The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 152-166

適合漸凍人使用之語音轉換系統初步研究

Deep Neural-Network Bandwidth Extension and Denoising Voice Conversion System for ALS Patients

黃百弘 Bai-Hong Huang, 廖元甫 Yuan-Fu Liao

國立臺北科技大學電子工程系

Department of Electronic Engineering, National Taipei University of Technology

tjtkng@gmail.com, yfliao@ntut.edu.tw

Matúš Pleva, Daniel Hládek

Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia

matus.pleva@tuke.sk, daniel.hladek@tuke.sk

摘要

漸凍人症（肌萎縮性脊隨側索硬化症，Amyotrophic lateral sclerosis， ALS）為一種神經退化性疾病，這種疾病目前還沒有治癒的方法，並會讓漸凍人慢慢失去說話能力，最終導致無法利用語音與人溝通，而失去自我認同。因此，我們需要為漸凍人建立適合其使用之語音溝通輔具（voice output communication aids, VOCAs），尤其是讓其能具有個人化的合成語音，即病友發病前的聲音，以保持自我。但大部分在 ALS 後期，已經不能講話的病友，都沒有事先妥善保存好個人的錄音，最多只能找出有少量大約 20 分鐘的低品質語音，例如經過失真壓縮（MP3）、只保留低頻寬（8 kHz），或是具有強烈背景雜訊干擾等等，以致無法建構出適合 ALS 病友使用的個人化語音合成系統。針對以上困難，本論文嘗試使用通用語音合成系統搭配語音轉換演算法，並在前級加上語音雜訊消除（speech denoising），後級輔以超展頻模組（speech super-resolution）。以能容忍有背景雜訊的錄音，並能將低頻寬的合成語音加上高頻成分（16 kHz）。以盡量能從低品質語音，重建出接近 ALS 病友原音的高品質合成聲音。其中，speech denoising 使用 WaveNet，speech super-resolution 則利用 U-Net 架構。並先以 20 小時的高品質（棚內錄音）教育電台語料庫，模擬出成對的高雜訊與乾淨語音語句，或是低頻寬與高頻寬語音，

分別訓練 WaveNet 與 U-Net 模型，再用以處理病友的低品質語音錄音音檔。實驗結果顯示，訓練出來的 WaveNet 與 U-Net 模型，可以相當程度還原具雜訊或是低頻寬的教育電台語音檔。並能用來替 ALS 病友重建出高品質的個人化合成聲音。

關鍵詞：類神經網路、ALS、WaveNet

一、緒論

漸凍症全名，肌萎縮性脊隨側索硬化症(Amyotrophic lateral sclerosis， ALS)，為一種逐漸且要命的神經退化性疾病，ALS 患者因中樞神經系統內控制骨骼肌的神經元退化使大腦逐漸完全喪失控制肌肉運動的能力，病程晚期會影響語言能力、進食、和呼吸系統的運行，但此疾病並不見得會影響 ALS 患者的思考能力；相反，ALS 患者晚期依然維持完整的思緒、具有發病前的人格、智力及記憶，所以病友即便影響到聲帶無法發出聲音，但病友的思緒依然是然全不受影響的，在病友無法自身發出聲音與他人溝通的情況下就需要用到音溝通輔具（voice output communication aids, VOCAs），目前 VOCAs 常用的方法是使用文字轉語音(Text To Speech,TTS)來作為病友語音輸出，病友經過前端裝置輸入文字再由 TTS 轉換成聲音，即便目前 TTS 已能模擬出非常接近人類的聲音，卻無法還原出病友在未發病時的聲音特色，滿足病友想用發病前聲音說話的自我認同需求，且病友家屬也強烈表示希望病友能用具有發病前個性的聲音與他們溝通。

如果要為病友建立個人的 TTS 需要病友在擁有大量的高品質音檔下才能完成，在一般的情況下聲音的資料是非常容易取得的，只要將說過的話錄下來便能成為語料，但這件看似簡單的事情對於晚期的漸凍病友卻不是如此，在病情發展到影響聲帶時，病友將失去語言能力，或在配帶呼吸器的情況下要正常發音是非常困難的事情，所以只能仰賴平日錄製的語音，但漸凍病友的病期難以推敲，且每位病友受病情影響的部位無法預測，所以有預前錄製音檔的病友相當稀少，即便有數量也非常少。

在病友先前自行錄製中我們發現，病友家屬將以手邊最容易取得之錄音系統錄製音檔，如手機或錄音筆，所以錄音語料面臨兩大問題，第一環境音雜音大無法正確分析語音特徵，在實驗中有一位病友家屬提供我們的音檔為測錄音檔，冷氣的低頻聲與環境回音發常嚴重甚至已經超過病友的聲音，聲碼器在提取聲音特徵時大受影響，完全無法有效的提取特徵參數，使得後面的訓練一蹋糊塗，轉換出來的聲音富含嚴重雜音，根本無法辨識為病友聲音，但病友無法再提供其他錄音檔，只有嚴重雜訊音檔為病友轉換音

檔，第二病友語料取樣頻率不足高頻完全消失，情況是病友家屬有事先幫病友保存語音，病友提供我們一小時的語音，但經過處理後只剩下 16 分鐘可用音檔，音檔為 AMR 破壞性壓縮，並且取樣頻率只有 8kHZ，聲音輸出少了高頻聲音顯得不真實，無法完整還原病友發病前聲音，所以本論文將傳統語音轉換系統前級增加 Speech Denosing[1]與在後級再以 Super-Resolution[2]來改善，如圖一所示。

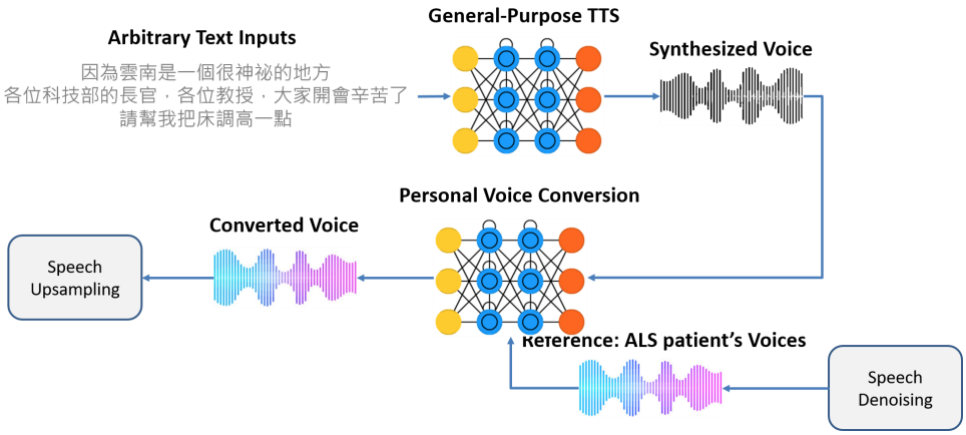


圖 一、ALS 病友語音轉換系統圖

本論文將病友語音輸入前級加上 Speech Denosing 確保病友音檔輸入為無雜訊干擾語音，在轉換後音檔後級加上 Super-Resolution 確保病友轉換音檔輸出為具有高頻的較高品質音檔，期望經過這套系統不僅能確保病友已剩不多的語音，保持輸入訓練音檔品質，也確保輸出為病友具有高頻還原語音。

二、基於類神經網路病友語音轉換系統

本論文提出的 ALS 語音轉換系統如圖 1，輸入語音使用微軟釋出的 Hanhan TTS(Text To Speech)[3]，將病友語音經過前級 Speech Denosing 系統之後作為目標語者(病友)之語音，建立個人語音轉換系統，最後經過後級 Super-Resolution 系統將病友高頻補足。

(一) ALS 語音轉換系統

1 bi-LSTM 語音轉換系統架構

Bi-LSTM 語音轉換系統架構，如圖二，先將 TTS 來源語者經由 Vocoder 取出語者特

徵，在經過 bi-LSTM 模型進行學習，轉換成病友語音特徵，最後經由 Vocder 將特徵組合回來，輸出病友轉換語音。

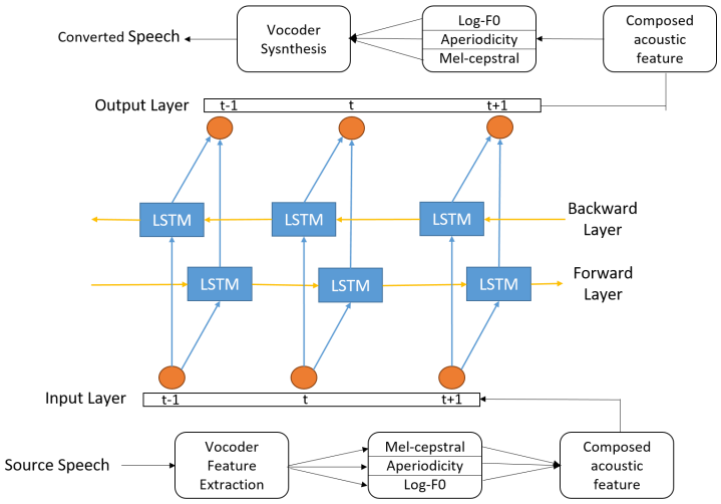


圖 二、BLSTM 語音轉換系統架構

2 實現方法

本論文語音轉換系統使用平行語料架構，系統第一步是要擷取來源與目標語者每一個音框的特徵參數，再利用 Dynamic Time Warping（DTW)[4]進行音框對齊。在語音轉換的過程中，首先透過 Vocoder 來分析音檔裡面的資訊，分別包含了下列三個重要的特徵 Mel-cepstral[5]、Band Aperiodicity、Log-F0，其中 Mel-cepstral 是聲音的頻譜，Band Aperiodicity 是非週期性的特徵用來判斷聲音是否發聲，最後 Log-F0 決定聲音的音高，這些特徵參數將作為模型的輸入進行訓練，在訓練之前因為來源語者與目標與者說話的速度不會相同，Dynamic Time Warping（DTW)解決來源與目標音框長度不同的問題。再以 DTW 的 alignment 結果，計算每一個來源與目標對應音框，應該收縮（shrinking）、拉長（stretching）、還是保持不變（kept）如下圖三所示。

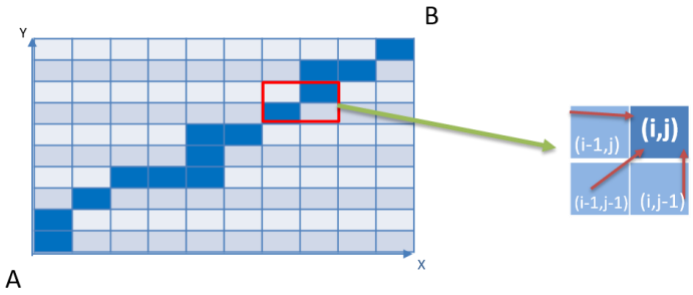


圖 三、 Dynamic Time Warping(DTW)演算法示意圖

接下來將 DTW 對其資料丟入 bi-LSTM 模型中訓練，LSTM 網路是一種特殊的 RNN 結構，可以描述時變的語音訊號，並且把之前的資訊帶到當前的訓練任務中，，LSTM 的結構能夠用來防止長距離依賴問題，也就是可以解決梯度消失的問題，bi-LSTM 是將 LSTM 與 BRNN 結合在一起，這種方法可以在輸入的方向或的長時的上下文信息，效果優於 LSTM，經過系統訓練後會產生病友的個人語音模型，系統後級會將目標語者之 Mel-cepstral、Band Aperiodicity、Log-F0，重新經過 Vocder 組合，將病友轉換語音輸出，最後我們就可以達到使用來源語者說任何語句都能轉換成具有病友語音個性的病友專屬語音轉換系統。

(二) 強化病友轉換語音系統

在摘要我們提到 ALS 病友後期已經無法在錄音，但病友卻都沒有妥善的保存聲音，在處理病友音檔時遇到兩大問題，第一環境音雜音大無法正確分析語音特徵，第二病友語料取樣頻率不足高頻完全消失，在病友已經無法在錄製音檔的情況下，我們希望可以最大限度使用病友已所剩不多的音檔，所以我們將這兩套系統加入傳統的轉換系統中。

1 Speech Denosing(消雜訊系統)

語音消雜(Speech Denosing)的問題上，大多數用於語音消雜的技術使用頻譜圖作為前端 [6,7,8,]。然而，這種做法帶來了潛在丟棄的缺點 有價值的信息（階段）和利用通用特徵提取器（頻譜圖分析）而不是學習給定數據分佈的特定特徵表示，最近，神經網絡已經證明在處理離散化音頻信號的樣本之間的結構化時間依賴性方面是有效的，有趣的是， 大多數這些生成模型都是自回歸模型[9,10]，WaveNet[10]是在自然語音上被廣泛利用得自回歸模型，本論文使用 WaveNet 模型建置 Speech Denosing，目標為最大保留已剩不多的病友語音。

(1)系統架構

圖四為 Speech denoising 系統架構圖，將病友雜訊聲音經過系統消雜後，可以輸出較為乾淨的病友語音，目標為保留病友乾淨語音供後端轉換系統使用。

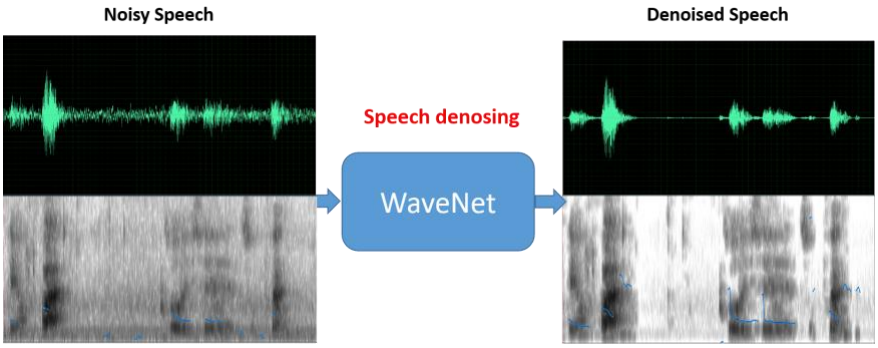
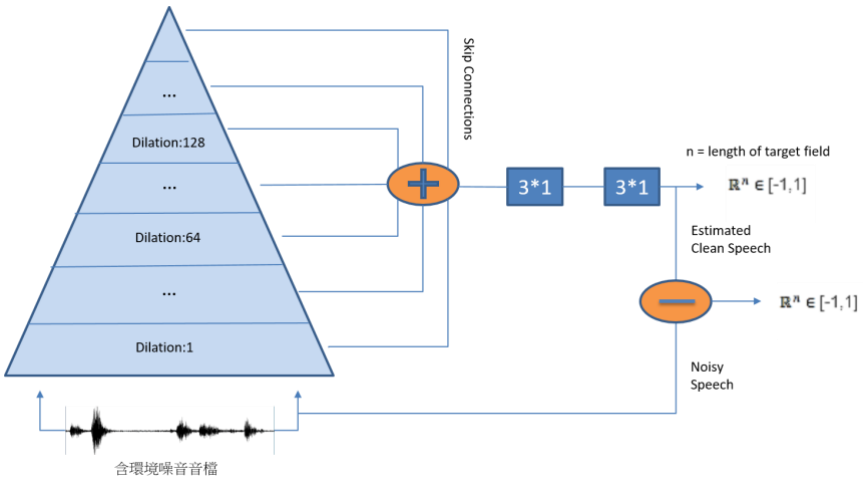


圖 四、Speech denoising 系統架構圖

(2) 實現方法

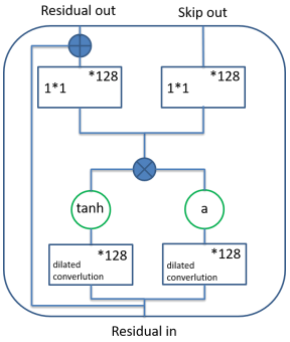
問題表述如下：mt = st + bt，其中：mt＝混合信號，st＝語音信號，bt＝背景噪聲信號。 目標是估計給定的 mt，本論文使用 FaNT[18]將雜音添加入乾淨音檔，所以將混雜噪音的音檔當做 noisy voice，將乾淨的音檔作為 clean voice，系統先經過對齊確定兩個音檔長度是否一致，再以 clean voice 減去 noisy voice 就可以得到我們事先添加的雜訊，將得到資料丟入 Wavenet 架構下訓練，WaveNet 能夠合成自然的發聲語音。 這種自回歸模型的形狀給出先前樣本的一些片段的下一個樣本的概率分佈，本論文的模型的描述在圖五中，



圖五、WaveNet 模型示意圖

本論文所提出的模型具有 30 個殘餘層如圖六，每層中的膨脹因子 以 2 為倍數增加 1，2，...，256,512。該模式重複 3 次（3 個堆疊）。 在第一次擴張卷積之前，1 通道輸入被線性投影到 128 個通道 標準的 3x1 卷積，以符合每個殘留層中的濾波器數量。跳

過 連接是 1x1 卷積，還有 128 個濾波器，在匯總所有後應用 RELU 跳過連接。最後兩個 3x1 卷積層未擴張，包含 2048 和 256 個濾波器，分別由 RELU 分隔。輸出層將要素圖線性投影到 a 使用 1x1 濾波器的單通道時間信號，該參數化導致感受野 共 6,139 個樣本（≈384ms），目標字段由 1601 個樣本（≈100ms）組成。



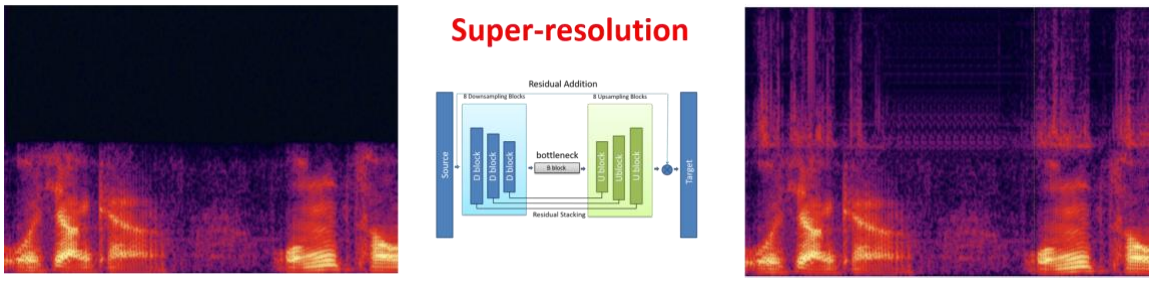
圖六、剩餘層示意圖

2 Super-Resolution(展頻系統)

超分辨率(Super-Resolution)廣泛被使用來解決許多應用中的低分辨率問題[13]，這些技術重建或學習消失的高頻信息以增強成像系統的分辨率。最近，這些技術已被有效地用於提高分辨率，主要是為了提高生物識別系統的識別性能包括 face [14]和 iris [15]，但多半是圖像上的利用，在音頻上一樣可以使用超分辨率(Super-Resolution)來回復消失的高頻音頻，本論文使用 U-Net[16]架構來完成超分辨率(Super-Resolution)系統建置，目標為有效的恢復病友高頻的聲音。

(1) 系統架構

本系統將病友轉換語音輸入為 8kHZ 的低頻音檔，經過 Super-Resolution 系統後可以展頻出 16kHZ 的具有高頻的音檔，圖七為展頻系統架構圖，目標為系統可以有效的還原出病友高頻音檔。



圖七、展頻系統架構圖

(2)實現方法

模型架構是 U-Net，使用虛擬的一維子像素卷積 Sub-Pixel Convolutions[17]，圖八為 U-Net 架構圖，系統經過左側八個下採樣過程分八組捲積來執行，每組捲積後進行 maxpool，將音檔進一步縮小為原本的二分之一，通過八次操作將 39998*8 語音計算成 154*2048，右側的採樣程，使用 8 組反捲積，每次上採樣將語音擴展成 2 倍然後將對應層的語音進行剪輯與複製，然後 concat 到上捲積結果上，完成上採樣後得到 40000*16，最後使用 1*1 捲積合將通道數減為 2 來代替捲積層。

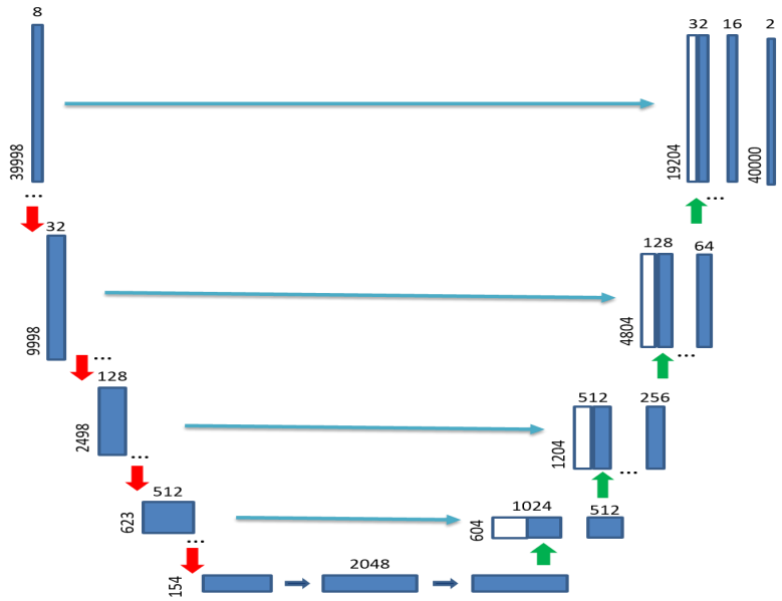


圖 八、 U-net 架構圖

系統將數據經過採樣剪輯並分批丟入模型中以更新其權重。保存驗證分數最低的模型儲存為最佳模型，經過每個下取樣波形(Downsampled waveform)通過八個下取樣塊(Downsampling blocks)發送，經由瓶頸層(Bottleneck)連接到八個上取樣塊(Upsampling blocks)，瓶頸層跟下取樣塊之間有殘差連結(Residual connections)，這些殘差連結彼此

共享特徵資訊。上取樣塊使用子像素卷積，沿一個維度重新排序信息以擴展其他維度，最後由最終一層卷積層將重新排序堆疊後的模塊與殘差加成(Residual addtion)加在原始輸入音檔中產出升頻後得音檔。

三、強化病友合成語音實驗

強化病友合成語音實驗中，目的是為了使病友合成語音更接近病友發病前聲音，本文將比較合成語音經過 Super-Resolution 與 Speech Denosing 系統強化前後是否有效改善語料高頻不足與雜音干擾等問題，在實驗中會比較上述兩個系統的主觀客觀偏好，來評斷結果。

(一) 訓練與測試語料

本文建立展頻系統與消雜訊系統皆使用教育電台廣播節目播出時所錄製的音檔，我們從中挑出其中的 7 個節目共 20 小時，並經過人工剪輯成只有人聲的音檔當作 Data 使用，表 1 為訓練資料統計表，表 2 為測試資料統計表。

表一、訓練資料統計表

類型	節目名稱	集數	音檔總時間(minute)
純人聲	創設市集	6	155
	國際教育心動線	7	185
	技職最前線	5	105
	青農市集 On Air	20	457
	晨間新聞	10	258

表二、 測試資料統計表

類型	節目名稱	集數	音檔總時間(minute)
純人聲	教官不說教	5	92
	兒童新聞	4	60

(二) 實驗設定 - Speech Denosing 消雜訊系統

消雜訊系統實驗中，本文先將音檔消完雜訊完後再輸入 ALS 語音合成系統中重新訓練模型，為滿足大多是環境音條件，本文模擬各種病友在自行錄音時可能會遇到的環境雜訊如表 3。

表三、添加雜訊表

雜訊名稱	音檔時數(hr)
環境有人走動聲響(運動場)	20hr
環境有他人說話雜訊(餐廳)	20hr
音檔總時數	40hr

(三) 實驗設定 - Super-Resolution 展頻系統

展頻系統實驗中，我們將教育電台語料處理成低頻音檔 8kHZ 取樣頻率和高頻音檔 16kHZ 取樣頻率，作為 Data 如圖九，十。

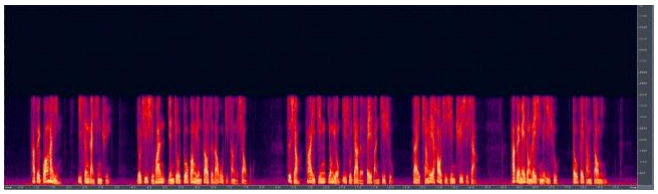


圖 九、 8kHZ 音檔頻譜圖

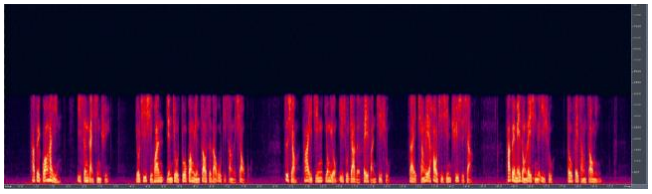


圖 十、 16kHZ 音檔頻譜圖

(四) 評估方法

1 主觀分數評估

本文系統偏好的評估方式，將測試音檔給 5 位包含病友家屬及母語為中文的人員進

行評分，系統偏好度測試是 2 選 1 的方式，為標準的 A/B/X 測試，系統評分採用平均主觀值分數(mean opinion score , MOS)進行評估，分為可理解度評分、相似度評分和自然度評分，評分方式為 1~5 分，分數越高則為越好。

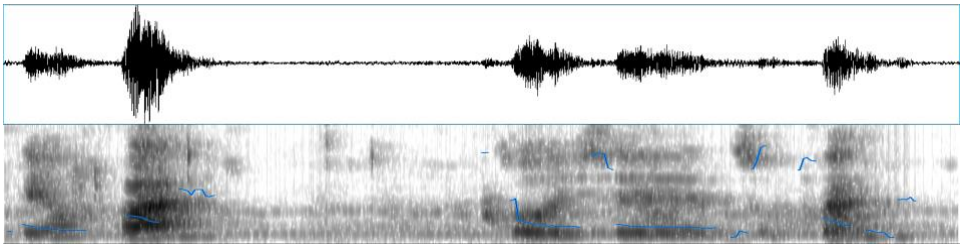
2 客觀分數評估

客觀評估方式比較聲音的參數在訓練後與目標值的差異，客觀評估的分數直接對應到轉換的效果程度好壞評估的內容包含了 Mel-cepstral distortion (MCD): 均方根誤差單位 dB、BAP: 均方根誤差 單位 dB、F0-RMSE: 均方根誤差 單位 Hz 以及 VUV: %(以百分比表示)。

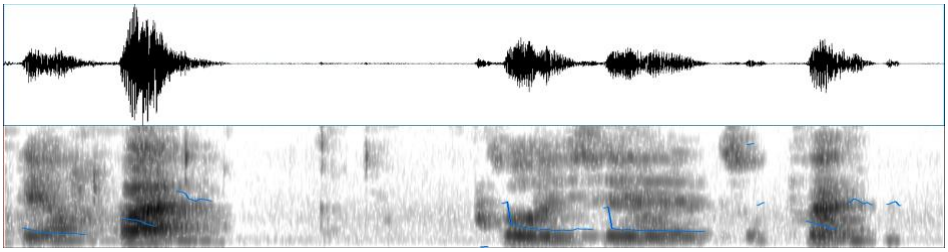
四、實驗結果

Speech Denosing 強化病友合成語音實驗

由圖十一，十二可發現，病友音檔經過消雜訊系統後，低頻雜音已明顯濾除，音檔 F0 部分也可以穩定的求出，表四為未經消雜與消雜與音合成系統客觀評估表，數據顯示 MCD、BAP、F0-RMSE、VUV 的錯誤率明顯降低。經過消雜訊系統後合成聲音在 A/B/X 偏好評估或 MOS 主觀評估中都優越於無消雜續系統，結果如下圖與表



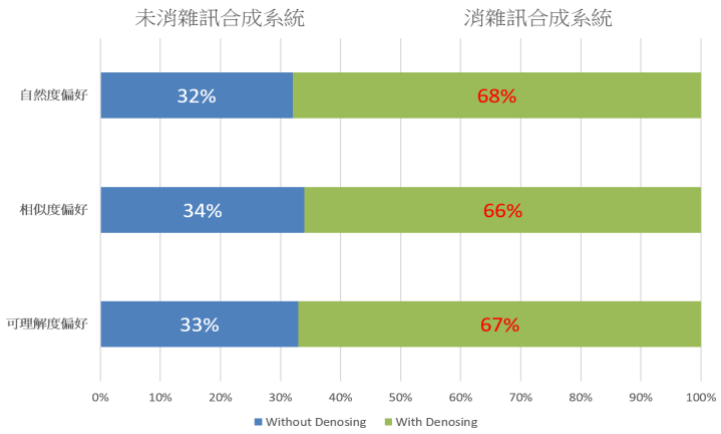
圖十一、未消音音檔圖



圖十二、已消雜訊音檔圖

表四、消雜訊系統客觀評估表

	未消雜與音合成系統	消雜語音合成系統
MCD	9.265 dB	6.869 dB
BAP	0.542 dB	0.208 dB
F0-RMSE	68.597 Hz	49.443 Hz
VUV	62.851%	25.935%



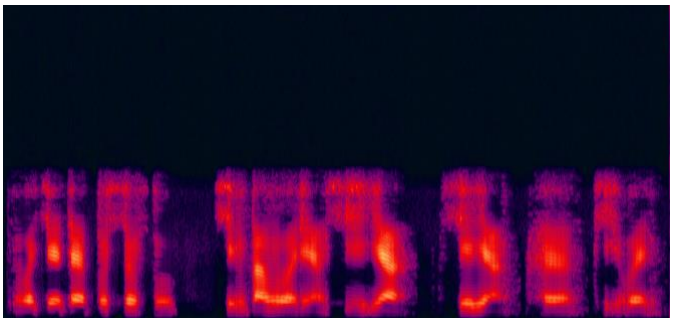
圖十三、消雜尋合成系統 A/B/X 偏好測試圖

表五、消雜訊系統 MOS 主觀評估表

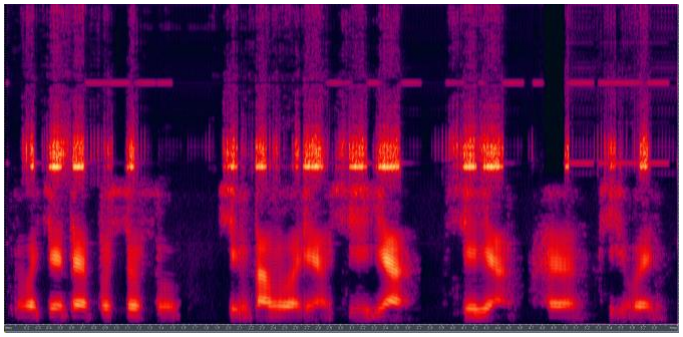
	未消雜訊合成系統	消雜訊合成系統
自然度評分	2	4
相似度評分	2.1	4
可理解度評分	2	4.1

Super-Resolution 強化病友合成語音實驗

經由圖十四，十五比較明顯可以發現，經過展頻系統，系統有相當程度的將病友原先缺少的高頻音域模擬出來。

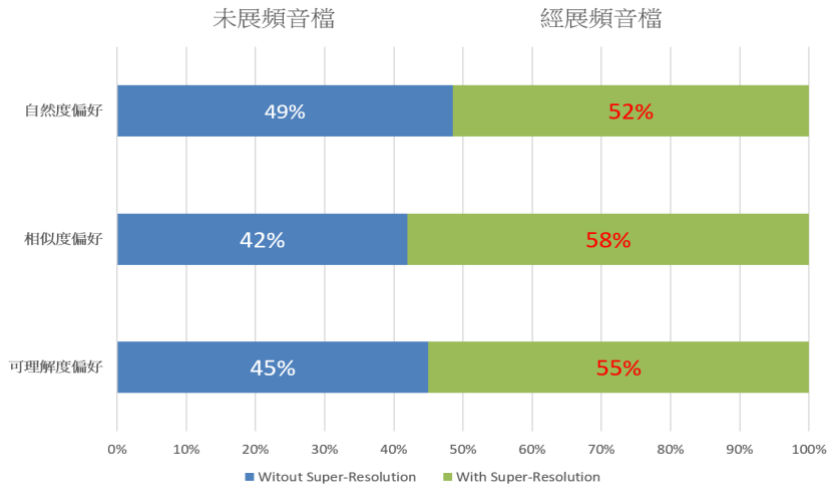


圖十四、未經過展頻系統之病友合成語音頻譜圖



圖十五、經由展頻系統補救之病友合成語音

經由展頻系統之病友合成語音比未展頻合成語音在可理解度、相似度以及自然度偏好效果都比較好，更重要的是無論在 A/B/X 偏好評估或 MOS 主觀評估中，相似度偏好都明顯高過未展頻之音檔。



圖十六、展頻音檔 A/B/X 偏好測試圖

表六、展頻音檔 MOS 主觀評估表

	未展頻音檔	展頻音檔
自然度評分	3.4	3.6
相似度評分	3.1	3.8
可理解度評分	3.3	3.6

五、結論

在實驗中我們成功將病友語音經過 Super-Resolution 系統與 Speech Denosing 系統的強化，不管是相似度或可理解度上都有顯著的提升，病友與病友家屬也給予我們正面的評價。本論文所提出 Super-Resolution 系統與 Speech Denosing 系統僅解決了兩種病友語問題，著重於語料品質上得改善，但事實上在聲音需求上不只漸凍病友需要其他重大傷病病友也有需求，還有許多語音問題需要處理，如:病友還能發聲，但聲音卻不如之前，說話音調與耐力受到影響，是否可以從病友目前語料與發病前語料中找到改善方法，可以從現有資料解決病友需求。

Acknowledgements

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and finally by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067.

參考文獻

[1]. Volodymyr Kuleshov, S. Zayd Enam, and Stefano Ermon : AUDIO SUPER-RESOLUTION USING NEURAL NETS ICLR 2017

[2]. Dario Rethage, Jordi Pons, Xavier Serra : A Wavenet for Speech Denoising arXiv:1706.07162v3

[3]. Heiga Zen : Acoustic Modeling for Speech Synthesi Dec. 14th, 2015@ASRU

[4]. Stan Salvador , Philip Chan. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space. Dept. of Computer Sciences Florida Institute of Technology Melbourne, FL 32901

[5]. Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition

algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[6]. Anurag Kumar and Dinei Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. arXiv preprint arXiv:1605.02427, 2016.

[7]. Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In Interspeech, pages 436–440, 2013.

[8]. Shahla Parveen and Phil Green. Speech enhancement with missing data techniques using recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages I–733, 2004.

[9]. Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837, 2016

[10]. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[11]. Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759, 2016.

[12]. Kominek, John, Tanja Schultz, and Alan W. Black. "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." Spoken Languages Technologies for Under-Resourced Languages. 2008

[13]. S.C. Park, M.K. Park, and M.G. Kang, “Super-resolution image reconstruction: a technical overview,” IEEE Signal Processing Magazine, vol. 20, no. 3, pp. 21 – 36, 2003.

[14]. F. Lin, C. Fookes, V. Chandran, and S. Sridharan, “Super-resolved faces for improved face recognition from surveillance video,” in Lecture Notes in Computer Science, Seoul, Korea, Republic of, 2007, vol. 4642 LNCS, pp. 1 – 10.

[15]. Kwang Y.S., Kang R.P., Byung J.K., and Sung J.P., “Super-resolution method based on multiple multi-layer perceptrons for iris recognition,” in Ubiquitous Information Technologies Applications, ICUT ’09. International Conference on, 20-22 2009, pp. 1 –5

[16]. Olaf Ronneberger，Philipp Fischer，Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation arXiv:1505.04597

[17]. Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, Zehan Wang: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

[18]. Andreas Kitzig: Nieder Rhein University, Room Impulse Response-package: NRU-RIR-package.