基於多視角注意力機制語音增強模型於強健性自動語音辨識

Multi-view Attention-based Speech Enhancement Model for Noise-robust Automatic Speech Recognition

趙福安 Fu-An Chao,
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
fuann@ntnu.edu.tw

洪志偉 Jeih-weih Hung
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
jwhung@ncnu.edu.tw

陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

仰賴深度學習(Deep Learning)的發展，近年來許多研究發現相位(Phase)資訊在語音增強(Speech Enhancement, SE)中至關重要。亦有學者發現，透過時域單通道語音增強技術，可以有效地去除雜訊，進而顯著提升語音辨識的精確度。啟發於此，本研究從時域及頻域面分別探討兩種考慮相位資訊的語音增強技術，並提出多視角注意力機制語音增強模型、融合時域及頻域兩者特徵運用於語音辨識中。我們藉由 Aishell-1 中文語料庫評估這些語音增強技術，透過使用各種雜訊源，模擬不同的雜訊狀態作為訓練及測試，進而驗證所提出的新方法皆優於基於其他時域及頻域的方法。具體而言，當測試於訊噪比為-5dB、5dB、15dB 的三種環境下，使用新提出之方法中重新訓練(Retraining)之聲學模型(Acoustic Model, AM)，與基於時域的方法相比較，在已知雜訊的測試集，分別使相對字錯誤率下降 3.4%、2.5%及 1.6%；而在未知雜訊的測試集，則使相對字錯誤率分別

下降了 3.8%、4.8%及 2.2%。

Abstract

Recently, many studies have found that phase information is crucial in Speech Enhancement (SE), and time-domain single-channel speech enhancement techniques have been proved effective on noise suppression and robust Automatic Speech Recognition (ASR). Inspired by this, this research investigates two recently proposed SE methods that consider phase information in time domain and frequency domain of speech signals, respectively. Going one step further, we propose a novel multi-view attention-based speech enhancement model, which can harness the synergistic power of the aforementioned time-domain and frequency-domain SE methods and can be applied equally well to robust ASR. To evaluate the effectiveness of our proposed method, we use various noise datasets to create some synthetic test data and conduct extensive experiments on the Aishell-1 Mandarin speech corpus. The evaluation results show that our proposed method is superior to some current state-of-the-art time-domain and frequency-domain SE methods. Specifically, compared with the time-domain method, our method achieves 3.4%, 2.5% and 1.6% in relative character error rate (CER) reduction at three signal-to-noise ratios (SNRs), -5 dB, 5 dB and 15 dB, respectively, for the test set of pre-known noise scenarios, while the corresponding CER reductions for the test set of unknown noise scenarios are 3.8%, 4.8% and 2.2%, respectively.

關鍵詞：語音強化、自動語音辨識、深度學習、單通道語音增強、重新訓練、聲學模型

Keywords: Speech Enhancement, Automatic Speech Recognition, Deep Learning, Single-Channel Speech Enhancement, Re-training, Acoustic Models

一、緒論

近年來，隨著深度學習的蓬勃發展，現今使用之基於深度學習架構的自動化語音辨識(ASR)系統在無雜訊干擾的情況下，已可達到近乎人類感知水平的辨識水準。但是，在真實環境中往往存在背景雜訊等聲學干擾，若在此環境下，使用事先訓練的 ASR 系統，其辨識性能可能會嚴重下降。為了降低這類干擾效應，長期以來已經發展了許多相關研究及技術，而在這些研究相關技術中，語音增強(Speech Enhancement, SE) 是一個主要的類別，其作為語音在訓練聲學模型(Acoustic Models, AM)前消除雜訊干擾的預處理(Preprocessor)，是一項備受重視的研究方向。

然而，在大多數缺少多麥克風陣列的真實情況，單通道 SE 技術(Single-channel SE)通常效果較差，只能提高語音品質和理解度指標，例如語音品質的感知評估(Perceptual Evaluation of Speech Quality, PESQ)和短時客觀理解度(Short-time Objective Intelligibility,

STOI)，而在這些前端 SE 的指標有所提升，並不能有效地反映在後端 AM 的辨識結果。其原因大致分為兩種：一、在使用前端 SE 模型時產生額外的失真(Distortion)以及雜訊殘留(Artifacts)，進而影響了後端 AM 的辨識效果；二、在訓練 SE 模型時，目標函數(Objective)常設計為優化或近似原乾淨訊號的品質，但 AM 的訓練目標則為最小化分類錯誤、即降低辨識之錯誤率，兩模型訓練目標不一致，造成前後端不匹配的問題。為了解決這樣的現象，儘管許多學者提出使用聯合訓練(Joint Training)的方法，如[8]，針對如何訓練一個強健(Robust)的單通道 SE 模型以減少輸出的失真，還是一項值得研究的方向[9]。

近年來，從時域分析的角度出發，來處理語音訊號的做法，引起了越來越多的關注[6][7]，其中，全摺積時域音訊分離網絡(Fully Convolutional Time-domain Audio Separation Network, Conv-TasNet)[7]在語音分離(Speech Separation, SS)任務上取得了顯著的成果，且超越了頻域上的相關方法。這個想法也被用於單通道 SE 的相關研究，並同時驗證在 ASR 系統上[10]，其對應之使用多情境訓練(Multi-condition Training, MCT) 的 AM，使辨識率獲得顯著進步。然而，此架構若以時域的訊號作為輸入，其特徵抽取(Feature Extraction)的步驟須仰賴摺積神經網路(Convolution Neural Network, CNN)的處理。因此，當沒有足夠的資料數據來找到適當的特徵分佈時，此方法不利於進行小規模資料集的訓練，且增強效果較頻域方法差[11]。

有鑑於此，在本論文中，我們研究了兩種語音特徵，一種是藉由一維摺積神經網路(Convolution Neural Network, CNN)的濾波器組得到的時域特徵，另一種為傳統 STFT 得到的頻域特徵，後者為人工特徵(Hand-crafted Feature)，因此這兩種特徵分別為可訓練的以及固定的。接著我們提出了多視角注意力機制編碼器(Multi-view Attention-based Encoder)，動態地選擇逐音框(Frame-wise)之特徵，合併為單一強健的語音表示作為最終特徵，以輸入至 SE 模型。

為了評估我們提出的方法其效能，我們使用開源中文語料庫 AISHELL-1[12]，並收集多種雜訊資料庫，合成各種帶噪的語音作為訓練以及測試語料。根據評估實驗結果得知，相較於時域上的對應方法，使用新方法之重新訓練之聲學模型(AM)，在已知雜訊的測試集，於三種訊噪比： -5dB、5dB、15dB 之情境中，分別得到相對字錯誤率 3.4%、2.5%及 1.6%的下降；而在未知雜訊的測試集，則使相對字錯誤率分別下降了 3.8%、4.8%及 2.2%。

二、文獻回顧

考慮一段在單通道麥克風中，受加成性雜訊干擾的帶噪離散語音訊號$y(t)$，其公式如下：

$$y(t) = x(t) + n(t) \tag{1}$$

其中$x(t)$是目標之乾淨語音訊號，$n(t)$是加成性雜訊(Additive Noise)，t是時間索引。我們旨在消除帶噪語音訊號$y(t)$中的加成性雜訊$n(t)$，以恢復乾淨語音訊號$x(t)$。

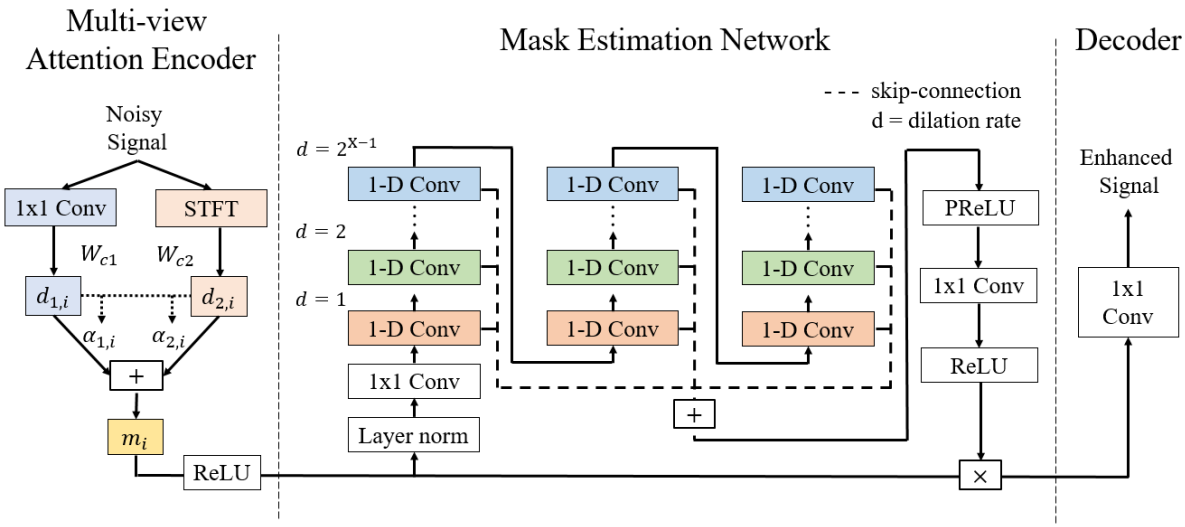
在大多數的研究中，選擇透過短時傅立葉轉換(short-time Fourier transform, STFT)，以在頻域上利用時頻分析(Time-Frequency Analysis)分析帶噪語音訊號，其可視為一特徵抽取的步驟，其公式如下：

$$X(t,f) = \int_{-\infty}^{\infty} w(t-\tau)x(\tau)e^{-j2\pi f\tau}d\tau \tag{2}$$

其中 $w(\cdot)$為窗函數，用於將訊號截短，$x(\tau)$為待轉換的訊號，窗函數隨著時間在時間軸上位移，因此訊號將只留下窗函數截取部分做傅立葉轉換。式(2)是將一維的時域訊號$x(t)$透過短時傅立葉轉換求得$X(t,f) \in \mathbb{C}^{t,f}$，稱為複數時頻圖(complex-valued spectrogram)，當以極座標表示，可得$X(t,f) = |X(t,f)|e^{j\theta_X(t,f)}$，即其可拆解成訊號隨著時間與頻率變化的幅度(Magnitude)$|X(t,f)|$及相位(Phase) $\theta_X(t,f)$。一般來說，相位在估測上比較困難，因此頻域上的語音增強技術通常只針對頻譜幅度做調整、保留原始相位不加以更動。

傳統的 SE 技術主要針對輸入訊號作統計分析，試圖還原乾淨語音訊號，如頻譜消去法(Spectral Subtraction)、維納濾波器(Wiener Filter)等[1]，相較於這些傳統方法所使用的統計分析技術，著名的深度類神經網路(Deep Neural Network, DNN)有更好的非線性轉換能力，已被廣泛應用於 SE 並取得了顯著成果。[2]首先引入了深層去噪自動編碼器(Deep Denoising Auto Encoder, DDAE)，[3]提出了時頻遮罩法(Time-frequency Masking, T-F Masking)。隨後便沿著這個脈絡發展了許多後續研究，這些方法都透過各式 DNN 模型得到優異的去噪效果，大致可以依據其目標分成映射(Mapping)方法或時頻遮罩(T-F Masking)方法[4][5][6]。

近期許多研究發現，相位資訊對語音強化的效能至關重要[4][5][13]。為考慮相位資



圖一、多視角注意力機制語音增強模型

訊，我們需在短時訊號的頻域中同時考慮幅度(Magnitude)和相位(Phase)，或是選擇直接在時域中分析訊號，兩種方法均被證明是有效的，並且優於先前僅考慮頻譜幅度的語音強化技術。在頻域語音強化法中，[13]提出了使用兩個分支結構分別預測幅度遮罩和相位的方法，同時重建幅度和相位，[5]使用遮罩方法，預測複數比率遮罩(Complex Ratio Mask, CRM)。[10]提出了 Denoising-TasNet，其採用[7]中摺積編碼器、解碼器的架構，在時域上對語音訊號做強化，與頻域上的方法相比，其透過控制摺積核(Kernel)以及步長(Strides)大小做摺積運算(Convolution)，類似於短時傅立葉運算(Short-time Fourier Transform, STFT)，隱性地考慮了訊號的相位(Phase)資訊、將其編碼成潛在表示式(Latent Representation)。其去噪效果不僅在訊號失真比(Signal-to-distortion Ratio, SDR)指標有所提升，更驗證在自動語音辨識(ASR)上有效地降低了詞錯誤率(Word Error Rate, WER)。

三、所提出的新方法：多視角注意力機制語音增強模型

在本論文中，我們提出多視角注意力機制語音增強模型，其架構可見圖一，可以拆解成多視角注意力機制編碼器(Multi-view Attention-based Encoder)、遮罩估計網路(Mask Estimation Network)以及解碼器(Decoder)如[7]，藉此對輸入訊號進行建模。首先，我們將輸入之帶噪語音在時域上切割為許多長度為 L 的相互重疊片段，各片段以列向量$x_m \in \mathbb{R}^{1\times L}$表示，其中$m$表示每個片段的索引，$m \in \{1, \dots, M\}$，$M$為總片段數。我們將所有片

段縱向串連在一起，表示為一矩陣$X \in \mathbb{R}^{M\times L}$，其每一列即為$x_m$。

（一） 多視角注意力機制編碼器

多視角注意力機制編碼器考慮了時域以及頻域兩種不同的語音特徵，如圖一，它們分別以C_1與C_2表示，兩種特徵抽取的步驟可視為一線性轉換，首先，時域特徵是藉由矩陣相乘將X轉換至N列的特徵矩陣$C_1 \in \mathbb{R}^{N\times L}$如下：

$$C_1 = \mathcal{F}(UX) \tag{3}$$

其中$U \in \mathbb{R}^{N\times M}$為轉換之矩陣，包含 N 個基函數(Basis Function)，它們對應到一系列可訓練的一維摺積層(1-D Convolution Layer)係數，而$\square(\cdot)$為可選擇的激活函數(Activation Function)，這裡通常選擇線性整流單元(Rectified Linear Unit, ReLU)[7]為激活函數$\square(\cdot)$，以確保特徵為非負值。此外，頻率特徵矩陣C_2是透過傳統的 STFT 對時域訊號矩陣X轉換而得，相較於可訓練的時域特徵矩陣C_1，頻譜特徵矩陣C_2是固定式的，因其求取所用的傅立葉轉換矩陣為一常數矩陣。

根據近期研究，注意力(Attention)模型廣泛被應用於序列至序列(Sequence-to-sequence)的任務中，其中在語音處理領域也獲得了優異的成果。有鑑於此，我們提出多視角注意力機制編碼器，其利用注意力機制將語音特徵，透過注意力加權後，得到單一特徵表示作為輸入之特徵。希望透過注意力機制，使 SE 模型可以更加關注輸入特徵之變異(Variants)並提高 SE 的性能。為了計算注意力權重，我們首先將個別音框特徵透過投影層(Projection Layer)投影至相同維度的向量，如下式：

$$d_{k,i} = W_{ck}c_{k,i} + b_{ck} \tag{4}$$

其中，$c_{k,i}$為語音特徵矩陣C_k之個別音框的特徵向量、i 表示音框索引(Frame index)、W_{ck}，b_{ck}分別是投影層的權重矩陣(Weight Matrix)和偏差(Bias)。藉由投影得到之向量$d_{k,i}$，我們可以透過時序注意力機制(Temporal Attention Mechanism)計算注意力權重$\alpha_{k,i}$：

$$\alpha_{k,i} = \frac{\exp(v_{k,i})}{\sum_{k=0}^{K-1}\exp(v_{k,i})} \tag{5}$$

其中$v_{k,i}$是$d_{k,i}$和其他$d_{k',i}(k' \neq k)$間的相似度(Similarity)，K是特徵種類的個數，在考慮時域及頻域兩種特徵情況下，$K=2$。在我們的實驗中，我們考慮了三種不同相似度分數計算公式，如表一所示：

表一、不同注意力機制相似度分數之計算 (對兩種特徵而言，$k = 0, 1$)

注意力機制	公式
加成性(Additive)	$v_{k,i} = \omega_A^{\mathrm{T}} \tanh\left(W_A d_{k,i} + B_A d_{1-k,i} + b_A\right)$
串接性(Concatenate)	$v_{k,i} = \omega_C^{\mathrm{T}} \tanh\left(W_C [d_{k,i}; d_{1-k,i}] + b_C\right)$
縮放點積(Scaled dot-product)	$v_{k,i} = \frac{d_{k,i}^T d_{1-k,i}}{\sqrt{n_d}}$

其中W_A、W_C、B_A為權重矩陣，b_A、b_C為偏差，ω_A、ω_C為向量，n_d為投影向量之維度，$\frac{1}{\sqrt{n_d}}$為控制兩向量點積(Dot-product)後值的縮放係數。

最後我們利用計算完的注意力權重$\alpha_{k,i}$將特徵之投影向量$d_{k,i}$加權後相加，得到融合後的特徵表示式z_i：

$$z_i = \sum_{k=0}^{1} \alpha_{k,i} d_{k,i} \tag{6}$$

（二） 遮罩估計網路

在基於遮罩的 SE 模型中，通常採用 DNN 模型來估計輸入帶噪語音特徵的遮罩，藉此來分離雜訊與語音。為了有效地捕獲時間資訊並考慮音框之間的長期依賴性(Long-term Dependency)以預估遮罩，大部分的研究透過堆疊雙向長短期記憶層(Bidirectional Long-short Term Memory, BLSTM)或擴張摺積層(Dilated Convolution Layer)來實現。其輸出可以直接是目標語音的單個遮罩M_x，也可以選擇輸出語音和雜訊兩個各別遮罩M_x, M_n：

$$[M_x, M_n] = \mathcal{M}_\theta(Z), \tag{7}$$

其中$\mathcal{M}_\theta(\cdot)$為遮罩估計網路，$Z$為輸入語音特徵矩陣，可為式(6)所得之個別音框特徵的排列：$Z = [z_1, z_2, \ldots, z_L]$。在得到估計之遮罩後，我們可以將語音遮罩與輸入特徵Z逐項相乘得到強化後的語音特徵矩陣$D_x \in \mathbb{R}^{N \times L}$：

$$D_x = M_x \odot Z \tag{8}$$

其中，⊙表示逐元素相乘運算(Element-wise Product)。

（三） 解碼器

得到強化後的語音特徵D_x後，我們透過解碼器(Decoder)將特徵表示轉換並重建時域之波形，此步驟可以視為矩陣相乘運算，如下式：

$$\hat{S}_x = VD_x \tag{9}$$

其中$\hat{S}_x \in \mathbb{R}^{M\times L}$為重建的語音片段組成之矩陣，而解碼器矩陣$V \in \mathbb{R}^{M\times N}$是由長度為$M$的$N$個基函數排列而成。在頻域中，矩陣$V$可以是短時逆傅立葉轉換(Inverse Short-time Fourier Transform, iSTFT)，在時域中，矩陣V則對應至一維轉置摺積運算(1-D Transpose Convolution)。最後我們使用重疊相加(Overlap-add)的方法從片段矩陣$\hat{S}_x$重構語音訊號。

四、實驗設置

我們使用 AISHELL-1[12]語料集來執行評估實驗，其為北京希爾貝殼科技有限公司提供的開源中文 ASR 語料庫，包含 400 位語者和 170 個小時的中文語音。為了評估我們提出的方法，我們從原始的訓練集中生成了一些模擬雜訊的訓練資料，並使用各種雜訊資料集在不同的訊雜比(SNR)及不同雜訊種類之狀態下設計了四種類型的測試集。

為了進行訓練，我們採用的雜訊來自 MUSAN[14]、DEMAND[15]、QUT-NOISE[16]和環境背景雜訊資料集[17][18]，總共 2553 種雜訊，乾淨語音和雜訊音檔皆重新採樣至 16kHz。我們將這些雜訊以 5dB 之 SNR 值混入原始訓練資料中的每個語句，合成與原資料相同(即 SNR 5 dB 的 120098 個帶噪語句)的多條件訓練(Multi-condition Training, MCT)資料(以下稱為”MCT 資料”)。

在測試資料的準備上，我們將原始測試集語音及上述之雜訊，根據三種 SNR 值設定: -5 dB、5 dB 與 15 dB 加以混合，分別建立了三個額外的測試集。另外，為模擬測試於未知雜訊的情形，我們使用了不同的雜訊資料集：Nonspeech 雜訊資料集[19]，以 SNR 5dB 來合成第四種測試集。

（一）語音增強模型設置

針對語音增強系統，我們在原始 AISHELL-1 訓練集和發展集中，每個語者隨機挑出 20 則語句(分別為 6800 及 800 則語句)來訓練我們的語音增強系統，所有資料均使用前述

MCT 雜訊以 5dB 之 SNR 加以混合。訓練的所有語音增強模型只輸出乾淨語音，訓練皆不採用語音分離問題[7]中的置換不變訓練(Permutation invariant Training, PIT)，且損失函數皆為負比例無關訊噪比(Scale-invariant Signal-to-noise ratio, SISNR)，相關公式如下：

$$\mathbf{s}_{target} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \tag{10}$$

$$\boldsymbol{e}_{noise} = \hat{\boldsymbol{s}} - \boldsymbol{s}_{target} \tag{11}$$

$$SISNR = 10\, log_{10} \frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \tag{12}$$

$$\mathcal{L}_{SISNR} = -SISNR(\mathbf{s}, \hat{\mathbf{s}}) \tag{13}$$

其中$\hat{\mathbf{s}} \in \mathbb{R}^{1\times T}$、$\mathbf{s} \in \mathbb{R}^{1\times T}$，分別為預測語音訊號向量及乾淨語音訊號向量，$T$為訊號之長度，$\mathbf{s}_{target}$為$\hat{\mathbf{s}}$於$\mathbf{s}$之投影向量。

我們考慮了三種不同特徵的語音增強模型，所有模型遮罩估計網路皆使用時間摺積網路(Temporal Convolution Network, TCN)[7]，包含殘差連結(Residual-connection)以及跨層連結(Skip-connection)，可見圖一。相關之超參數 (hyperparameter)設置為$X = 8$、$R = 3$、$B = 128$、$H = 512$、$S = 128$、$P = 3$，為[7]中最佳的模型設置，不同之處在於編碼器與解碼器之結構：

1. **STFT-TCN**

   在頻域 SE 模型設定上，我們使用 STFT 以及 iSTFT 作為編碼器與解碼器，藉由 STFT 我們得到複數頻譜(Complex-valued Spectrogram)作為頻域特徵，特徵的前半部$(1, \dots, N/2)$表示實數(Real Value)部分，後半部$((N/2+1, \dots, N)$表示虛數(Imaginary Value)部分，與[11]相同。我們設置傅立葉轉換點數為 512 點，窗函數為漢寧窗(Hanning Window)，取窗長為 64 (個樣本數)，窗移為 32 (個樣本數)擷取特徵，得到 512 維之頻域特徵。

2. **Denoising-TasNet(M)**

   在時域 SE 模型設定上，我們參考[10]提出 Denoising-TasNet 之架構，使用一維摺積層及轉置摺積層作為編碼器與解碼器。且根據[7]，在提取時域特徵時使用較小的窗長可以有較佳的效果，因此我們設置窗長為 16 (個樣本數)，窗移為 8 (個樣本數)擷取特徵，濾波器數為 512，最後得到 512 維之時域特徵。由於這裡使用的模型較原[16]提出的模型，多了跨層連結且有更好的效果，我們將其稱之為 Denoising-

TasNet(M)。

3. Multi-view-TCN

在我們新提出之多視角注意力機制語音增強模型中，我們使用了多視角注意力機制編碼器及 1-D 轉置摺積層作為解碼器，取 256 維的頻域特徵與 256 維的時域特徵進行融合，為了固定特徵之總音框數，我們將窗長及窗移皆設置為 16 (個樣本數)以及 8 (個樣本數)，映射層之維度設置為 128，最後得到 128 維的融合特徵。

（二）語音辨識模型設置

於後端語音辨識系統，我們使用 Kaldi 建構 Hybrid DNN-HMM 聲學模型，遵循[18]提供的 GMM-HMM 訓練流程。在 DNN 模型的訓練，我們堆疊分解式時延神經網絡(Factorized Time Delay Neural Network, TDNN-F)，採取詞圖無關最大交互資訊(Lattice-free Maximum Mutual Information, LF-MMI)目標函數進行訓練，其可在原官方提供之測試集獲得更好的結果(7.46%及 6.51%之字錯誤率，CER)，我們將此系統作為我們實驗的基線(Baseline)。另外，我們訓練了兩種聲學模型(AM)進行比較，一種訓練資料包含原始訓練資料和 MCT 資料，為多情境訓練之 AM，以 MCT-AM 表示；另一種訓練資料包含原始訓練資料，MCT 資料和使用對應 SE 模型增強 MCT 資料得到的 ENH 資料，以彌除增強後的特徵與模型不匹配的問題，為重新訓練(Retraining)之 AM，以 ENH-AM 表示。

五、實驗結果及分析

我們使用 SISNR 作為 SE 之評估方式，見式(15)，單位為 dB 其值越高越好；於 ASR，我們則採用字錯誤率(Character Error Rate, CER)作為評估以百分比表示，值為越低越好。

（一）不同注意力機制之比較

表二、不同注意力機制之比較

AM Model	SE Model	Attention Mechanism	Test (5dB)	
			CER	SISNR
Baseline	–	–	43.81	5.02
MCT-AM	–	–	18.75	5.02
MCT-AM	Multi-view-TCN	Additive	17.05	14.81
		Concatenate	16.92	14.88
		Scaled dot-product	**16.49**	**14.91**

我們首先針對提出的 Multi-view-TCN 模型，比較不同注意力機制的效果，於此實驗我們測試在已知雜訊為 5dB SNR 之測試集，SE 法只作用於測試集，實驗數據如表二所示。由此表之第一、二列可以發現，相較於基線結果，使用 MCT-AM，在帶噪的環境中可以大幅地降低 CER。此外，根據第二至四列之數據可知，在不重新訓練聲學模型的前提下，所新提出之 Multi-view-TCN 法在三種不同注意力機制下，不僅提升語音增強指標 SISNR，同時也增加了語音辨識率(即 CER 降低)。其中，以伸縮點積(Scaled dot-product)注意力機制效果最為顯著，因此，我們在此後的實驗，於 Multi-view-TCN 模型上皆使用伸縮點積注意力機制。

（二）已知雜訊環境之語音辨識結果

在第二部分的實驗中，我們測試各模型在已知雜訊且不同訊噪比的環境，分別針對語音增強及語音辨識的表現進行探討，實驗結果請見表三。

表三、已知雜訊環境之語音辨識結果

AM Model	**SE Model**	**Test (-5dB)**		**Test (5dB)**		**Test (15dB)**	
		CER	**SISNR**	**CER**	**SISNR**	**CER**	**SISNR**
Baseline	–	81.41	-4.94	43.81	5.02	15.14	15.02
MCT-AM	–	58.19	-4.94	18.75	5.02	**8.66**	15.02
	STFT-TCN	49.78	**5.43**	18.92	13.71	9.42	18.73
	Denoising-TasNet(M)	47.47	4.52	16.99	14.69	9.14	19.52
	Multi-view-TCN	**46.62**	5.01	**16.49**	**14.91**	8.90	**19.83**

根據表三之數據，在語音增強的效果上，所提出的 Muliti-view-TCN 模型相較於其他兩方法幾乎都可得到更佳的 SISNR 值，唯在 -5dB 之 SNR 的測試集較基於頻域的方法(STFT-TCN)差；在語音辨識效能上，Multi-view-TCN 相較其他兩方法皆得到更高的辨識精確率，由此可知，SISNR 指標的提升雖不能完全反映在 CER 的下降，但大致有一致的現象。然而，所有模型在 15dB 之 SNR 環境下，語音辨識表現皆比原始 MCT-AM 差，我們猜測在雜訊干擾較少的環境，語音透過語音增強模型會產生較多額外的失真，導致前後端較顯著之不匹配現象。因此我們採取重新訓練的方法，使用各模型對原 MCT 資料增強後加入訓練資料，重新訓練聲學模型，進行後續的實驗。

（三）重新訓練聲學模型之結果

表四、重新訓練聲學模型之結果

AM Model	SE Model	Test (-5dB)	Test (5dB)	Test (15dB)
		CER		
MCT-AM	–	58.19	18.75	8.66
MCT-AM	STFT-TCN	49.78	18.92	9.42
ENH-AM		40.20	12.69	7.81
MCT-AM	Denoising-TasNet(M)	47.47	16.99	9.14
ENH-AM		40.41	12.02	7.79
MCT-AM	Multi-view-TCN	46.62	16.49	8.90
ENH-AM		**39.03**	**11.71**	**7.66**

根據表四可發現，在重新訓練聲學模型後，在不同訊噪比之測試環境下皆有顯著進步的語音辨識率，且在高訊噪比的環境(SNR 15dB)，所有模型皆比 MCT-AM 佳，因此驗證了重新訓練聲學模型的方法，可以彌除前後端不匹配的現象。而所新提出的 Multi-view-TCN 在各測試集皆表現最佳，相較於 Denoising-TasNet(M)法，在-5dB、5dB 及 15dB 三種 SNR 測試環境下可獲得 3.4%、2.5%及 1.6%相對字錯誤率下降。

（四）未知雜訊環境之語音辨識結果

表五、未知雜訊環境之語音辨識結果

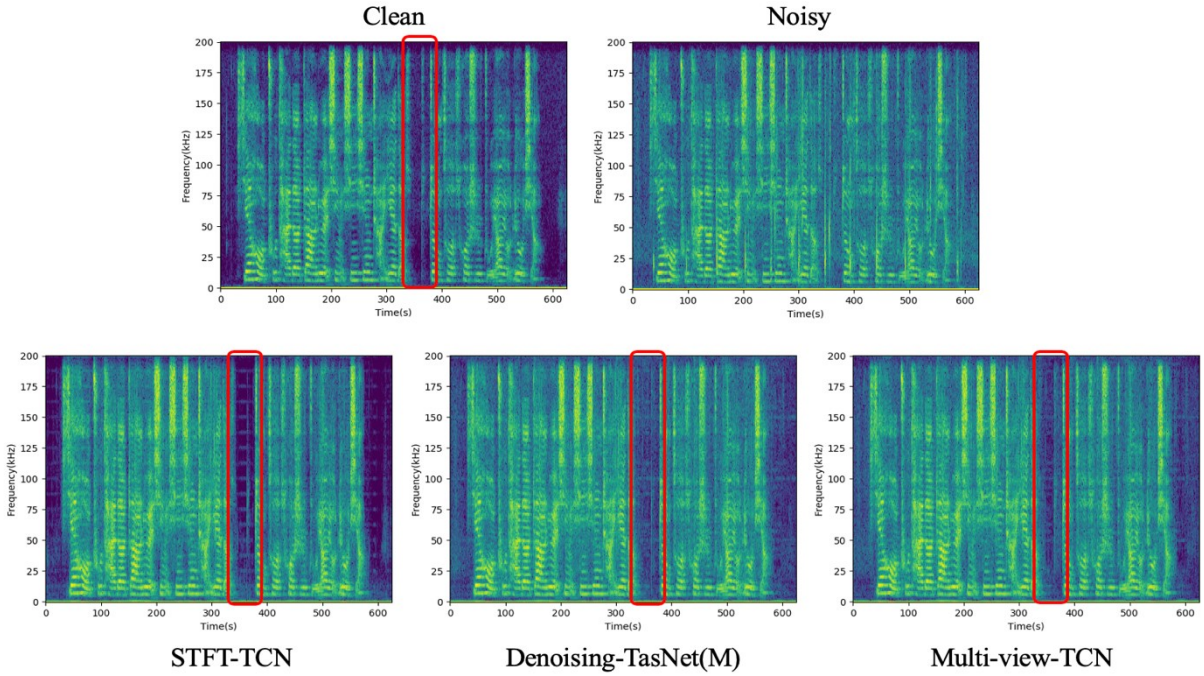
AM Model	SE Model	Test (unseen) (-5dB)		Test (unseen) (5dB)		Test (unseen) (15dB)	
		CER	SISNR	CER	SISNR	CER	SISNR
Baseline	–	88.42	-4.97	50.58	5.01	16.92	15.01
MCT-AM	–	70.90	-4.97	25.37	5.01	10.45	15.01
MCT-AM	STFT-TCN	60.56	5.98	25.86	13.79	11.80	19.09
ENH-AM		47.40	5.98	15.51	13.79	8.62	19.09
MCT-AM	Denoising-TasNet(M)	56.88	5.73	22.82	14.54	11.44	19.96
ENH-AM		47.31	5.73	14.88	14.54	8.63	19.96
MCT-AM	Multi-view-TCN	55.38	6.10	22.08	14.71	10.92	20.15
ENH-AM		**45.49**	**6.10**	**14.16**	**14.71**	**8.44**	**20.15**

表五為未知雜訊環境的實驗結果，從此表我們可以觀察到，在雜訊未知的環境下，各方法於語音辨識的表現與前述實驗有一致的趨勢，且在低訊噪比環境(-5dB SNR)時，新方

法其表現皆優於頻域的 STFT-TCN 法及時域的 Denoising-TasNet(M)法，因此，所新提出的 Multi-view-TCN，展現了在未知雜訊的環境下之泛化(Generalization)的能力，在 SISNR 及 CER 的評估上皆得到最好的結果，相較於 Denoising-TasNet(M)，在 SNR 為-5dB、5dB 與 15dB 之雜訊環境下，可得到 3.8%、4.8%及 2.2%的相對字錯誤率下降率。

（五）語音增強效果比較

最後，我們比較不同系統在語音增強上的效果，其音檔範例可見[1]。另外，我們比較不同系統於同一音檔增強後的時頻圖(Spectrogram)，如圖二所示。可以觀察紅色粗體方框的範圍中， 頻域之 STFT-TCN 法會殘留許多雜訊頻譜，並將某些原音框中乾淨語音的頻譜一併消除；而基於時域的模型 Denoising-TasNet(M)，雖保留了較多乾淨頻譜成分，但同時也包含了許多雜訊頻譜成分；Multi-view-TCN，則似乎是在兩者之中權衡，因此獲得了比較好的增強效果。



圖二、語音增強系統之時頻圖比較 (BAC009S0764W0193, 腳步聲, 5dB SNR)

六、結論

在本研究中，我們提出了多視角注意力機制語音增強模型於強健語音辨識，同時考慮了頻域特徵(複數時頻圖)以及時域的特徵，透過注意力機制將兩者特徵融合為單一特徵，

[1] https://smildemo.csie.ntnu.edu.tw/rocling_demo/index.html

藉由 Aishell-1 開源中文語音語料庫的評估實驗，透過使用各種不同的雜訊源，模擬不同的雜訊情形作為訓練及測試，充分顯示此新方法皆優於基於時域的語音增強方法。而在已知雜訊的測試集中，於-5dB、5dB、15dB 三種 SNR 環境，使用重新訓練之聲學模型，相對於時域上的方法，新方法可分別下降相對字錯誤率 3.4%、2.5%及 1.6%；而在未知雜訊的測試集，則獲得相對字錯誤率 3.8%、4.8%及 2.2%的下降。

從實驗結果而論，於時域處理的語音增強模型，不僅對於語音增強效果優異，同時也較適用於後端語音辨識模型。於未來，我們將專注在時域特徵之研究，並考慮更多語音特徵應用在所提出之基於注意力機制之特徵融合方法，並將問題進一步延伸至處理摺積型雜訊(Convolution Noise)，如混響(Reverberation)之干擾等。

參考文獻

[1] N. Wiener, “Extrapolation, Interpolation, and Smoothing of Stationary Time Series”, *NewYork: WILEY*, 1949.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. INTERSPEECH*, pp. 436–440, 2013.

[3] Y. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381–1390, 2013.

[4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Proc. LVA/ICA*, pp. 91–99, 2015.

[5] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[6] D. Rethage, J. Pons, and X. Serra, “A Wavenet for speech denoising,” in *Proc. ICASSP,* pp. 5069–5073, 2018

[7] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[8] T. Menne, R. Schlüter and H. Ney, “Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR,” *arXiv preprint arXiv:1904.09049, 2019*.

[9] K. Tan and D. Wang, “Improving robustness of deep learning based monaural speech enhancement against processing artifacts,” in *Proc. ICASSP*, 2020.

[10] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, “Improving noise robust automatic speech recognition with single-channel time-domain enhancement network,” in *Proc. ICASSP*, pp. 7009–7013, 2020

[11] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, “Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks,” *arXiv preprint arXiv:2005.11611*, 2020.

[12] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL1: An open-source Mandarin speech corpus and a speech recognition baseline,” in *Proc. O-COCOSDA*, pp. 1–5, 2017

[13] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “Phasen: A phase-and harmonics-aware speech enhancement network,” *arXiv:1911.04697*, 2019.

[14] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,”*arXiv preprint arXiv:1510.08484*, 2015.

[15] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[16] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” in *Proc. INTERSPEECH*, pp. 3110–3113, 2010

[17] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, “Smart phone-based real-time classification of noise signals using subband features and random forest classifier,” in *Proc.. ICASSP*, pp. 2204–2208, 2016

[18] F. Saki and N. Kehtarnavaz, “Automatic switching between noise classification and speech enhancement for hearing aid devices,” in *Proc. EMBC*, pp. 736–739, 2016

[19] G. Hu and D. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067-2079, 2010.