是本研究團隊過去的研究成果，結合 MLSA [7] 和 MAPA [8] 兩種噪音預估模型的演算法。MLSA 和 MAPA 兩種噪音預估模型各有優缺點，MLSA 模型對於訊號刪減幅度較低，因此在訊號品質較佳的環境可以保留較多的語音資訊，但在訊號品質較差的狀況對於雜訊消除能力較低；而 MAPA 模型對於雜訊的刪除能力較好，但對於高訊雜比的訊號容易過度刪減，造成較大的訊號失真。GMAPA 演算法使用動態調整事前機率比例的機制，在較高訊雜比的條件下，GMAPA 採用較小的事前機率比例，以防止過度語音失真。另一方面，在較低訊雜比的條件下，GMAPA 使用較大的事前機率比例，以提升增強後語音訊號的訊雜比。此外，根據語音訊號的訊雜比（SNR），我們設計一個映射函數（如圖一）來決定最佳的事前機率比例。

$$G_{GMAPA} = \frac{\xi[m,l] + \sqrt{\xi^2[m,l] + (2\alpha - 1)(\alpha + \xi[m,l])\xi[m,l]/\gamma[m,l]}}{2(\alpha + \xi[m,l])}$$



圖一 GMAPA 映射函數示意圖。

機器學習語音增強方法

傳統的語音增強是針對訊號進行噪音與訊號的預估，但對於非穩態的噪音(例如:警笛聲、人聲等)降噪效果較差。近幾年快速發展的機器學習中，DDAE (deep denoising auto-encoder) [9][10] 和 NMF (non-negative matrix factorization) [11-13]兩種技術也常被應用於訊號增強的領域中。非負矩陣分解技術藉由基底矩陣 W 與編碼矩陣 H 相乘以近似輸入頻譜 V ，如式(1)：

$$V \approx WH, \quad (1)$$

基底矩陣 W 的維度為 $F \times R$ 、編碼矩陣 H 的維度為 $R \times T$ 與輸入頻譜 V 的維度為 $F \times T$ ；且

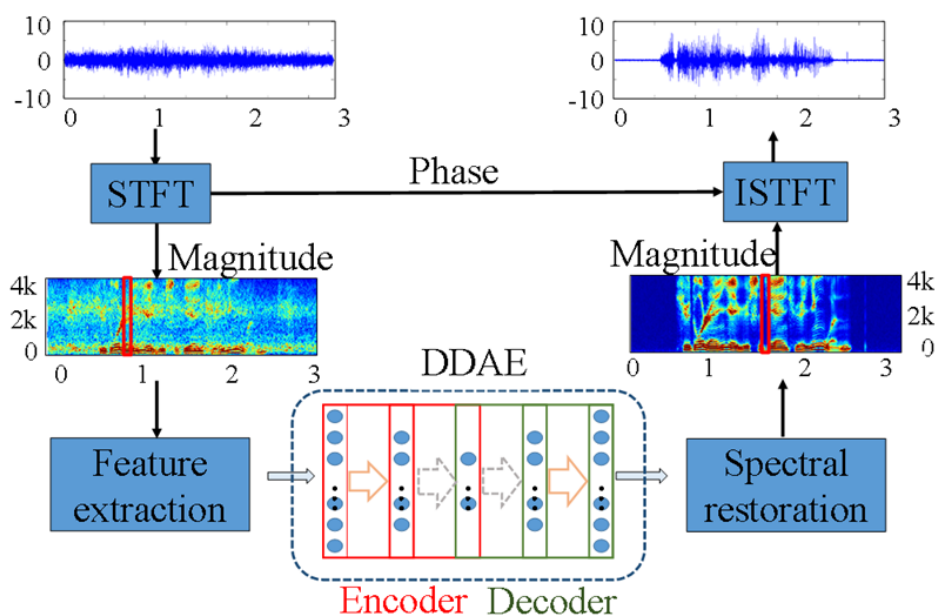
W 、 H 與 V 中所有的元素皆為正實數。在訓練階段，分別使用乾淨語音與雜訊做為訓練語料，取得代表語音與雜訊的基底矩陣 W_S 與 W_N 。在測試階段，帶噪聲音頻譜 Y 藉由 W_S 與 W_N 取得編碼矩陣 H ，如式(2)所示。

$$Y \approx WH = [W_S \ W_N][H_S \ H_N]^T \quad (2)$$

最後，增強後的聲音頻譜可由式(3)求得。

$$S' = \frac{W_S H_S}{W_S H_S + W_N H_N} \times Y, \quad (3)$$

隨著機器學習(machine learning)的進展，語音增強的效能已經有大幅的提昇，在眾多機器學習理論中，又以深層學習理論(deep learning)最為受到矚目。相較於傳統的機器學習理論，深層學習理論利用多層式結構架構出一個非線性且複雜的模型，對於多項標準的訊號處理、模式識別等測試項目，已有諸多優異的研究成果與表現，甚至是這些領域最先進的技術。圖二為流程圖，基於深層學習理論的深層去噪自編碼模型應用於語音增強技術。



圖二 應用深層去噪自編碼模型於語音增強。

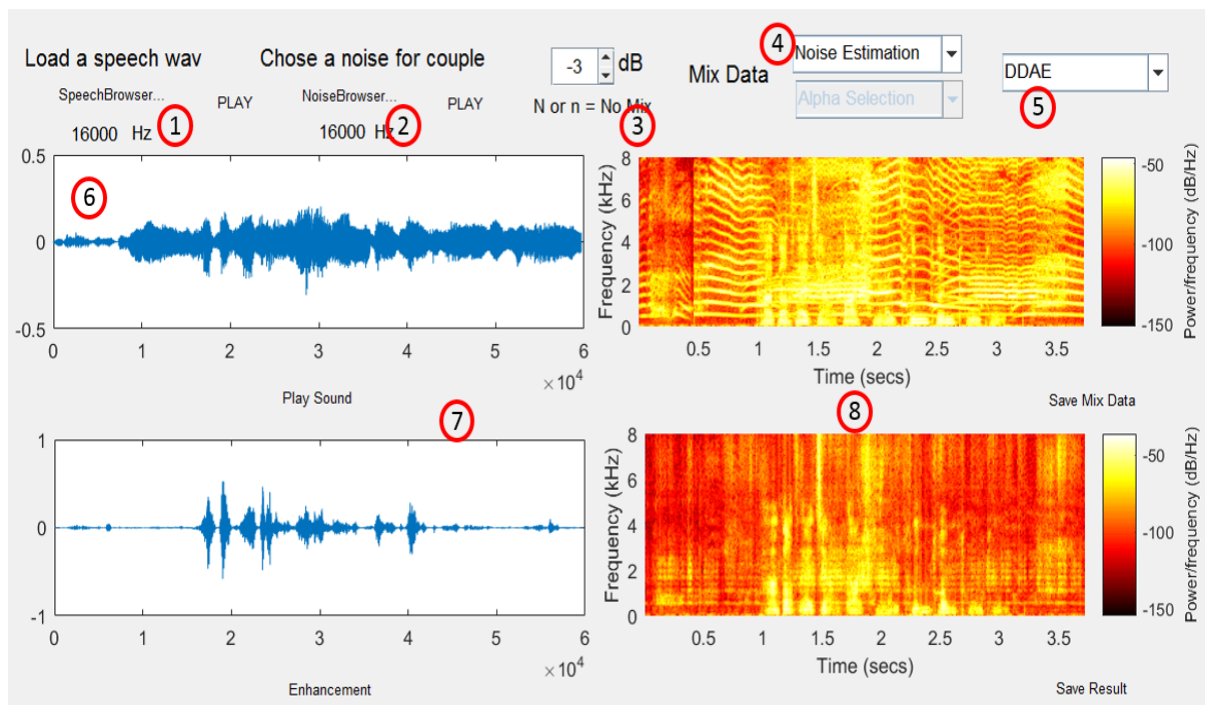
如圖二所示，類似於傳統語音增強技術，帶噪語音訊號首先經由短時傅利葉轉換拆解時域訊號至其頻譜成份並保留相位資訊。振幅資訊輸入深層去噪自編碼模型降噪，獲得較為乾淨的振幅頻譜，最後由反短時傅利葉轉換將較為乾淨的振幅頻譜與相位重建為

時域訊號。

由雜訊語音估測乾淨語音可視為一個函數近似的問題，該函數則用來描述雜訊語音與乾淨語音之間的映射關係。傳統增強技術使用線性映射函數，但由於近來深度學習在訊號處理及物體識別的發展，以類神經網路等複雜的非線性映射函數為基礎的降噪法受到關注。使用深層去噪自編碼模型的優點是易於擴展為不同地類神經網路架構來改善其表現。但不論是基於深層去噪自編碼模型或類神經網路模型降噪法，學習雜訊語音及乾淨語音之間的轉換函數都是基於收集大量的乾淨及雜訊語音資料訓練而得。

三、語音增強使用者介面

此研究所開發的語音增強介面包含上述的傳統方法與機器學習降噪方法，介面如圖三，使用者可在圖三①②輸入欲處理的語音與加成的噪音，在③輸入合成的 SNR 進行混噪，若不混噪可輸入‘N’，混噪後的語音波形圖與聲譜圖將顯示在⑥。在輸入語音與噪音時，介面會顯示輸入訊號的頻率資訊，若兩者的頻率不同則無法進行混噪處理。降噪的方法分為兩部分，在④選擇噪音估測模型並於⑤選擇語音增強方法。處理後的結果如圖三，在⑦與⑧分別顯示時域語音波形圖與聲譜資訊。此介面也提供了 NMF 與 DDAE 模型訓練功能，使用者可以自由設定參數和選擇訓練資料即可生成模型。

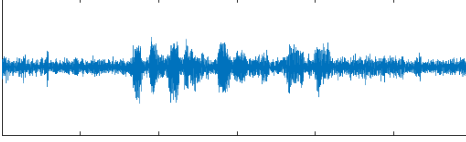
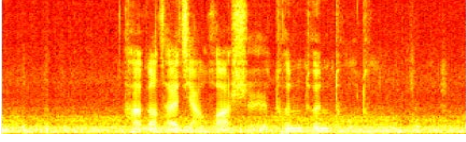
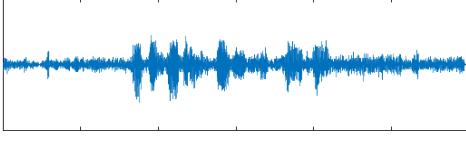
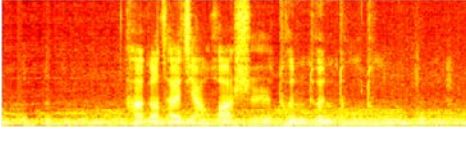
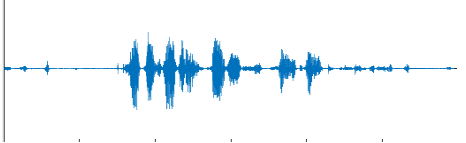
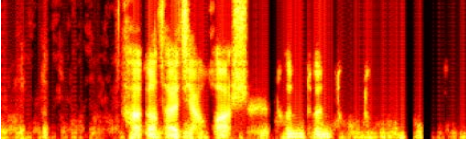
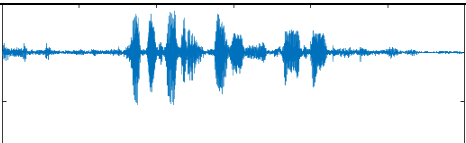
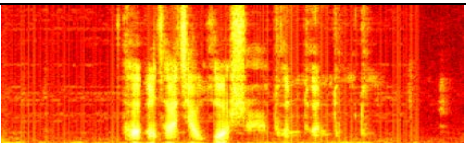
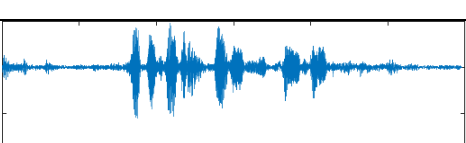
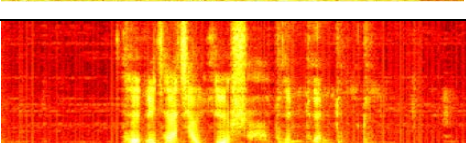
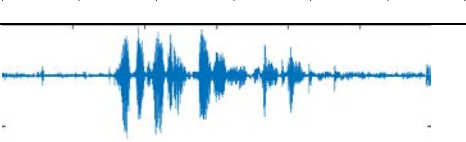
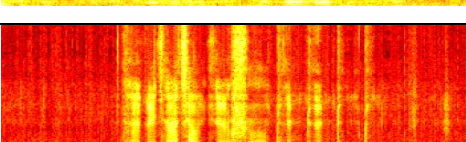


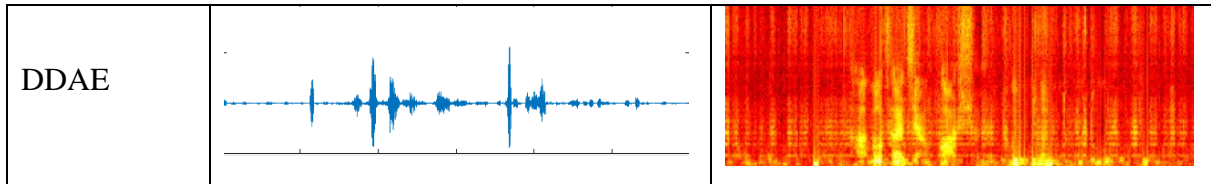
圖三 使用者介面示意圖。

四、語音增強方法比較

為了驗證此介面的正確性，我們使用 TMHINT 的語料為，以 -5dB、0dB、5dB 的 SNR 值分別混上車噪音與嬰兒哭聲噪音做為測試語料，透過本研究所建立的語音增強介面進行處理。在 NMF 方法的維度設定為 300、100 迴圈數，訓練階段選用 320 句乾淨語音與欲處理的噪音訓練基底矩陣；DDAE 的模型架構為 [400 200 100 200 400] 的五層模型，訓練資料為 7dB、3dB、-3dB、-7dB 的混噪語音-乾淨語音對，訓練的噪音選擇為車噪音和嬰兒哭聲。本研究的目的是在於開發使用者介面，因此在機器學習方法的測試語料與訓練語料相同，其結果以語音波形圖和聲譜圖(0dB)呈現於表一與表三，用 PESQ 當作客觀評量指標，顯示於表二和表四。表一、表二為車噪音而表三、表四為嬰兒哭聲。從表一的結果可發現傳統與機器學習方法都能有效抑制車噪音。我們選用 PESQ 做為客觀評量(如表二)，從客觀評量的結果可以觀察出傳統方法與機器學習方法在 PESQ 上的表現無顯著的差異。

表一 各語音增強方法對車噪音的處理成效。

處理前語音		
韋納濾波器		
KLT		
MMSE		
GMAPA		
NMF		

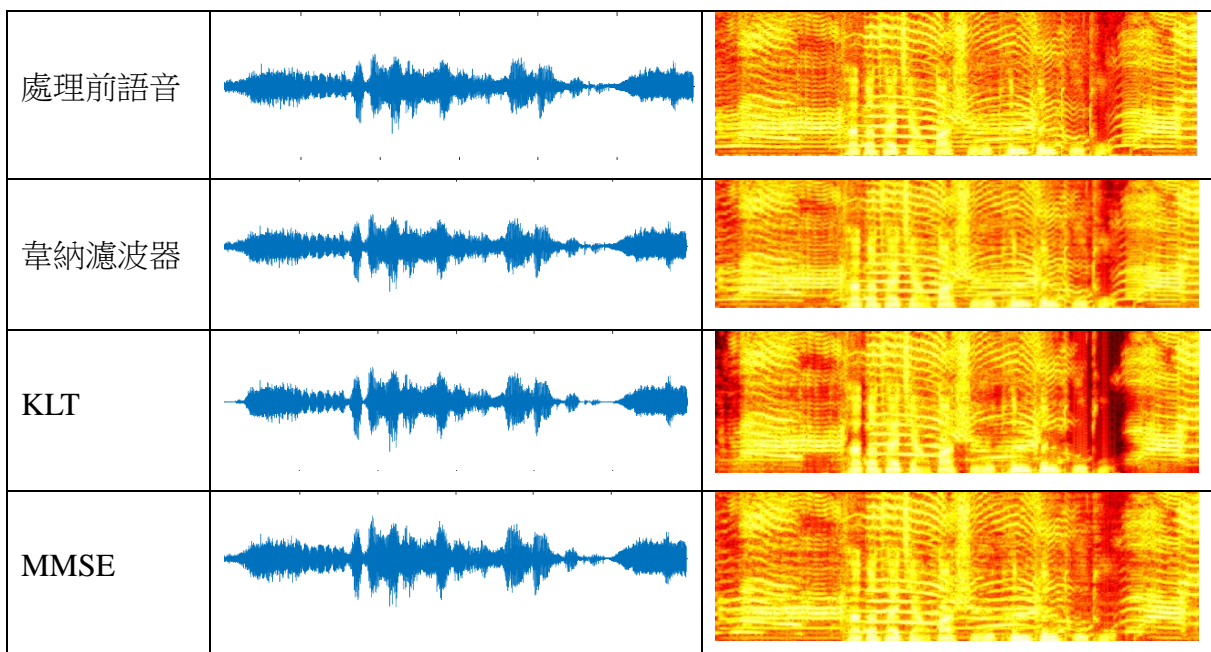


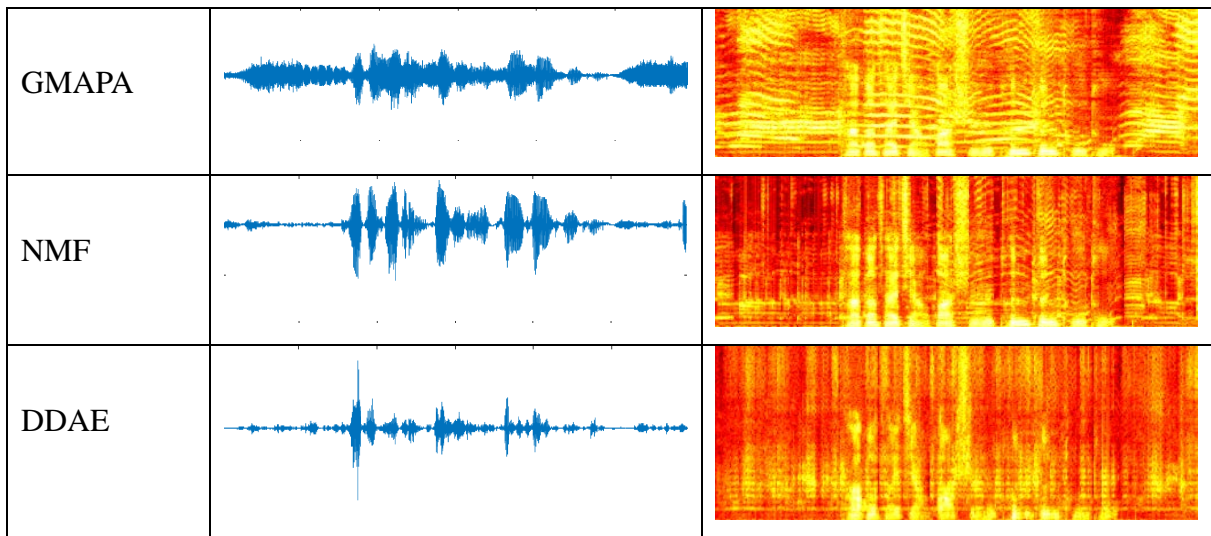
表二 各語音增強方法對車噪音處理後的 PESQ

	-5dB	0dB	5dB
Wiener Filter	1.55	1.65	2.29
KLT	1.74	1.83	2.60
MMSE	1.95	2.19	2.43
GMAPA	1.88	2.15	2.44
NMF	1.91	2.18	2.52
DDAE	2.18	2.41	2.67

表三的噪音為嬰兒哭聲，從表三可看出傳統的方式並無法有效的降低嬰兒哭聲，而機器學習的語音增強方法可以有效的降低嬰兒哭聲。嬰兒哭聲為非穩態的噪音，且嬰兒哭聲的聲學特徵類似人聲，使用傳統方法較難準確的預估並消除。從客觀評量中(如表四)也可得到相同的結論。

表三 各語音增強方式對嬰兒哭聲的處理成效。





表四 各語音增強方法對嬰兒哭聲處理後的 PESQ

	-5dB	0dB	5dB
Wiener Filter	1.45	1.55	1.59
KLT	1.47	1.50	1.59
MMSE	1.47	1.49	1.49
GMAPA	1.46	1.49	1.52
NMF	1.99	2.11	2.57
DDAE	2.05	2.29	2.62

五、結論

在本研究中我們成功建構了語音增強系統使用者介面，在此介面上包含傳統語音增強與機器學習所開發出來的語音增強方法，簡單的操作介面提供使用者可快速的進行模擬與方法的選擇。目前仍然沒有一個演算法可以有效的抑制所以有噪音，在做語音增強演算法的選擇上還是根據噪音做選擇，此平台可以做為一輔助工具，快速檢視該噪音以何種演算法較為有效。

參考文獻

- [1] Su, Y.-C., Tsao, Y., Wu, J.-E., Jean, F.-R., "Speech enhancement using generalized maximum a posteriori spectral amplitude estimator," in Proc. ICASSP, pp. 7467-7471, 2013.
- [2] Scalart, P., et al., "Speech enhancement based on a priori signal to noise estimation," in Proc. ICASSP, pp. 629-632, 1996.

- [3] Hsu, C.-C., Cheong, K.-M., Chien, J.-T., and Chi, T.-S., “Modulation Wiener filter for improving speech intelligibility,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 370-374, 2015
- [4] Boll, S., “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [5] Ephraim, Y. and Malah, D., “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [6] Tsao, Y. and Lai, Y.-H., "Generalized Maximum a Posteriori Spectral Amplitude Estimation for Speech Enhancement," *Speech Communication*, vol. 76, pp. 112–126, 2016.
- [7] McAulay, R. and Malpass, M., “Speech enhancement using a soft decision noise suppression filter,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.
- [8] Lotter, T. and Vary, P., “Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model,” *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.
- [9] Ozerov, A. and Févotte, C., “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
- [10] Hsu, C.-C., Chien, J.-T., and Chi, T.-S., “Layered nonnegative matrix factorization for speech separation,” *InterSpeech*, pp. 628-632, 2015
- [11] Lu, X., Tsao, Y., Matsuda, S., and Hori, C., “Ensemble modeling of denoising autoencoder for speech spectrum restoration,” in *Proc. INTERSPEECH*, pp. 885–889, 2014.
- [12] Lee, Y.-S., Wang, C.-Y., Wang, S.-F., Wang, J.-C., and Wu C.-H., “Fully complex deep neural network for phase-incorporating monaural source separation,” *Proceedings of ICASSP2017, New Orleans, USA, March 5~9, 2017*
- [13] Huang, K.-Y., Wu, C.-H., Su, M.-H., and Fu, H.-C., “Mood detection from daily conversational speech using denoising autoencoder and LSTM,” in *Proceedings of ICASSP2017, New Orleans, USA, March 5~9, 2017*