藉由壓縮性之頻譜損失函數以學習 DEMUCS 語音強化模型之初步研究
Exploiting the compressed spectral loss for the learning of the DEMUCS speech enhancement network

戴麒恩
Chi-En Dai
暨南大學電機系
National Chi Nan University
s108323060@mail1.ncnu.edu.tw

洪啟瑋
Qi-Wei Hong
暨南大學電機系
National Chi Nan University
s108323024@mail1.ncnu.edu.tw

洪志偉
Jeih-Weih Hung
暨南大學電機系
National Chi Nan University
jwhung@ncnu.edu.tw

摘要

本研究針對著名的 DEMUCS 語音強化模型、藉由修改其訓練時所需的損失函數，來提升其效能。DEMUCS 由 Facebook 團隊開發，主要由卷積層組成其編碼模組與解碼模組，而兩模組之間則以長短時記憶模型來對編碼模組之輸出加以分解或降噪。雖然 DEMUCS 是一個純時域處理的語音強化架構，其訓練所使用的損失函數，卻同時涵蓋了時域和頻域的特徵，其中頻域上的特徵即為訊號經短時間傅立葉轉換所得的頻譜。

我們探討當 DEMUCS 之損失函數中的頻譜其強度值做壓縮時，對於所訓練而得的模型其效能是否有明顯的改變，我們採用的壓縮運算主要是對頻譜強度取一個小於一的正冪次方值，或對頻譜強度取其對數值。

當在 VoiceBank-DEMAND 之資料集上進行評估實驗時，初步結果表明，上述之壓縮運算為取正冪次方值時，其損失函數能使所學習的 DEMUCS 模型比原 DEMUCS 模型更有效地提升測試語音的客觀品質與可讀性指標(PESQ 與 STOI)，充分顯示引入次方壓縮性的頻譜強度於損失函數中能得到語音強化效能更佳的 DEMUCS 模型。相較而言，當壓縮運算為對數函數時，則沒有改進的效果。

Abstract

This study aims to improve a highly effective speech enhancement technique, DEMUCS, by revising the respective loss function in learning. DEMUCS, developed by Facebook Team, is built on the Wave-U-Net and consists of convolutional layer encoding and decoding blocks with an LSTM layer in between. Although DEMUCS processes the input speech utterance purely in the time (wave) domain, the applied loss function consists of wave-domain L1 distance and multi-scale short-time-Fourier-transform (STFT) loss. That is, both time- and frequency-domain features are taken into consideration in the learning of DEMUCS.

In this study, we present revising the STFT loss in DEMUCS by employing the compressed magnitude spectrogram. The compression is done by either the power-law operation with a positive exponent less than one, or the logarithmic operation.

We evaluate the presented novel framework on the VoiceBank-DEMAND database and task. The preliminary experimental results suggest that DEMUCS containing the power-law compressed magnitude spectral loss outperforms the original DEMUCS by providing the test utterances with higher objective quality and intelligibility scores (PESQ and STOI). Relatively, the logarithm compressed magnitude spectral loss does not benefit DEMUCS. Therefore, we reveal that DEMUCS can be further improved by properly revising the STFT terms of its loss function.

關鍵字：語音強化、DEMUCS、短時傅立葉轉換、損失函 數、壓縮頻譜損失、對數頻譜距離、感知語音品質、短時語音可讀性

Keywords: speech enhancement, DEMUCS, STFT, loss function, compressed spectral loss, logarithmic spectral distance, PESQ, STOI

1 簡介

語音強化 (speech enhancement, SE) 目的在於抑制語音中的加成性雜訊、通道干擾或混響、以優化語音的品質或可讀性。現今，由於語音分析技術的高度發展，逐漸落實於智慧生活的諸多應用，例如手機的語音輸入、語音檢索、語音操控機器人、助聽器等，但雜訊的存在使這些應用或設備無法精準捕捉原始乾淨語音之資訊、進而限制了語音技術在實用性上的擴展，因此直接在接收端抑制雜訊的語音強化技術相當重要性。

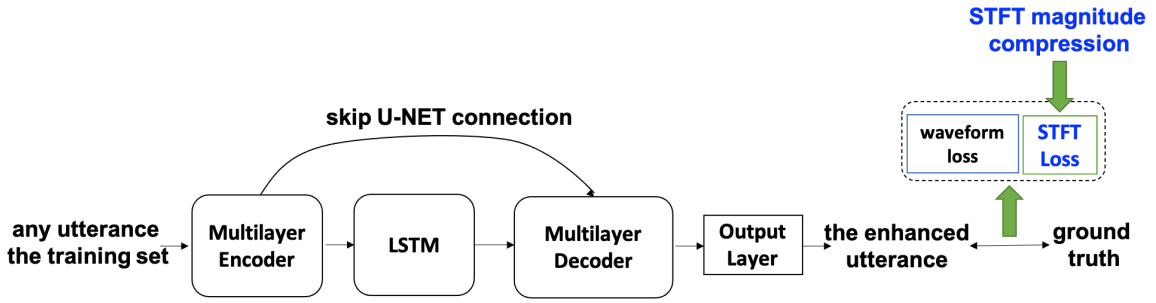
經典的語音強化技術通常是基於語音或雜訊干擾的統計特性以建構模型，如頻譜消去法(spectral subtraction) (Boll, 1979)、維納濾波法(Wiener filter) (Scalart and Filho, 1996)、卡爾曼濾波法(Kalman filter) (Dionelis and Brookes, 2018)等，為了在處理上減少延遲，它們所擷取的訊號通常較短、其統計特性無法精準呈現，進而限制了這些技術的效能，且其在非穩態的雜訊環境中表現通常較差。

近年來由於深度類神經網路(deep neural network, DNN) (Goodfellow et al., 2016; Hinton et al., 2012)之理論及技術的高度發展，基於 DNN 的語音強化技術也如雨後春筍地不斷湧現，且其在非穩態雜訊場影中的強化效能通常優於經典之基於統計特性的方法。一般來說，基於 DNN 之語音強化法，根據其對於輸入語音之特徵擷取的角度，大致可分為兩類：時頻域(time-frequency domain, T-F domain)法和端到端(end-to-end)的時域(time domain)法(Luo and Mesgarani, 2019; Yin et al., 2020)，二類型的方法在不同的雜訊場景或語音資料的表現各有優劣。時頻域法通常藉由短時傅立葉轉換(short-time Fourier transform, STFT) 對輸入語音創建時頻圖(spectrogram)，亦稱時頻域(T-F)特徵，在對這些 T-F 特徵加以消噪或強化後，進而用反短時離散傅立葉轉換(inverse STFT)重建原始的時域訊號波形(waveform)。可是使用 STFT 建構的時頻域特徵可能存在缺點或限制：首先，STFT 是一種固定的訊號轉換法（使用固定的弦波基底），相較於時域法其基底是由資料學習而得，時頻域法未必較佳。其次，時頻法對於乾淨語音之頻譜的相位成分通常無法精準的重建。

在諸多時域的語音強化法中，由 Facebook 團隊所研發的 DEMUCS 法(Defossez et al., 2019; Defossez et al., 2020)（全名：Deep Extractor for Music Sources with extra unlabeled data remixed）效能相當優異且廣為使用，它原本的目標是對於不同音源的混合分離成個音源，但也可以直接用於語音強化上（相當於把雜訊看成與乾淨語音不同的另一個音源）。它主要是由卷積層搭配跳躍連接(skip connection)的 U-net 模組來組成其編碼器-解碼器架構，而其訓練所使用的損失函數則主要包含了兩項：時域（或稱波域：wave domain）上的損失及時頻域上的損失，後者之時頻域所用特徵即為 STFT 求得的時頻圖。

諸多文獻(Ephraim and Malah, 1985; Yu and Deng, 2015)提到，人耳對於語音強度之感知程度(perceptual nature)是近似於一對數函數，亦即人耳感知的語音強度的提升程度不及真實語音強度的提升程度，可謂是人耳自我保護的機制、避免為太大聲的語音傷害。在語音強化與強健語音辨識的領域，許多方法(Lee et al., 2018; Braun and Ivan Tashev, 2021)皆利用這個特點、將語音頻譜之強度取對數或取小於一的正次方值，壓縮其動態範圍(dynamic range)，多數因而得到更佳的效果。在近期的研究(Hong et al., 2022; Wu et al., 2022)中，也提出了直接使用客觀語音品質指標(PESQ) (Ruder, 2017)與語音可讀性指標(STOI) (V-Botinhao et al., 2016)來監測 DEMUCS 在訓練上的收斂程度，而使 DEMUCS 達到適度提升的效果。

基於上述觀察，本研究提出：將 DEMUCS 訓練時在損失函數中所使用的短時傅立葉轉換求得的各個音框頻譜強度做動態壓縮，探討此壓縮的運算是否能使訓練而得的 DEMUCS 法在語音強化上的表現更佳，整體示意圖如圖一所示。我們分別使用前述的取冪次方值與取對數的壓縮法。值得一提的是，相較於最近的研究(Hong et al., 2022; Wu et al., 2022)、語音品質與可讀性指標只用於監測 DEMUCS 的收斂程度、而沒有實際加入其損失函數求其梯度來更新模型參數，這裡所提的頻譜強度動態壓縮法則藉由更動 DEMUCS 的損失函數直接參與其模型參數的訓練。而當上述的更新法藉由 VoiceBank-DEMAND 資料集來實現評估時，我們初步發現此更新版的 DEMUCS 相較於原 DEMUCS 能夠達到更佳的客觀語音強化指標、亦即 PESQ 與 STOI 值。



圖一: DEMUCS 訓練流程、及本文所提之處理 STFT 頻譜強度的壓縮用於損失函數

在之後的章節中，我們將先簡要介紹 DEMUCS 法的流程，接著陳述我們提出之更動損失函數的新方法，隨後是實驗環境介紹、實驗結果呈現分析與討論，最後則是結論與未來展望。

2 DEMUCS 法簡要介紹

最初，DEMUCS 是為多音源分離而設計。它主要是由卷積層搭配跳躍連接(skip connection)的 U-net 模組來組成其編碼器-解碼器架構，其中搭配了長短時記憶模型(LSTM)用於修改編碼器輸出來分離音源，當 DEMUCS 應用於單聲道語音強化時，我們將 DEMUCS 的輸入與輸出通道數設置為 1、直接以乾淨語音為逼近的真實目標(ground truth)。關於 DEMUCS 模型安排的細節，可參考文獻(Defossez et al., 2019; Defossez et al., 2020)。

在原始 DEMUCS 網路模型的學習中，用以最小化的損失函數包括了在（時域）波形上的 L1 損失和（時頻域）多解析度短時傅立葉轉換(multi-resolution STFT)之頻譜損失。對任一其長度（點數）為T的輸入訊號，其增強後的時域訊號$\boldsymbol{y}$和原始乾淨的語音信號$\tilde{\boldsymbol{y}}$之間的損失函數表示為：

$$L_{DEMUCS}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|_1 + \sum_i L_{stft}^{(i)}(\boldsymbol{y},\tilde{\boldsymbol{y}}), \quad (1)$$

其中$\|.\|_1$是 L1 範數運算，而式子右邊第一項是時域波形上的 L1 損失，第二項是多解析度時頻域上的損失和，下標 "$stft$"為短時間傅立葉轉換(STFT)，上標"(i)"是對應不同音框長度(frame size)與音框平移(frame shift)的索引(index)。此外，第二項中個別索引(i)項的$L_{stft}^{(i)}(\boldsymbol{y},\tilde{\boldsymbol{y}})$又包含兩部分：頻譜收斂(spectral convergence, 以下標sc表示)之損失$L_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}})$和頻譜強度(spectral magnitude, 以下標mag表示)之損失$L_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}})$，呈現如下式：

$$L_{stft}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = L_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) + L_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}), \quad (2)$$

其中

$$L_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{\|\,|STFT(\boldsymbol{y})| - |STFT(\tilde{\boldsymbol{y}})|\,\|_F}{\|STFT(\boldsymbol{y})\|_F}, \quad (3)$$

而

$$L_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\left\|\log\left(\frac{|STFT(\boldsymbol{y})|}{|STFT(\tilde{\boldsymbol{y}})|}\right)\right\|_1, \quad (4)$$

其中$STFT(.)$是求取音框之頻譜的短時間傅立葉轉換(short-time Fourier transform), $|.|$是逐項(element-wise)取絕對值的運算，$\|.\|_F$是 Frobenius 範數運算。

3 提出之新方法

DEMUCS 網路訓練中使用的損失函數綜合衡量了強化語音與其原始乾淨語音在時域與頻域上的差異。一般而言，修改損失函數的做法有兩種：第一種是在原損失函數上加上其他的損失函數並取加權，而第二種是對原始損失函數中的個別項目作修改，我們採取的是第二種，其相對第一種修改法的可能好處是在訓練中求取損失函數之梯度時，運算複雜度的增加幅度較小。

我們具體的作法是將 DEMUCS 原損失函數中使用的短時傅立葉轉換求得的音框頻譜的強度做壓縮(compression)，使其動態範圍(dynamic range)降低，這樣的潛在優點是：一方面可模擬人耳對於音量強度的感知效應（感知強度的變化低於真實強度的變化），另一方面也有助於避免訓練模型時，損失函數之梯度下降值過大導致不易收斂至最佳值的問題(Ruder, 2017)。

參照文獻 (Braun and Ivan Tashev, 2021)，我們使用兩種強度壓縮的運算。第一種是乘冪運算，且使用的冪次方值為小於一的正數，即 $f(x)=x^r, 0<r<1$。第二種則是使用 log1p 函數，即 $f(x)=\text{log1p}(x)=\log(1+x)$，其中取對數前加一的目的是使函數的最小值為 0，而非負無窮大（當 $x \geq 0$ 時）。

根據使用乘冪運算的壓縮法，我們將原(3)與(4)式修改為：

$$\hat{L}_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{\| |STFT(\boldsymbol{y})|^r - |STFT(\tilde{\boldsymbol{y}})|^r \|_F}{\| |STFT(\boldsymbol{y})|^r \|_F}, \tag{5}$$

$$\hat{L}_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\left\| \log\left(\frac{|STFT(\boldsymbol{y})|^r}{|STFT(\tilde{\boldsymbol{y}})|^r} \right) \right\|_1, \tag{6}$$

其中的次方值 $0<r<1$。值得注意的是，雖然原式(3)中的 $STFT(\boldsymbol{y})$ 項是複數陣列，但它的 Frobenius 範數 $\|STFT(\boldsymbol{y})\|_F$ 只和 $STFT(\boldsymbol{y})$ 其中每項的絕對值 $|STFT(\tilde{\boldsymbol{y}})|$ 有關，亦即 $\|STFT(\boldsymbol{y})\|_F = \||STFT(\boldsymbol{y})|\|_F$，因此，式(5)中的分母項可以直接表達成 $\||STFT(\boldsymbol{y})|^r\|_F$。

同理，對於第二種使用 $\text{log1p}\ (x)=\log(x+1)$ 函數的壓縮法，我們將原式(3)與式(4)修改為：

$$\hat{L}_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{\|\text{log1p}\ |STFT(\boldsymbol{y})|) - \text{log1p}\ (|STFT(\tilde{\boldsymbol{y}})|)\|_F}{\|\text{log1p}\ (|STFT(\boldsymbol{y})|)\|_F}, \tag{7}$$

$$\hat{L}_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\left\| \log\left(\frac{\text{log1p}(|STFT(\boldsymbol{y})|)}{\text{log1p}(|STFT(\tilde{\boldsymbol{y}})|)} \right) \right\|_1, \tag{8}$$

在更動式(3)(4)、如式(5)(6)或是式(7)(8)，我們將式(2) 對應的時頻域損失更新為：

$$\hat{L}_{stft}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \hat{L}_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}}) + \hat{L}_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}}), \tag{9}$$

其中 $\hat{L}_{sc}(\boldsymbol{y},\tilde{\boldsymbol{y}})$ 與 $\hat{L}_{mag}(\boldsymbol{y},\tilde{\boldsymbol{y}})$ 分別如式(5)(6)、或是式(7)(8)所示，則 DEMUCS 使用之整體的損失函數可以更動為：

$$\hat{L}_{DEMUCS}(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \frac{1}{T}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|_1 + \sum_i \hat{L}^{(i)}_{stft}(\boldsymbol{y},\tilde{\boldsymbol{y}}), \tag{10}$$

其中更動的 $\hat{L}^{(i)}_{stft}(\boldsymbol{y},\tilde{\boldsymbol{y}})$ 項來自式(9)。

4 實驗設定

我們使用 VoiceBank-DEMAND 語料庫(V-Botinhao et al., 2016; Thiemann et al., 2013)來評估我們所提的新方法、對於相對應的 DEMUCS 模型進行訓練與測試，所使用之乾淨無干擾的語句和雜訊分別來自 VoiceBank(V-Botinhao et al., 2016) 和 DEMAND (Thiemann et al., 2013)語料庫。訓練集包含了 28 個語者所生成的 11,572 個語句，其中摻入 10 種雜訊、訊雜比(SNR)分別為 0, 5, 10 與 15 dB，測試集則包含 2 個語者所生成的 824 個語句、摻入與訓練集不同之另 5 種雜訊、訊雜比(SNR)分別為 2.5, 7.5, 12.5 與 17.5 dB。此外，我們從訓練集中挑出 742 個語句作為驗證集。

我們採用文獻(Defossez et al., 2020)所介紹之具因果性(causal)的 DEMUCS 架構，學習迭代次數(epoch)設為 300，批量設為 32，其他參數設置，分別是 $U=4$, $S=4$, $K=8$, $L=5$ 及 $H=48$。依照式(5)-(8)，我們分別訓練對應至不同損失函數之 DEMUCS 模型。在測試上，我們使用客觀語音品質指標 PESQ (Rix et al., 2001)和可讀性指標 STOI (Taal et al., 2011)來評估強化後的語音，PESQ 分數介於-0.5 與 4.5 之間，STOI 分數介於 0 與 1 之間，三者分數越高皆代表語音強化效能越好。

5 實驗結果與討論

我們探討使用頻譜強度壓縮對應之 DEMUCS，其中，乘冪運算其冪次方 r 值分別設為 0.1, 0.3, 0.5 與 0.7，另外，為了觀察頻譜強度壓縮的逆運算、亦即將頻譜強度伸展對 DEMUCS 帶來的效應，我們額外求得冪次方 $r=1.1$ 對應的 DEMUCS 模型。使用乘冪運算與對數函數 log1p 之頻譜強度壓縮之 DEMUCS 對測試集強化所對應的 PESQ 與 STOI 指標數據列於表一，從此表我們有了以下的觀察：

1. 未任何強化處理之基礎實驗結果都遠比理想值差(PESQ 上限為 4.5，STOI 上限為 1)，足見雜訊干擾對於語音的破壞程度，而這裡所有的 DEMUCS 法都能得到較佳的 PESQ 與 STOI 分數，可見它們確實都能有效強化語音。
2. 對於使用小於 1 之冪次方 r 的頻譜壓縮之 DEMUCS 而言，它們幾乎都比原始

DEMUCS ($r = 1$)得到更佳的PESQ與STOI分數，例如當$r = 0.3$時，PESQ與STOI值為3.006與0.949為最佳，超越原始DEMUCS的2.923與0.947，這與文獻[13]所得到的結果大致相符，其頻譜強度其壓縮冪次設為0.3時語音強化效能也是最好。我們也觀察到，當壓縮冪次r從原始的1逐漸變小至0.3時，PESQ都有顯著的提升、STOI則維持或是小幅提升。而r值從0.3降為0.1時，雖然其PESQ變差，但仍高達2.988，超越原始DEMUCS ($r = 1$)的2.923。這些結果充分顯示了我們所提出、在DEMUCS的損失函數使用冪次壓縮頻譜強度確實能提升DEMUCS的語音強化效能。

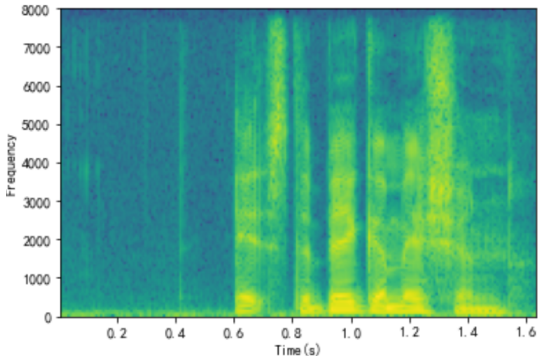
3. 當冪次值$r = 1.1$時，頻譜強度的動態範圍反而變大，實驗結果顯示，其反而使DEMUCS對應的PESQ分數下降，因此，延展頻譜強度的損失函數並無法改進DEMUCS。

若使用log1p函數來做頻譜強度壓縮時，其PESQ與STOI的分數都明顯變差，基於log1p函數與取冪次方皆有相似的壓縮輸出之動態範圍的效果，此結果令我們有些意外，可能的原因是當使用log1p函數時，在式(8)所呈現的頻譜強度損失連用了兩次的log函數、它可能對於頻譜強度造成過度壓縮而使對應的損失值失去鑑別力，在未來的研究裡，我們將對此問題進一步探討其解決方向。

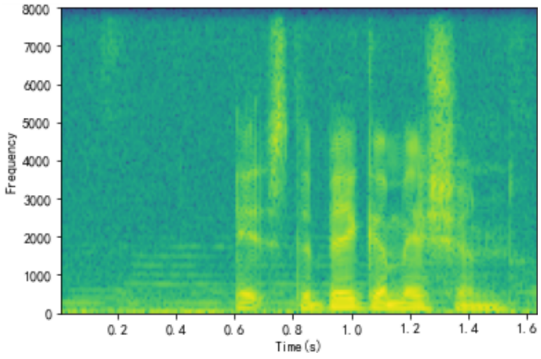
		PESQ	STOI
基礎實驗（未處理）		1.970	0.921
原始 DEMUCS ($r = 1$)		2.923	0.947
使用冪次壓縮頻譜強度之DEMUCS	$r = 0.1$	**2.988**	**0.948**
	$r = 0.3$	**3.006**	**0.949**
	$r = 0.5$	**2.979**	**0.948**
	$r = 0.7$	**2.974**	0.947
	$r = 1.1$	2.910*	0.947
使用 log1p 函數壓縮頻譜強度之 DEMUCS		2.735*	0.943*

表 1: 原始 DEMUCS 與各種壓縮頻譜強度之 DEMUCS 法對應的 PESQ 與 STOI 值

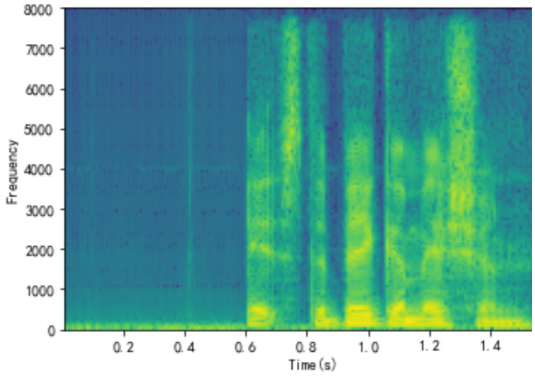
除了語音強化的相關評估數據外，我們也使用強度時頻圖(magnitude spectrogram)來驗證各方法在消噪上的效能，圖二(a)-(f)繪製了一VoiceBank的語音在各種環境下所得的強度時頻圖，首先，圖二(a)(b)分別對應乾淨環境與



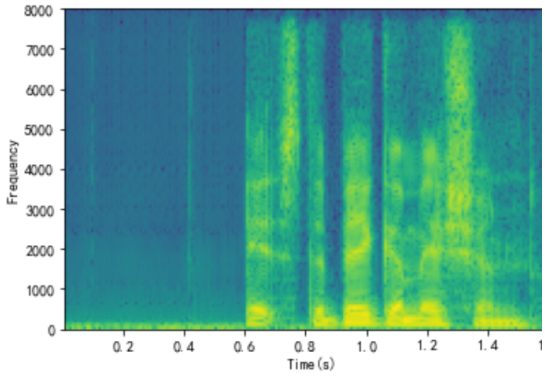
(a) clean



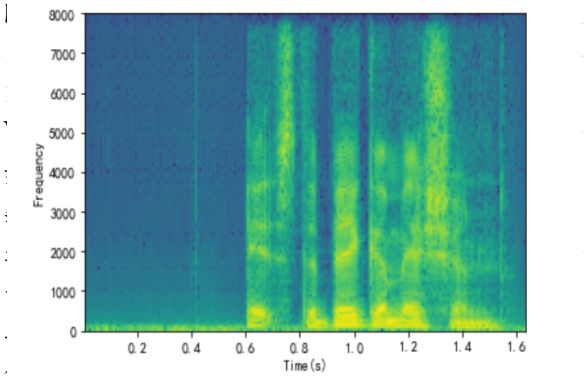
(b) noisy



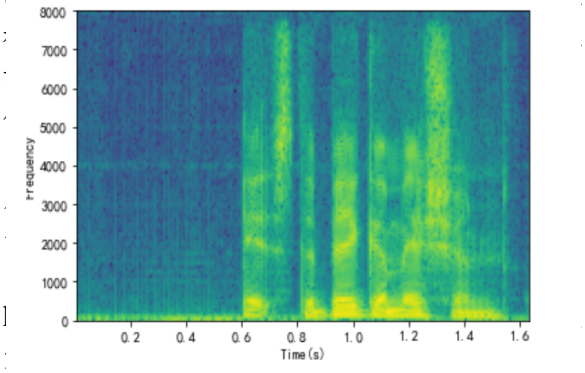
(c) enhanced by original DEMUCS



(d) enhanced by DEMUCS with power r = 0.3



(e) enhanced by DEMUCS with power r = 1.1



(f) enhanced by DEMUCS with log1p compression

圖二：一段語音在不同環境下的強度時頻圖：(a)乾淨、(b)雜訊干擾、(c) 雜訊干擾但使用原始 DEMUCS 強化、(d) 雜訊干擾但使用冪次為 0.3 作頻譜壓縮之 DEMUCS 強化、 (e) 雜訊干擾但使用冪次為 1.1 作頻譜壓縮之 DEMUCS 強化、(f) 雜訊干擾但使用 log1p 函數作頻譜壓縮之 DEMUCS 強化

雜訊環境，我們看到雜訊對於語音的時頻圖造成明顯的失真。其次，圖二(c)(d)為雜訊語音（圖二(b)）經過原始 DEMUCS 與冪次為 0.3 之頻譜壓縮之 DEMUCS 對應的時頻圖，將它們與圖二(a)(b)相比較，我們可看到雜訊被有效地抑制，然而似乎比圖(a)顯示的乾淨語音時頻圖更"乾淨"一些，意味著可能部分語音也被當作雜訊被消除或減低。圖二(e)(f)為雜訊語音（圖二(b)）經過冪次為 1.1 之頻譜延展與使用 log1p 函數壓縮之 DEMUCS 對應的時頻圖，它們似乎顯示不僅雜訊抑制、許多語音成分也被消除，例如在圖二(a)中 0.4 秒附近有一個短暫的語音頻譜成分，它在圖(f)中幾乎已經不見了。這呼應了之前的結果、即使用 log1p 函數做頻譜壓縮反而會造成 DEMUCS 效能的退步。

6. 結論與未來展望

此研究提出在訓練 DEMUCS 的損失函數時，預先對其 STFT 頻譜強度進行壓縮，藉此模擬人耳效應、提升其收斂穩定度以期提升 DEMUCS 語音強化的表現。我們藉由冪次方運算以及對數(log1p)運算分別來壓縮短時間頻譜強度。初步實驗結果顯示，當損失函數引用了冪次方運算壓縮之時頻圖，能使訓練而得的 DEMUCS 對輸入語音造就更佳的語音品質與可讀性，而對數運算的壓縮則表現不如預期。在未來的研究裡，我們會探討對數運算壓縮表現不佳的原因，同時，我們也將驗證上述的壓縮頻譜強度於其他類別的語音強化法是否也能發揮類似的效能、並進一步探究最佳化壓縮的方法。

參考文獻

Steven Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1979.

Pascal Scalart and Jozue VIEIRA Filho, “Speech enhancement based on a priori signal to noise estimation,” *in Proc. ICASSP*, 1996

Nikolaos Dionelis and Mike Brookes, "Speech Enhancement Using Kalman Filtering in the Logarithmic Bark Power Spectral Domain," *in Proc. EUSIPCO*, 2018

Ian Goodfellow, Yoshua Bengio and Aaron Courville, “Deep learning,” *MIT Press*, 2016

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl et al., “Deep neural networks for acoustic modeling in speech Recognition: the shared views of four research groups,” *IEEE Signal Processing Magazine*, 2012,

Yi Luo and Nima Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, 2019.

Dacheng Yin, Chong Luo, Zhiwei Hong and Wenjun Zeng, “Phasen: A phase-and-harmonics-aware speech enhancement network,” *in Proceedings AAAI*, 2020.

Alexandre Defossez, Nicloas Usunier, Leo’n Bottu and Francis Bach, “Music source separation in the waveform domain,” *arXiv:1911.13254*, 2019.

Alexandre Defossez, Gabriel Synnaeve and Yossi Adi, “Real time speech enhancement in the waveform domain,” *in Proc. Interspeech*, 2020

Yariv Ephraim, David Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985.

Dong Yu, Li Deng, "Automatic Speech Recognition: A Deep Learning Approach," *London: Springer*, 2015.

Jinkyu Lee, Jan Skoglund, Turaj Shabestary, Hong-Goo Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement,” *IEEE Signal Processing Letters*, 2018.

Sebastian Braun, Ivan Tashev, “A consolidated view of loss functions for supervised deep learning-based speech enhancement”, *in Proc. TSP*, 2021.

Qi-Wei Hong, Chi-En Dai, Hui-Chun Hsu, Zong-Tai Wu, Jeih-Weih Hung, "Leveraging the perceptual metric loss to improve the DEMUCS system in speech enhancement", *in Proc. ICASI*, 2022

Zong-Tai Wu; Yan-Tong Chen; Jeih-weih Hung, "Improving the performance of DEMUCS in speech enhancement with the perceptual metric loss", in Proc. ICCE-TW, 2022Sebastian Ruder, "An overview of gradient descent optimization algorithms", *arXiv:1609.04747v2,* 2017.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, Junichi Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust text-to-speech,” *in Proc. SSW*, 2016.

Joachim Thiemann, Nobutaka Ito, Emmanuel Vincen, “Demand: a collection of multi-channel recordings of acoustic noise in diverse environments,” *in Proc. ICA*, 2013

Antony W. Rix, John G. Beerends, Michael P. Hollier and Andries P. Hekstra, “Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs,” *in Proc. ICASSP*, 2001.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.