使用離散小波轉換特徵於 Conv-TasNet 語音強化模型的初步研究
A Preliminary Study of the Application of Discrete Wavelet Transform Features in Conv-TasNet Speech Enhancement Model

陳彥同
Yan-Tong Chen
暨南大學電機系
National Chi Nan University
s109323508@mail1.ncnu.edu.tw

吳宗泰
Zong-Tai Wu
暨南大學電機系
National Chi Nan University
s110323503@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

當前基於深度類神經網路架構之語音強化模型，常使用時域特徵來加以學習其模型參數，時域特徵如同經典的頻域特徵一般，能夠使所得模型達到優異的語音強化效果。基於此概念，本研究主要是探討如何從時域的語音中提取資訊、以在語音強化中創建更有效的特徵。我們提出了在時域中擷取短時間的子頻帶信號，並將它們融合成為單一特徵。具體方法是應用離散小波變換對每個輸入的音框信號進行分解、以獲得子頻帶信號，並對這些信號進行投影融合處理以創建最終小波域特徵。對應的融合處理法稱為雙投影融合(bi-projection fusion, BPF)法。同時，我們將藉由離散小波轉換之融合小波域特徵與原始時域特徵加以整合、來學習一高效的語音強化網路：全卷積時域音頻分離網路 (Conv-TasNet)，藉此來強化受雜訊干擾的語音訊號、提升其品質與可讀性。

我們在 VoiceBank-DEMAND 與 VoiceBank-QUT 兩個語音強化資料集上進行了評估實驗，初步結果表明，所提出的方法比原始單純使用時域特徵的 Conv-TasNet 實現了更高的客觀語音品質和可讀性指標，表明融合小波域特徵可以輔助原時域特徵、從輸入的雜訊語音中學習一個更有效的 Conv-TasNet 網路、達到更佳的語音強化效果。

Abstract

Nowadays, time-domain features have been widely used in speech enhancement (SE) networks like frequency-domain features to achieve excellent performance in eliminating noise from input utterances. This study primarily investigates how to extract information from time-domain utterances to create more effective features in speech enhancement. We present employing sub-signals dwelled in multiple acoustic frequency bands in time domain and integrating them into a unified feature set. We propose using the discrete wavelet transform (DWT) to decompose each input frame signal to obtain sub-band signals, and a projection fusion process is performed on these signals to create the ultimate features. The corresponding fusion strategy is the bi-projection fusion (BPF). In short, BPF exploits the sigmoid function to create ratio masks for two feature sources. The concatenation of fused DWT features and time features serves as the encoder output of a celebrated SE framework, fully-convolutional time-domain audio separation network (Conv-TasNet), to estimate the mask and then produce the enhanced time-domain utterances.

The evaluation experiments are conducted on the VoiceBank-DEMAND and VoiceBank-QUT tasks. The experimental results reveal that the proposed method achieves higher speech quality and intelligibility than the original Conv-TasNet that uses time features only, indicating that the fusion of DWT features created from the input utterances can benefit time features to learn a superior Conv-TasNet in speech enhancement.

關鍵字：語音強化、離散小波轉換、跨域、雙投影融合、全卷積時頻分離網路

Keywords: speech enhancement, discrete wavelet transform, cross-domain, temporal speech sequence, Conv-TasNet, bi-projection fusion

1 簡介

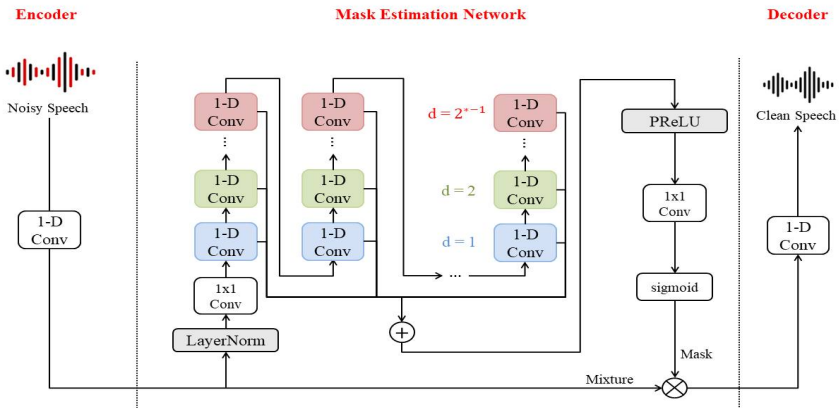
現今的語音處理技術已成功集成到智能 3C 設備、實現語音辨識、交互式語音聊天、智能機器人的語音指令控制、汽車或機車的語音控制等功能於眾多網路和多媒體視聽設備中。然而，在拓展語音相關之應用上，仍然存在許多關鍵性的挑戰。從信號處理的角度來看，雜訊(noise)干擾是信號傳輸和語音處理的首要問題之一。正如 2021 年暢銷書(Kahneman, 2021)中所述，雜訊(noise)和偏差(bias)是精準估測與決策上兩個主要的錯誤來源。然而，與偏差相比，雜訊的隨機性使其準確估測的可能性大大降低，從而加深了處理雜訊的難度。雖然這主要是對人類判斷的敘述，但它似乎同樣適用於基於機器之自動語音信號處理所面臨的處境。

在對應語音中的雜訊之各種議題中，語音強化 (speech enhancement, SE) (Philipos C. Loizou,2013)應可謂最直接的處理機制，其針對接收到的語音訊號加以處理、目標為提升輸出語音訊號的訊雜比(signal-to-noise ratio, SNR)，或改善語音的品質(quality)和可讀性(intelligibility)，使人們或機器更容易接受和理解語音訊號。傳統的語音強化演算法主要依賴於語音或雜訊的統計特性建構模型，然而它們在非穩態雜訊場景中的表現通常較差，主因之一在於非穩態的訊號特性較難以統計形式加以精準估測。

近十年來，由於深度學習和深度神經網路(deep neural networks, DNN)(Ian Goodfellow et al., 2016) 的理論與應用的飛快進步，學者和專家藉由 DNN 來建構語音強化的模型，與傳統基於統計的語音強化法相較，基於 DNN 的方法效果通常優異許多、特別是在非穩態的雜訊場景中。

在諸多基於 DNN 的語音強化法中，全摺積時域音訊分離網路法 (Conv-TasNet)(Yi Luo and Nima Mesgarani, 2019)相當著名且高效，因此廣為學者所探討並加以延伸。Conv-TasNet 採用編碼器串接解碼器(encoder-decoder)的架構，並應用堆疊的一維擴張卷積塊(dilated convolutional block)來執行其中的遮罩估測網路(mask estimation network)。原始的 Conv-TasNet 在編碼器端使用可訓練之一維卷積層來創建時域特徵，以作為編碼器輸出。而跨域時間卷積網路法(CD-TCN)(Fu-An Chao et al., 2019)則 進一步採用頻域特徵，結合時域特徵作為編碼器輸出。實驗結果表明，相較於使用單一時域的 Conv-TasNet，使用跨域特徵（時域與頻域）的CD-TCN達到更佳的語音強化效能。

在本研究中，我們參考了 CD-TCN 的方法脈絡，嘗試通過添加另一個特徵源來改進 Conv-TasNet，而這個特徵源來自輸入之音框訊號的離散小波變換 (discrete wavelet transform, DWT)(Stephane Mallat,2019)。我們使用 DWT 創建子頻帶特徵，然後將這些子頻帶特徵加以融合(fusion)、成為小波域特徵，最後將此小波域特徵與時域特徵加以串接、作為 Conv-TasNet 的編碼器輸出，我們參照了 CD-TCN 的方法，將其二元投影融合法(bi-projection fusion, BPF)用於小波子頻帶特徵的融合上。



圖一. 原始 Conv-TasNet 的流程圖（編碼端使用時域特徵）

我們將所提之新方法在 VoiceBank-DEMAND 和 VoiceBank-QUT 數據集上進行評估，初步實驗結果表明，所提出之新型跨域（小波域與時域）的 Conv-TasNet 模型在客觀語音品質和語音可讀性指標上，都呈現了優異的語音強化性能、優於單一時域的 Conv-TasNet 法，另外，藉由與 Conv-TasNet 架構類似、但於遮罩估測網路更細緻的 DPTNet (Jingjing Chen, 2020)模型之評估，我們也驗證了所提之小波域特徵的優越性。

2 提出之新方法

原始使用時域特徵作為編碼器輸出的 Conv-TasNet 其流程圖如圖一所示。我們看到其包含了編碼器、遮罩估測網路與解碼器三部分。在這裡，我們主要是針對其編碼器加以變化，提出採用小波域特徵的求取法、並將小波域特徵與時域特徵相結合，作為此網路的編碼器特徵。 這個新方法主要包括以下步驟：

2.1 建立時域特徵與一階離散小波特徵

從單通道麥克風接收到的受雜訊干擾的語音信號 y[n]可以表示如下：

$$y[n] = h[n] * x[n] + d[n], \qquad (1)$$

其中$x[n]$是乾淨的語音信號，$h[n]$是對應於通道效應或混響的捲積性雜訊，$d[n]$則是加成性雜訊，而n是時間索引。在這裡，我們忽略卷積性雜訊$h[n]$，專注處理加成性雜訊$d[n]$所干擾之語音訊號，以重建乾淨語音信號$x[n]$為目的的語音強化模型。

根據語音其短時穩態的特性，我們將輸入語音信號$y[n]$切割成M個長度為L的音框訊號，各音框訊號以向量 $\mathbf{x}_k \in \mathbb{R}^{L\times 1}$表示，其中 k為音框索引(frame index)。 因此，我們將各音框之向量橫排、構成一個原始資料矩陣 $X \in \mathbb{R}^{L\times M}$ 。

(1) 時域特徵

我們將輸入之原始資料矩陣 X 通過可訓練之一維卷積層運算，使其原本為L維的行向量x_k轉換為 N 維向量，橫向串接後得到時域特徵矩陣$\boldsymbol{W_T} \in \mathbb{R}^{N\times M}$，公式如下：

$$W_T = H(UX), \qquad (2)$$

其中$U \in \mathbb{R}^{N\times L}$ 為N個時域轉換的編碼器基底(basis)向量直排而成，即為卷積層之 kernel 函數，H 是一個非線性函數，如 ReLU 函數，以確保輸出 W_T的每項都是大於或等於零。

(2) 一階離散小波特徵

首先，我們對矩陣X的每一個行向量$\mathbf{x}_k$執行一階離散小波變換，得到其近似項係數(approximation coefficients)與細節項係數(detail coefficients)，如下式所示：

$$\left[\mathbf{c}_k^A, \mathbf{c}_k^D\right] = \boldsymbol{DWT}(X), \qquad (3)$$

其中$\boldsymbol{DWT}(\cdot)$代表一階離散小波變換，$\mathbf{c}_k^A$和 $\mathbf{c}_k^D$分別是近似項係數和細節項係數，可以看作是原始序列$\mathbf{x}_k$的低通(lowpass)子頻帶和高通(highpass)子頻帶。二者之頻寬都大約等於原始序列頻寬的一半，而它們的點數減為$\mathbf{x}_k$點數的一半，即$\frac{L}{2}$。

我們將高通與低通之各音框子頻帶信號分別橫排再一起、產生兩個特徵矩陣C_A和C_D，大小為 $\frac{L}{2}\times K$，它們的行向量分別為$\mathbf{c}_k^A$和 $\mathbf{c}_k^D$。之後，我們使用一維可訓練卷積層（連同非線性函數H）進一步處理C_A和C_D，以生成大小為$N \times M$的兩個矩陣W_A和W_D，兩者與時域特徵矩陣W_T大小相同。 其運算公式為：

$$W_A = H(U_A C_A),\ \ W_D = H(U_D C_D), \qquad (4)$$

其中U_A和U_D分別表示C_A和C_D對應的一維卷積層運算矩陣。

2.2 彙整時域特徵和小波域特徵

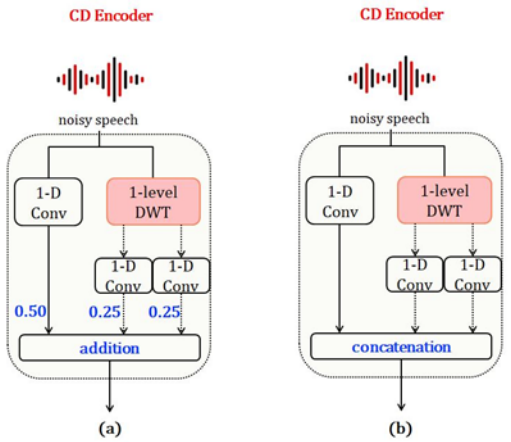
到目前為止，我們有三個特徵矩陣：W_T（時域特徵）、W_A（小波域低頻特徵）和W_D（小波域高頻特徵）。它們的尺寸相同。在這裡，我們提出了三種方法來彙整它們，以建構最終的編碼器輸出矩陣W_E，它們的流程圖分別繪製於圖二(a)、圖二(b)與圖三(b)。三個方法分述如下：

相加 (addition)

首先，最直觀的彙整方式是將W_E設為三個矩陣的加權和：

$$W_E = 0.5W_T + 0.25W_A + 0.25W_D, \qquad (5)$$

此步驟由圖一(a)所示。由此可看出最終的特徵矩陣W_E與每個分量矩陣的尺寸一致。



圖二. 更新後的 Conv-TasNet 編碼器部分示意圖，通過(a) 相加，或 (b)串接來彙整時域和小波域特徵

串接 (concatenation)

另一種直觀的彙整方式是將三個矩陣橫向串接：

$$W_E = [W_T; W_A; W_D], \qquad (6)$$

此步驟由圖一(b)所示。由此可看出最終的特徵矩陣W_E相對於每個分量矩陣而言，項數變為 3 倍

先融合再串接 (fusion and concatenation)

為了更有效地提取與整合兩個小波域特徵W_A和W_D的資訊，我們利用文獻(Fu-En Wang, 2020)所使用的二元投影融合法(bi-projection fusion, BPF) 將二者相融合。 BPF 已被用於集成時域和頻域特徵，並在 CD-TCN 法(Fu-An Chao et al., 2019) 中有優異的表現。此外，語音訊號中的低通成分和高通成分對應至不同的資訊、受雜訊影響也可能不同。例如，低通特徵W_A通常對應更多母音的成分，而高通特徵W_D可能對應到子音。此外，在常見的雜訊干擾場景中，低通特徵W_A的訊雜比 (signal-to-noise ratio, SNR) 通常高於低通特徵W_D（因為語音的低頻成分能量通常較大）。因此，BPF 模塊的兩個互補性遮罩矩陣（各項非負且二矩陣相同位置的二項和為 1）應適合用於融合這兩個不同頻帶的特徵來強化語音。

使用 BPF 法融合兩特徵的步驟如下：首先，我們串接W_A與W_D作為一個可訓練之卷積層的輸入，求取一遮罩(mask) M:

$$M = \sigma(\Psi_M([W_A; W_D], \theta_M)), \qquad (7)$$

其中σ是 sigmoid 函數，Ψ_M是卷積投影層運算，其參數為θ_M。 接著，我們將M和$1-M$與W_A和W_D分別相點乘，進而相加以生成小波域融合特徵：

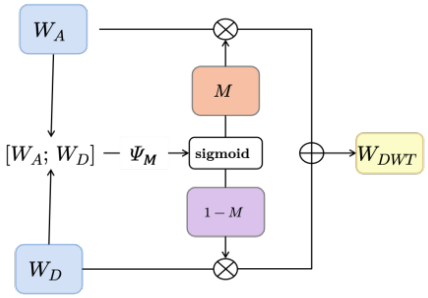
$$W_{DWT} = M \odot W_A + (1 - M) \odot W_D, \qquad (8)$$

其中 ⊙是點乘運算。

最後，我們橫向串接小波域融合 特徵和時域特徵以得到最終的特徵矩陣W_E：

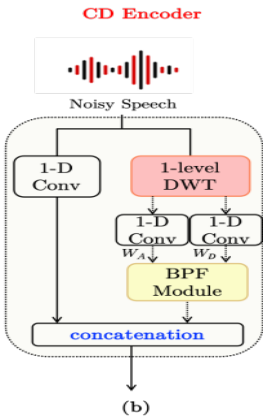
$$W_E = [W_T; W_{DWT}], \qquad (9)$$

因此，矩陣W_E的項數是各域特徵矩陣的兩倍。



(a)

圖三. (a) 融合高頻與低頻之小波域特徵的 BPF 模塊



(b) 更新後的 Conv-TasNet 編碼器部分示意圖，通過先融合再串接來彙整時域和小波域特徵

3 評估實驗與結果討論

3.1 實驗設置

我們首先使用 VoiceBank-DEMAND [9,10] 之資料集任務來評估我們提出的新方法，其中語音訊號來自 VoiceBank 語料庫 (Christophe Veaux et al., 2013)、而摻入的雜訊則來自 DEMAND 資料庫(Joachim Thiemann et al., 2013)。擷取自 Voicebank 語料庫，訓練集包括 28 個語者的 11,572 個語句，測試集包則包括 2 個語者的 824 個語句，訓練集語句由

DEMAND 資料庫中的十種雜訊所混合、其訊雜比(SNR)有四種：0、5、10 和 15 dB。測試集語句則分別以四種訊雜比、摻入五種來自 DEMAND 的雜訊：2.5、7.5、12.5 和 17.5 dB。此外，我們從訓練集中挑選 200 個語句作為驗證集。

為了進一步研究所提出方法的一般化能力(generalization capability)，我們建立了另一個測試集。該測試集使用與原始測試集相同的乾淨語音，但是摻入 QUT-NOISE 資料庫(Dean et al., 2010) 的不同類型的非穩態雜訊、以建構相對於 VoiceBank-DEMAND 任務更加複雜的場景，

且其面對的是訓練模型時未見的雜訊環境(unseen noise)，該測試集之訊雜比分別設為 -5、0、5 和 15 dB，我們稱之為 VoiceBank-QUI 資料集任務。在細項參數設定上，我們使用 db2 小波函數來實現所提方法中的離散小波轉換。對於 Conv-TasNet，我們使用的超參數：個別波段(repat)之捲積塊個數 X = 8、總波段數 R = 3、bottleneck 中的通道數B = 128、個別捲積塊中的通道數H = 256、在 skip-connection 路經之1 × 1 卷積塊的通道數 S = 128、以及捲積塊中的 kernel 大小 P = 3，此設置與文獻(Yi Luo and Nima Mesgarani, 2019) 中的最佳設置的唯一不同，是將參數 H的值減半，來加快訓練與測試之速度。

在評估語音強化的效能上，我們分別使用了 PESQ 分數(Antony W. Rix et al., 2001)作為語音品質(quality)的客觀指標、STOI 分數(Cees H. Taal et al., 2016)作為語音可讀性(intelligibility)的客觀指標、SI-SNR 分數(Yusuf Isik et al., 2016)作為度量語音失真的客觀指標，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。

3.2 實驗結果及討論

3.2.1 使用不同特徵之 Conv-TasNet

我們藉由各種不同的特徵來評估 Conv-TasNet 法。 這些特徵包含原始的時域特徵、CD-TCN 使用的時域與頻域之融合特徵、以及我們提出的三種時域與小波域之彙整特徵。相對應的測試集其 PESQ、STOI 和 SI-SNR 分數如 表一所示。 從這張表中，我們可看出以下幾點：

1. 在 DEMAND 雜訊場景下，與未處理的基礎實驗相比，這裡使用之對應不同特徵的

特徵域	整合方式	VoiceBank 測試集 (Conv-TasNet)					
		DEMAND			QUT-NOISE		
		PESQ	STOI	SI-SNR	PESQ	STOI	SI-SNR
未處理（基礎實驗）		1.970	0.921	8.445	1.247	0.784	3.876
時域	–	2.618	0.943	19.500	1.908	0.860	13.694
時域及頻域	–	2.648	0.942	**19.712**	**1.936**	**0.863**	13.779
時域及小波域	相加	**2.681**	0.942	19.352*	1.922	0.858*	13.645*
	連接	2.669	0.942	19.609	1.932	0.861	13.775
	先融合再串接	2.668	**0.943**	19.496*	**1.936**	0.862	**13.824**

表一. 各種特徵運用於 Conv-TasNet 法、在 DEMAND 與 QUT 雜訊場景下所得的語音強化評估指標值

特徵域	整合方式	VoiceBank 測試集 (DPTNet)					
		DEMAND			QUT-NOISE		
		PESQ	STOI	SI-SNR	PESQ	STOI	SI-SNR
未處理		1.970	0.921	8.445	1.247	0.784	3.876
時域	–	2.549	0.935	19.080	1.804	0.845	12.802
時域及頻域	–	**2.782**	**0.946**	**19.963**	2.019	0.870	14.500
時域及小波域	**先融合再串接**	2.724	0.945	19.960	**2.044**	**0.873**	**14.543**

表二. 各種特徵運用於 DPTNet 法、在 DEMAND 與 QUT 雜訊場景下所得的語音強化評估指標值

Conv-TasNet 法都達到明顯更佳的 PESQ 和 SI-SNR 分數，反映了這些特徵對應之 Conv-TasNet 其優異的語音強化能力。相比之下，它們對於 STOI 指標的改進較少，可能是因為基礎實驗對應的 STOI 分數已經高達 0.921（滿分為 1）。相對而言，在 QUT 雜訊場景中，各種特徵之 Conv-TasNet 法在三個指標(PESQ, STOI 與 SI-SNR)上都能有效提升。

2. 當 Conv-TasNet 法運用於 DEMAND 雜訊場景與 QUI 雜訊場景中，使用單一時域特徵相較於使用時域與頻域之融合特徵、以及三種時域與小波域之彙整特徵，對應的 PESQ 與 SI-SNR 分數大部分都較低，此顯示了頻域與小波域特徵都能補強原始時域特徵、使 Conv-TasNet 法達到更好的效果。
3. 相較於時域與頻域之融合特徵，三種時域與小波域之彙整特徵在 DEMAND 雜訊場景下得到較顯著的 PESQ 提升，而在 QUI 場景下則各有優劣。
4. 若比較所提出的三種時域與小波域之彙整特徵，在比較單純的 DEMAND 場景中，簡易的相加整合法可得到最佳的 PESQ 分數，而先融合再串接的整合法則在較複雜的 QUI 場景中得到較佳的 PESQ 與 STOI 分數，背後可能的原因是，先融合再串接之整合法的特徵數目是相加整合法的特徵數目的 2 倍，且前者有更多的模型參數（如 BPF 模型之參數）需學習，在單純的 DEMAND 場景中較容易使所學習之 Conv-TasNet 模型產生過擬合(over-fitting)的不良現象，但在未知且非穩態雜訊的 QUI 場景中，則是先融合再串接之整合法表現較佳。

3.2.2 使用不同特徵之 DPTNet

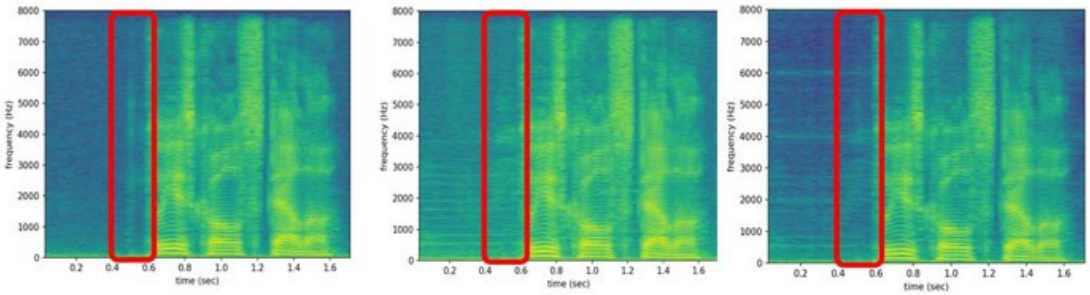
為了驗證所提之小波域特徵對於語音強化模型的效能，在這裡，我們額外採用一個與 Conv-TasNet 法架構相似、但其中的遮罩估測模組採用更細緻安排的架構，即 DPTNet 法(Jingjing Chen, 2020)，其主要參照當今效能卓越之 Transformer 架構(Ashish Vaswani et al., 20117)來構建其遮罩估測模組。類似 3.1 章節的安排，我們藉由相同的資料集安排，使用不同種類的編碼器特徵來訓練與測試 DPTNet 效能，其得到的實驗結果如表二所示，特別註明的是，為了簡單起見，這裡我們只採用了「先融合再串接」的方法求時域與小波域之彙整特徵。

1. 相較於表一與表二的結果，正如預期的那樣，除了時域特徵外，所有種類的編碼特徵在 DPTNet 中都比在 Conv-TasNet 中提供更好的 PESQ 分數。我們認為其可能的原因在於，在時間特徵的情況下，DPTNet 未最佳化之超參數設定造成不如 Conv-TasNet 的結果。
2. 相較於單一使用時域特徵，當與頻域特徵與小波域特徵分別與時間特徵相結合時，DPTNet 在三種語音強化指標上都有明顯的進步，由此可驗證頻域與小波域特徵對於時域特徵的加成性。
3. 若和時域與頻域之融合特徵相比較，時域與小波域之彙整特徵在 DEMAND 雜訊場景下其 PESQ 分數較低、STOI 與 SI-SNR 的分數則較高。然而在 QU 雜訊場景下，時域與小波域之彙整特徵則在三個指標(PESQ, STOI 與 SI-SNR)上都優於時域與頻域之融合特徵，此結果也大致與 Conv-TasNet 法的觀察類似，即採用「先融合再串接」的方法求時域與小波域之彙整特徵，在 QUT 此未知且非穩態雜訊的雜訊場景中表現幾乎是最佳的。

綜合上述的實驗結果，我們可看到，所提出的小波域特徵，在 Conv-TasNet 與 DPTNet 這兩種語音強化模型上，都能有效與原始時域特徵相加成、以達到更佳的語音強化效果，雖然在與時域特徵結合時，小波域特徵與頻域特徵對應上的效果各有優劣，然而本研究是凸顯小波域特徵在與時域特徵的整合上，是頻域特徵的另一個有效選擇，亦即小波域特徵和頻域特徵的加入都能有效改善 Conv-TasNet 與 DPTNet 這些著名的語音強化模型的效能。

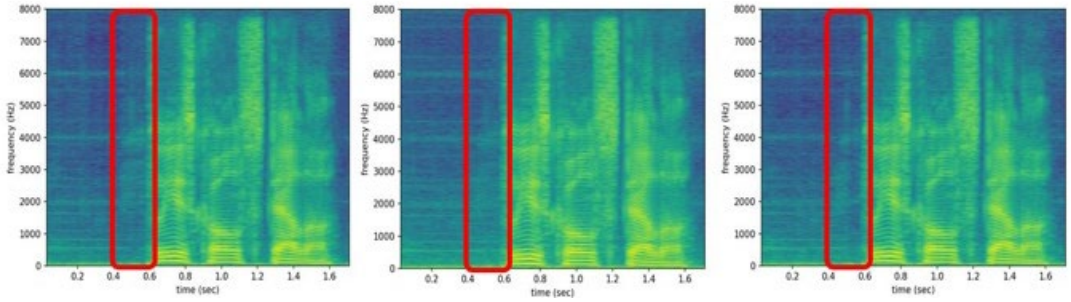
3.3 時頻圖之比較

除了語音強化的各項指標外，這裡我們額外使用語句的強度時頻圖 (magnitude spectrogram)來驗證所提之時域與小波域彙整特徵藉由 Conv-TasNet 所呈現的消噪(denoising)效能。在圖四中，我們分別呈現了一段語句在(a)乾淨環境、(b)雜訊環境、(c)雜訊環境但經時域特徵對應的 Conv-TasNet 強化



(a) clean (b) noisy (c) enhanced by time-domain Conv-TasNet

圖四. 各種場景下的語音強度時頻圖：(a)乾淨環境、(b)雜訊干擾 (c)雜訊干擾且由時域特徵對應之 Conv-TasNet 強化



(a) addition (b) concatenation (c) fusion and concatenation

圖五. 不同時域與小波域之彙整特徵對應的 Conv-TasNet 強化所得之強度時頻圖：(a) 相加法 (b) 連接法 (c) 先融合再連接法

後所得到的強度時頻圖，而圖五則分別對應了原圖四(b)之雜訊干擾語句經過三種時域與小波域之彙整特徵對應的 Conv-TasNet 強化所得之時頻圖強度。比較這些強度時頻圖，我們可以觀察到：

1. 雜訊的摻入對於乾淨語音造成顯著的干擾，特別是在前段的無聲片段中，時頻圖呈現的失真相當明顯。
2. 時域特徵對應的 Conv-TasNet 法能有效減少雜訊造成的時頻圖失真，如圖中的紅框所包含的區域，雜訊成分大幅減少。
3. 與雜訊語音時頻圖相較，三種時域與小波域之彙整特徵對應的 Conv-TasNet 也能帶來顯著的雜訊抑制，從紅框所包含的區域，我們看到第三種彙整法（先融合再串接）似乎比其他兩種彙整法達到更佳的雜訊抑制效果。

4 結論

在這項研究中，我們主要關注於處理語音信號中的雜訊干擾，進而討論一些著名的基於深度學習的語音強化法，並提出使用離散小波轉換為語音強化模型建立特徵。藉由廣泛使用於語音強化實驗的 VoiceBank-DEMAND 和 VoiceBank-QUT 兩種語料庫，我們評估了兩種語音強化模型架構 Conv-TasNet 和 DPTNet，在其上運用我們所提之小波域特徵整合原始時域特徵。初步實驗表明，小波域特徵可以有效與時域特徵相加成，使 Conv-TasNet 與 DPTNet 的語音強化效能更佳，特別是在提升語音的品質指標上。在未來的規劃中，我們將嘗試把小波域特徵應用於其他進階的語音強化模型、如 FullSubNet 與 MANNER 中，觀測其表現，同時也將使用更大型的語音資料評估任務（如 DNS challenge）來評估本研究所提之新方法的效能。

References

Ashish Vaswani et al., "Attention is all you need," *in Proc. NIPS*, 2017

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, Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans on Audio, Speech, and Language Processing*, 2011.

Christophe Veaux, Junichi Yamagishi and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *in Proc. O-COCOSDA/CASLRE*, 2013

Dean, David and Sridharan, Sridha and Vogt, Robert and Mason, Michael. "The QUT-NOISETIMIT corpus for the evaluation of voice activity detection algorithms," *in Proc. Interspeech*, 2010.

Fu-An Chao, Jeih-weih Hung and Berlin Chen, "Cross-Domain Single-Channel Speech Enhancement Model with BI-Projection Fusion Module for Noise-Robust ASR," *in Proc. ICME*, 2021.

Fu-En Wang et al., "BiFuse: Monocular 360° depth estimation via bi-projection fusion," *in Proc. CVPR*, 2020

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

Jingjing Chen, Qirong Mao, Dong Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," *in Proc. Interspeech*, 2020

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

Kahneman, D., Sibony, O., and Sunstein, C. R. "Noise: a flaw in human judgment," *First edition, New York, Little, Brown Spark*, 2021

Philipos C. Loizou, "Speech Enhancement: Theory and Practice," *CRC Press*, 2013

Stephane Mallat, "A Wavelet Tour of Signal Processing," *2nd ed., San Diego, CA: Academic*, 1999

Yi Luo and Nima Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans on Audio, Speech, and Language Processing*, 2019.

Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," *in Proc. Interspeech*, 2016