Design and Development of a Speech Assistive Device for Esophageal Speakers
食道語者語音輔助裝置之設計與開發

Yan-Zhi Chen, Yen-Ting Lin, Bo-Sen Liang, Ding-Lian Chen, Chen-Yu Chiang
National Taipei University, New Taipei City, Taiwan
oscar900517@gmail.com, d26923050@gmail.com, benson20030603@gmail.com,
shadow449515@gmail.com, cychiang@mail.ntpu.edu.tw

摘要

食道語者因疾病或手術等原因失去聲帶，無法像大多數人一樣使用聲帶發音，而是使用食道振動來取代聲帶振動。然而，食道語在發聲過程中需要頻繁吞氣，產生大量雜音和爆音，使得一般人較難以理解。爲了解決這一問題，我們開發了一個優化食道語的系統。該系統通過分析食道語的頻譜特性，並且利用機器學習技術來過濾雜音並抑制爆音。我們將食道語分爲四類：爆音（burst, B）、正常語音（voice, V）、雜音（noise, N）和靜音（silent, S）。這個系統的主要目的是在使用較少計算資源的前提下，消除食道語中的雜音並抑制爆音，從實驗結果顯示來看確實有顯著提高食道語的可理解性。

Abstract

Esophageal speakers lose their vocal cords due to conditions such as illness or surgery, making it impossible for them to use their vocal cords to produce sound like most people. Instead, they use esophageal vibrations to replace vocal cord vibrations. However, during the phonation process of esophageal speech, frequent air swallowing is required, which produces a lot of noise and burst sounds, making it difficult for the average person to understand. To address this issue, we developed a system to enhance esophageal speech. This system analyzes the spectral characteristics of esophageal speech and uses machine learning techniques to filter out noise and suppress burst sounds. We categorize esophageal speech into four types: bursts (B), normal voice (V), noise (N), and silence (S). The primary goal of this system is to eliminate noise and suppress burst sounds in esophageal speech while using fewer computational resources. The experimental results show a significant improvement in the intelligibility of esophageal speech.

關鍵字：語音轉換、語音增強、食道語

Keywords: Voice conversion, Speech enhancement, Esophageal speech

1 Introduction

1.1 研究背景及方向

食道語 (Kaye et al., 2017) 是利用食道內的空氣振動來產生聲音。食道語者通過吞咽或吸入空氣進入食道，然後利用食道和咽喉部的肌肉控制這些空氣的排出，在此過程中形成振動並產生聲音。這些聲音再通過口腔和鼻腔共鳴，產生可理解的語音，許多人因爲各種原因而失去了聲帶，不得不使用食道語與他人進行溝通。然而，學習並熟練掌握食道語需要努力且長久的練習，並不是一朝一夕可以精通的。食道語的發聲過程中，由於需要頻繁吞咽空氣來維持聲音的產生，這會導致顯著的噪音問題。吞咽動作本身會產生低頻的吞咽聲，而空氣在進入食道和從食道排出時，也會產生明顯的進氣和排氣聲。這些聲音，尤其是進氣聲，往往是突然且不規則的，並且可能帶有爆發性的音質，這會增加語音中的背景雜音，進而影響語音的清晰度和可理解性。這種噪音不僅讓聽者難以理解語音內容，還會增加交流的困難度，形成的雜音也更容易導致他人難以理解。

本研究錄製了一個食道語者的語音，使用 VoiceBank-2023 語料庫 (Su et al., 2023) 裡面總計 105 句的文本，總錄製時長約 18 分鐘。使用 VoiceBank-2023 語料庫是因爲語料庫有經過特別的設計，照順序累積語句，可以比較以較少的語句來包含中文的聲母和韻母，即是在語句檔案較少的情況下，也能得到更好的效果。

從食道語者的語音中可以觀察到，食道語者在發音時一開始的聲音振幅偏大，這是因爲中文大部分的音節從 consonant（子音）開始，加上他們只能利用食道內的空氣壓力來擠

出聲音，這導致接在子音後面的 vowel（母音）或韻母音量較小，可能在時域上產生了「遮蔽效應」，使母音相較子音顯得較不清晰。我們觀察到無聲子音如ㄐ、ㄑ、ㄒ、ㄓ、ㄔ、ㄕ、ㄗ、ㄘ、ㄙ等容易會產生爆音，因此我們把上述歸類在爆音（B）這個標籤，而無聲子音像是ㄅ、ㄆ、ㄈ、ㄉ、ㄊ、ㄍ、ㄎ、ㄏ雖是無聲子音但是根據觀察，發生爆音的概率較低，因此我們把上述類別和有聲子音ㄇ、ㄋ、ㄌ、ㄖ及母音歸類在正常語音（V）的標籤裡面，只降低容易爆音的無聲子音振幅並預期會有顯著改善食道語的清晰度。Table 1 是我們統計 105 句語音檔案，判斷爆音的標準爲振幅超過我們所設定的閥值 0.7 時，即被判定爲爆音。語音振幅超過 0.7 時對應的標籤結果如下:

Label	count	Probability (%)
B	1603	87.452%
V	220	12.002%
S	1	0.055%
N	9	0.491%

Table 1: 爆音對應標籤的發生機率

根據上表可以看出，我們這樣的分類方法是可行的，也能發現歸類在爆音 (B) 類別中的無聲子音確實容易導致爆音的現象。這說明無聲子音對於食道語者發音時，更容易觸發爆音的產生。透過這樣的分類，我們可以更清楚地了解各類聲音在語音處理中的行爲，並進一步優化語音處理系統的性能。

「遮蔽效應」廣義上指的是能量較大的聲音能夠掩蓋或抑制能量較小的聲音。在音頻信號處理中，遮蔽效應通常可以細分爲頻域遮蔽和時域遮蔽。頻域遮蔽指的是在頻率域中，能量較強頻帶聲音會掩蓋能量較弱的頻帶，而時域遮蔽則是指在時域中，可以分成兩種類型: 前遮蔽（pre-masking），主要是發生在能量較強的聲音出現之前的短時間內，能量較較弱的目標聲音會被遮蔽。以及後遮蔽（post-masking），發生在能量較強的聲音出現之後的一段時間內，較弱的目標聲音被遮蔽。對於食道語者而言，主要面臨的是時域遮蔽的問題。當食道語者使用食道發聲時，由於聲音的振幅和頻率特徵與正常語音存在差異，這會導致食道語中的聲音在時間域上出現遮蔽效應。具體而言，人類在接收到聲音信號後需要一定的時間進行處理，這意味著如果前後的聲音有較大的能量差異，則會導致前後聲音的辨識及處理上的困難。如 Figure 1 所示:

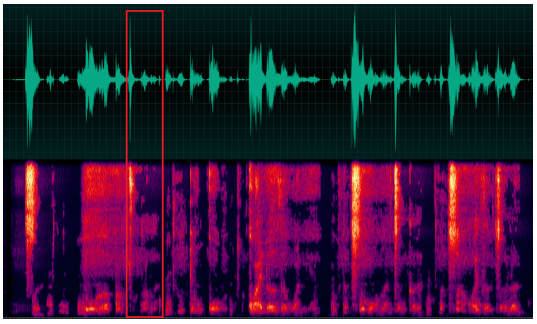


Figure 1: 食道語者語音的波形與頻譜

從 Figure 1 紅框部分可以發現當爆音出現後，較強的能量遮蔽了接在後面的母音導致可能出現了後遮蔽（post-masking），使得降低了整段話的可理解性。

基於上述觀察，本研究期望開發一款輕量語音演算方法，未來用於輕便的裝置，如領夾式麥克風，體積小巧可藏於衣領後方，通過過濾雜音並調整音量來放大被遮蔽的聲音和縮小爆音，從而提升食道語的音質和清晰度。這種聲音輔助裝置的應用有望可以改善食道語者的生活品質，使他們能夠更自信地表達自己的想法。

1.2 相關研究

綜觀食道語的研究，從早期食道語音增強技術基於線性預測編碼（Linear predictive coding, LPC）(Kenji et al., 2002) 和共振峰合成的傳統方法開始，通過對 30 毫秒的語音樣本，每 10 毫秒位移進行一次 LPC 和音高分析，窗口之間會有 20 毫秒的重疊，以提高食道語者的語音品質。儘管這些方法成功改善了音高的準確性，但由於食道語者的聲音特性不穩定，容易出現爆音問題，導致音高提取的穩定性受到影響，最終造成合成語音的品質下降。此外，此技術還帶來了較高的延遲，並且處理後的聲音與原始聲音有明顯差異，這些都限制了其在實際應用中的普及性和實用性。隨著技術的進步，研究逐漸轉向了基於統計模型的語音轉換方法，特別是高斯混合模型（Gaussian mixture model, GMM）的應用 (Doi et al., 2010)。GMM 技術通過對頻譜特徵、基頻和非週期訊號進行建模，並利用平行語料進行訓練，實現語音的轉換。

後來的研究進一步引入了多重回歸高斯混合模型（MR-GMM）和核回歸高斯混合模型（KR-GMM），並加入了語音品質控制向量，旨在提高食道語音的自然度和語音內容理解度 (Yamamoto

et al., 2012)。這些方法通過控制語音的 male/female(gender), husky/clear (clearness), elder/younger(age),deep/thin(deepness) 等因素，成功提升了語音的個人化和適應性，但也對數據需求和計算資源提出了更高的要求。隨著深度學習技術的引入，研究者開始探索使用聲學隱性表徵（phonetic posteriorgrams, PPGs）(Chen et al., 2020) 結合各類模型來進行比較和食道語音的轉換。

總體而言，食道語音增強技術從傳統方法到統計模型，再到近年來的深度學習模型，顯示了技術發展帶來的顯著進步。然而，隨著技術的進步，這些方法的計算複雜度也逐漸增加。早期的 LPC 和共振峰合成方法相對計算簡單，但其缺點是容易受限於語音特性不穩定性，導致增強效果有限。隨後的 GMM 和 MR-GMM 方法通過更精細的建模，提高了語音的自然度和可懂性，但也導致了計算需求的增加，尤其是在多維特徵和大規模數據的訓練過程中。最近的深度學習技術，如 PPGs 結合各類神經網路模型，雖然在語音增強上展現了卓越的性能，但其龐大的計算量和對高效能硬體的依賴，成爲了在實際應用中推廣的瓶頸。

2 研究方法

2.1 系統目標及流程

我們的系統目標是透過聲學特徵提取和模型學習技術，專注於解決食道語者在發音過程中常遇到的「容易爆音的無聲子音」問題。這些無聲子音在語音中容易產生過大的振幅或過度的雜訊，從而影響語音的清晰度和可理解性。爲了有效地解決這些問題，系統將運用神經網路（neural network, NN）模型進行語音信號的分類和增強。模型將通過標註數據的訓練，學會辨識無聲子音中的爆音。系統主要目標是專注於降低語音訊號中過大的振幅，並有效地抑制食道語者吸氣所造成的吞口水雜訊，從而提升語音信號的整體音質，以確保語音在各種環境下都能保持清晰和自然。

2.2 系統架構

如 Figure 2 所示，系統主要分爲兩個階段：訓練階段和測試階段。在這個架構中，我們首先處理音檔數據，然後將其輸入到深度學習模型中進行分類。以下是系統架構的詳細說明：

訓練階段我們先對音檔做預處理，將 48kHz 取樣率的音檔降取樣至 16kHz 來進行本次實驗。並透過 Praat 軟體對音檔的每個段落進行標記，分配四種類型的標籤：1) 爆音（B）: 容易爆音的無聲子音、2) 正常語音（V）: 包含母音及其他子音、3) 雜音（N）和 4) 靜音（S）來讓模型學習特徵。並且抽取音檔的聲學特徵轉換成梅爾時頻譜 (Mel spectrogram)，其參數設置爲: 梅爾時頻譜的維度爲 80，音框長度爲 512 點，音框位移爲 160 點。我們將音檔每個音框以 50% 的重疊形式進行分割。並從音檔中提取出每 20 個音框 (約 0.032*20=0.64 秒) 組合成一張 20×80 維的 Mel spectrogram。這些 Mel spectrogram 與對應的標籤對齊後，輸入到神經網路（NN）中進行訓練。模型的輸出是四類分類：爆音（B）、正常語音（V）、雜音（N）和靜音（S）。

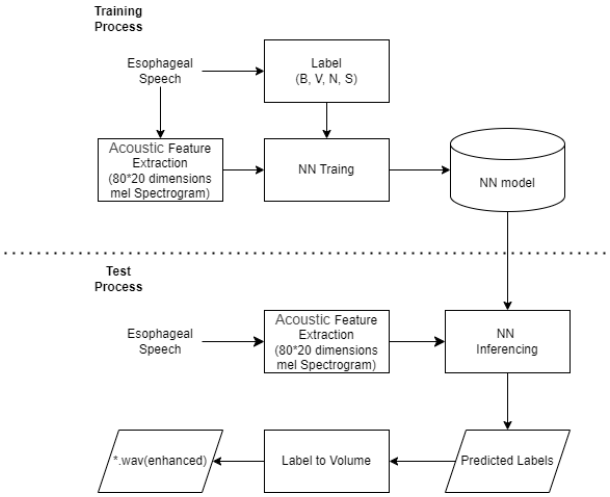


Figure 2: 系統架構圖

在測試階段，使用與訓練階段相同的數據預處理步驟來準備測試數據。然後將此輸入到已訓練好的模型中，獲取預測標籤（Predicted Labels）和加權音量（Volume）。爲了保證音檔的自然性，在分段處理時加入 Hann 窗以減少邊界效應，並重疊 50% 的形式，其數學式表示:

$$x_m[n] = x[n] \cdot \omega[n - mR] \tag{1}$$

其中，m 是當前窗的位置索引，R=256。$\omega(n)$ 代表 hanning window，數學式：

$$\omega[n] = \begin{cases} 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) & \text{if } n \in [0, N-1] \\ 0 & \text{o.w.} \end{cases} \tag{2}$$

其中，N=512。Hann 窗有助於降低信號邊緣的強度，避免因分段造成的不連續性和微小爆音，保持信號的平滑性和自然性。在本研究中，我們對標籤進行了加權處理，以更好地反映不同類別的重要性。爲了實現這一點，我們爲每個標籤分配了相應的加權係數。其中，N

（噪音）和 S（靜音）被賦予加權係數爲 0，而 B（爆音）和 V（正常語音）分別被賦予加權係數爲 0.3 和 1。因爲我們希望可以抑制會爆音的無聲子音，來達到減少遮蔽效應的問題，其數學式表示:

$$y[n] = \sum_{m=0}^{M-1} g_m x_m[n] \tag{3}$$

其中 M 是總窗口數，g_m 是第 m 個窗口的加權係數，數學式表示:

$$g_m = \begin{cases} 1 & \text{if label}_m = \text{V} \\ 0.3 & \text{if label}_m = \text{B} \\ 0 & \text{if label}_m = \text{N or label}_m = \text{S} \end{cases} \tag{4}$$

2.3 模型訓練

2.3.1 CNN 架構

本研究所使用的卷積神經網路（CNN）架構設計如 Table 2 所示：模型接受形狀爲 20×80×1 的梅爾頻譜圖像作爲輸入，並通過兩個卷積層進行特徵提取。第一個 2 維卷積層使用 8 個 5×5 的卷積核，第二個 2 維卷積層使用 10 個 5×5 的卷積核，並使用 ReLU 激活函數。每個卷積層後面接有 2×2 的最大池化層來進行下採樣，並在每個卷積層後應用 60% 的 Dropout 操作以減少過度擬合。在經過展平層（Flatten）後，輸入到一個包含 128 個神經元的全連接層，該層也應用 Dropout。最終的輸出層包含 4 個神經元，使用 Softmax 激活函數進行多類別分類。使用 Adam 來做爲 Optimizers，學習率設爲 0.0001，損失函數選擇分類交叉熵（categorical crossentropy），並通過準確率（accuracy）來評估性能。

Layer	Output Shape	Param#
Conv2D	(20, 80, 8)	208
MaxPolling2D	(10, 40, 8)	0
Dropout	(10, 40, 10)	0
Conv2D	(5, 40, 10)	2010
MaxPolling2D	(5, 20, 10)	0
Dropout	(5, 20, 10)	0
Flatten	(1000)	0
Dense	(128)	128128
Dropout	(128)	0
Dense	(4)	516

Table 2: CNN 模型架構

2.3.2 GRU 架構

本研究也使用門控循環單元（GRU）架構進行對比設計如 Table 3 所示：模型將輸入的形狀從 (20, 80, 1) 重新塑形爲 (20, 80)，以適應後續的處理步驟。隨後，應用了一層一維卷積層（Conv1D），此層使用 128 個濾波器，核大小爲 3。接著，模型包含了三層 GRU 層，均設置爲 64 個單元，並加入了 60% 的 Dropout 層，以防止過度擬合。此後，輸出經過展平層（Flatten），並傳遞至一個包含 128 個神經元的全連接層，再次加入 60% 的 Dropout 層。最後，通過一個 Softmax 激活函數的輸出層，將輸入數據分類爲四個類別。爲了最佳化模型，我們選擇了 Adam 來做爲 Optimizers，設置學習率爲 0.00001。損失函數則採用了多類別交叉熵損失（categorical crossentropy），並將模型的評估指標設定爲準確率。

Layer	Output Shape	Param#
Reshape	(20, 80)	0
Conv1D	(20, 128)	30848
GRU	(20, 64)	37248
Dropout	(20, 64)	0
GRU	(20, 64)	24960
Dropout	(20, 64)	0
GRU	(64)	24960
Dropout	(64)	0
Flatten	(64)	0
Dense	(128)	8320
Dropout	(128)	0
Dense	(4)	516

Table 3: GRU 模型架構

3 實驗結果及討論

3.1 模型評估方法

爲了評估建構的模型性能，本研究我們採用了混淆矩陣（Confusion Matrix）作爲主要的評估工具並且計算了準確率（Accuracy）。混淆矩陣是一種常用的分類模型評估方法，拿預估結果（Predicted）和答案（True）進行對比，並計算出每個 label 各自的轉換率，這樣可以提供關於模型預測準確性和錯誤類型的詳細資訊。我們是多類別的分類模型，所以 Accuracy 爲:

$$\text{Accuracy} = \frac{\text{正確分類的樣本}}{\text{全部樣本}} \tag{5}$$

我們也計算了「個別」類別的 precision，數學式如下:

$$\text{Precision of class A} = \frac{\text{TP}_{\text{classA}}}{\text{TP}_{\text{classA}} + \text{FP}_{\text{classA}}} \tag{6}$$

3.2 實驗結果

實驗結果我們採取主觀層面跟客觀層面的分析。在客觀層面，是測試由食道語者提供的 105 句語音音檔，總計 18 分鐘左右，並使用我們訓練的模型輔助切割後並使用 Praat 軟體進行調整來得到各類別的切割位置，並將資料拆成 train 85 句、dev 10 句、test 10 句三個集合分辨用於訓練、評估、測試模型。在研究中，爲了證明 VoiceBank-2023 語料庫在語句數量較少的情況下，通過有順序地累積語句，可以更迅速且有效地涵蓋中文中的聲母和韻母，我們通過實驗比較了隨機選取和有序選取的訓練狀況。資料隨機選取的詳細資料如 Table 4 所示。

Sets	Duration	#Sentences	#Syllables
train	14	85	6393
dev	2	10	913
test	2	10	916

Table 4: 隨機選取資料集，食道語者各資料集音檔長度（分鐘）、句數、音節數

此外，按順序選取的資料集詳細資料如 Table 5 所示。

Sets	Duration	#Sentences	#Syllables
train	14	85	6654
dev	2	10	739
test	2	10	829

Table 5: 有序選取資料集，食道語者各資料集音檔長度（分鐘）、句數、音節數

在隨機選取的狀況下，使用 CNN 模型進行訓練後，我們獲得了如下的最佳結果如 Table 6 所示。

P\T	B	V	S	N	#labels
B	0.872	0.119	0.001	0.007	838
V	0.001	0.932	0.031	0.058	5625
S	0.001	0.024	0.915	0.058	2375
N	0.003	0.062	0.134	0.800	1866
Acc.	0.9005				

Table 6: Confusion matrix for CNN model trained with random-ordered corpus.

此外，我們也測試了 GRU 模型作爲對照，以評估不同模型架構在相同語料配置下的性能表現。經過相同的隨機選取和訓練過程，使用 GRU 模型的最佳結果如 Table 7 所示。

P\T	B	V	S	N	#labels
B	0.896	0.095	0.000	0.001	838
V	0.017	0.952	0.018	0.012	5625
S	0.001	0.049	0.858	0.092	2375
N	0.004	0.012	0.007	0.800	1866
Acc.	0.9003				

Table 7: Confusion matrix for GRU model trained with random-ordered corpus.

在按序選取的狀況下，CNN 模型的最佳結果如 Table 8 所示。

P\T	B	V	S	N	#labels
B	0.920	0.069	0.000	0.013	578
V	0.023	0.941	0.017	0.019	5567
S	0.002	0.031	0.888	0.078	2122
N	0.002	0.043	0.070	0.883	1482
Acc.	0.9195				

Table 8: Confusion matrix for CNN model trained with ordered corpus.

GRU 模型在按序選取的狀況下的最佳結果如 Table 9 所示。

P\T	B	V	S	N	#labels
B	0.927	0.055	0.002	0.016	578
V	0.025	0.932	0.015	0.027	5567
S	0.001	0.033	0.879	0.086	2122
N	0.003	0.023	0.076	0.898	1482
Acc.	0.9150				

Table 9: Confusion matrix for GRU model trained with ordered corpus.

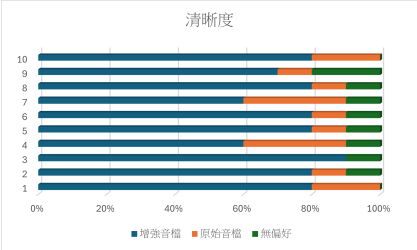
根據上述實驗結果，我們發現隨機選取語料會降低模型的準確度。因此，我們選擇了按照順序選取語料的策略，並對兩個模型的準確度和相關參數進行了詳細比較。以下是比較結果：

	CNN	GRU
B	0.920	**0.927**
V	**0.941**	0.932
S	**0.888**	0.879
N	0.883	**0.898**
Acc.	**0.9195**	0.9150
Network parameters	130k	127k

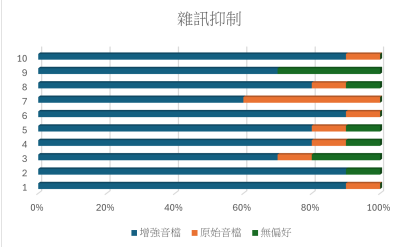
Table 10: Comparison of the CNN and GRU in the precisions of target labels, accuracies, and number of parameters.

在主觀層面的分析中，我們進行了聽感測

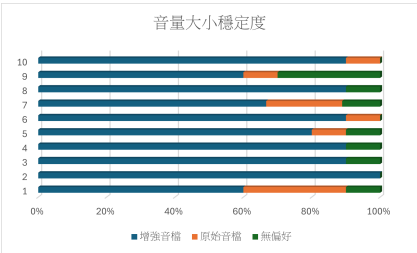
試。具體而言，我們將經過模型分類和語音增強處理後產生的音檔進行隨機排序，然後讓 10 位測試者試聽各 10 個原始音檔及增強後的音檔。測試者被要求在成對的音檔樣本中選擇一個他們認爲較優的樣本。如果測試者無法分辨兩個音檔之間的差異，他們可以選擇“無偏好”。我們根據三個方面進行詢問，進而得到 Figure 3 的結果。



(a) 清晰度



(b) 雜訊抑制



(c) 音量大小穩定

Figure 3: 三種評估指標分別是 (a) 清晰度、(b) 雜訊抑制、(c) 音量大小穩定

根據 Figure 3，我們可以觀察到大多數測試結果顯示，增強後的音檔在音量穩定性方面有顯著改善。然而，少數音檔在清晰度和雜訊抑制上表現未達預期，特別是某些音檔仍然存在雜訊或聲音斷斷續續的情況。儘管如此，整體結果仍顯示出正向的改善。

3.3 實驗討論

從實驗結果可以觀察到，使用 VoiceBank-2023 語料庫時，有序選取語料對模型準確度的提升非常顯著。

對於有序選取語料的情況，CNN 模型在正常語音（V）的準確率相對較高。部分辨認失敗的原因在於，雖然被歸類在正常語音（V）中的無聲子音不常出現爆音，但在極少數情況下仍有可能出現爆音。此外，可能由於類似遮蔽效應的存在，母音有時會被子音遮蔽，這可能導致在頻譜轉換後，母音被誤辨認爲雜音或靜音。儘管模型在大多數情況下能正確辨認正常語音（V），但仍需進一步改進以提高在此類情況下的辨識精度。至於靜音與雜音的辨認率，CNN 模型均接近九成，誤辨率主要表現爲雜音被辨認爲靜音，或靜音被辨認爲雜音。然而，由於我們將雜音與靜音的加權係數設爲 0，因此這些誤辨率對最終結果並不會造成實質性影響，即使在某些情況下發生誤辨識，也不會顯著影響整體性能。另一方面，GRU 模型在辨識爆音（B）方面有明顯更好的效果，這是因爲 GRU 可以利用其記憶單元捕捉前後文的信息，從而更好地理解爆音（B）在不同語境中的表現。這對於提高子音分類的準確度非常重要，也是 GRU 在子音辨識結果上表現良好的原因。相比之下，CNN 通過卷積提取局部特徵，這些特徵可能不足以捕捉爆音（B）的快速變化和短時動態，導致效果不如正常語音（V）。

至於根據聽覺測試結果，我們可以看到語音增強處理在不同的指標上有著不同的效果。在清晰度方面，大多數測試音檔的結果是增強後的音檔被受測者認爲更清晰，這表明增強處理在改善清晰度方面總體上是有效的。然而，在某幾個音檔約有 20% 到 30% 的受測者在清晰度評估中選擇了原始音檔，這可能是因爲增強過程中把一些被認爲是雜訊的正常語音 (V) 給濾掉，使得音檔的聽感變得斷斷續續，從而影響了清晰度的感受。在雜訊抑制方面，雖然大部分音檔經過增強後的效果是正面的，但有一個音檔卻有 40% 的受測者選擇了原始版本，這說明這個特定音檔在增強過程中可能未能有效地抑制雜訊，反而在某種程度上保留了原始雜訊。至於音量大小的穩定性，大多數受測者更傾向於選擇增強後的版本，這表明在這一指標上，增強處理成功地平衡了音量，使得音檔的音量聽起來更穩定。

總體而言，從三個方向來看，增強處理的效果都是正面的。特別是在音量大小的穩定性方面，處理得非常成功，大多數受測者都選擇了增強後的音檔，顯示出增強技術在平衡音量方面的顯著優勢。雖然在清晰度上有少數音檔的處理效果不如預期，但整體上仍然能是正面的。雜訊抑制方面也顯示出良好的效果，但某

些音檔在這方面的表現略顯不足，提示我們在未來的改進中需要更加關注這一指標。總的來說，這些結果表明增強技術在大多數情況下能有效提升音質，尤其是在音量穩定性上，未來可以進一步優化處理過程以提高清晰度和雜訊抑制的效果。

4 結論

在本研究中，根據上述實驗結果，考量到相對參數下計算量跟的訓練跟準度，我們最後選擇使用卷積神經網路（CNN）來識別食道語中的各種聲音成分，包括爆音、正常語音、靜音和雜音。實驗結果顯示，該模型在各類聲音成分的識別上均表現出色，在會爆音 (B) 跟正常語音 (V) 識別方面達到了高準確率。正常語音 (V) 的準確率相對較高，部分辨認失敗的原因在於有些無聲子音偶爾出現的爆音以及遮蔽效應的影響，導致母音有時被誤認爲雜音或靜音。靜音與雜音的辨認率接近九成，誤辨主要集中在雜音和靜音之間的相互誤認，這對最終結果影響不大。爆音的準確率達到了 92%，顯示了該模型在處理子音方面的強大能力。總體來說，該系統在辨認爆音、正常語音、靜音及雜音方面均表現良好，低計算量跟準確度也證明了其可行性和可靠性。未來的工作可以集中在進一步優化模型，特別是針對遮蔽效應和少數爆音情況進行改進，以提升系統的整體性能和準確性。此外，擴充語者和語句的多樣性也是一個重要方向，這樣可以提高模型在不同情境下的穩健性和適應性。通过這些措施，可以進一步提高語音增強技術的有效性和廣泛適用性。

這項研究除了開發在小型的發聲輔助裝置外，或許可以應用在一般麥克風上來消除雜音與爆音，甚至部分還可以結合手機或網頁開發成一個應用軟體，使人與人在用該軟體交流時可以避免突然有雜訊或爆音的情況發生。

References

Chen-Yu Chen, Wei-Zhong Zheng, Syu-Siang Wang, Yu Tsao, Pei-Chun Li, and Ying-Hui Lai. 2020. Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system. In *Interspeech*, pages 4686–4690.

Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. 2010. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. *IEICE TRANSACTIONS on Information and Systems*, 93(9):2472–2482.

Kadria Ezzine, Joseph Di Martino, and Mondher Frikha. 2022. Intelligibility improvement of esophageal speech using sequence-to-sequence voice conversion with auditory attention. *Applied Sciences*, 12(14):7062.

Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. 2017. The electrolarynx: voice restoration after total laryngectomy. *Medical devices: evidence and research*, pages 133–140.

Matsui Kenji, Hara Noriyo, Kobayashi Noriko, and Hirose Hajime. 2002. Enhancement of esophageal speech using formant synthesis. *Acoustical Science and Technology*, 23(2):69–76.

Luis Serrano, Sneha Raman, David Tavarez, Eva Navas, and Inma Hernaez. 2019. Parallel vs. non-parallel voice conversion for esophageal speech. In *INTERSPEECH*, pages 4549–4553.

Jia-Jyu Su, Pang-Chen Liao, Yen-Ting Lin, Wu-Hao Li, Guan-Ting Liou, Cheng-Che Kao, Wei-Cheng Chen, Jen-Chieh Chiang, Wen-Yang Chang, Pin-Han Lin, et al. 2023. Voicebank-2023: A multi-speaker mandarin speech corpus for constructing personalized tts systems for the speech impaired. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Kenzo Yamamoto, Tomoki Toda, Hironori Doi, Hiroshi Saruwatari, and Kiyohiro Shikano. 2012. Statistical approach to voice quality control in esophageal speech enhancement. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4500. IEEE.