ESC MA-SD Net: Effective Speaker Separation through Convolutional Multi-View Attention and SudoNet
通過卷積多視角注意力和 SudoNet 進行高效的人聲分離

Che-Wei Liao
Dept. of Electrical Engineering
National Chi Nan University
Nantou County, Taiwan
s109323018@mail1.ncnu.edu.tw

Aye Nyein Aung
Dept. of Electrical Engineering
National Chi Nan University
Nantou County, Taiwan
s111356509@ncnu.edu.tw

Jeih-weih Hung
Dept. of Electrical Engineering
National Chi Nan University
Nantou County, Taiwan
jwhung@ncnu.edu.tw

摘要

本研究以人聲分離(speech separation)為主題，研究如何將混合的多個人聲信號成功分離。我們是利用端到端(end-to-end)的高效語音分離模型 SuDoRM-RF 做為基礎，並結合了 MANNER 模型中的殘差卷積轉換器區塊(Residual Conformer Block)以及多視角注意力區塊(Multi-view Attention block)來達到高效的語音分離模型 ESC MA-SD Net。本模型中殘差卷積轉換器區塊在於移除無用資訊的同時還能保留重要語音信息，而透過多視角注意力模塊則用以關注擷取對各個面向語音特徵，如此一來，我們將可以得到相較原本 SuDoRM-RF 模型更加高效的語音分離模型 ESC MA-SD Net。在我們的實驗中，分別從驗證資料(Validation dataset)以及時頻圖(Spectrogram)來展示提出之方法的良好的語音分離成效。

Abstract

This study focuses on speaker separation, investigating how to successfully separate mixed multiple speech signals. We build upon the efficient end-to-end speech separation model SuDoRM-RF and integrate the Residual Conformer Block from the MANNER model along with the Multi-view Attention block to create the efficient speech separation model ESC MA-SD Net. The Residual Conformer Block in this model eliminates irrelevant information while preserving crucial speech details. The Multi-view Attention module is employed to capture diverse aspects of speech features. By doing so, we achieve a more efficient speech separation model, ESC MA-SD Net, compared to the original SuDoRM-RF model. In our experiments, we demonstrate the effectiveness of the proposed method using validation data and spectrograms to showcase the improved speech separation performance.

關鍵字：語音分離、殘差連接法、端到端模型
Keywords: Speech separation, Residual connect method, End to end module

1 緒論 (Introduction)

語音處理的技術隨著科技的進步不斷地在更新，之前傳統的語音處理技術通常都是由數個不同功能的模塊所組合而成，這些模塊都需要個別去訓練，但這些模塊都是分開訓練、使用不同訓練資料、要調整的參數也都不同，這樣對整體模型的最佳化將造成困難。而近年基於深度學習之模型架構、並使用了端到端(end-to-end)(Amodei, D., Anubhai, R., Battenberg, E., et al. 2016) 的整體訓練模式，其對應的優點是可以直接從原始的語音信號生成對應的輸出，直接最佳化整體模型的輸出結果、使其中各個模塊能夠在訓練過程中、同時更新並彼此配合來使模型最終輸出趨於目標輸出(ground-truth output)。本文所提出之 ESC MA-SD Net 語音分離模型的訓練即採用端對端的模式。語音分離演算法可以依照訓練目標(training target)、分成對映式(mapping)以及遮罩式(masking)(Wang et al., 2014) 前者直接求取輸入混合語音與輸出之分離語音的對映函數，而這些對映函數所要轉換的語音特徵通常包括了耳蝸時頻譜圖 (cochleagram) 、梅爾倒頻譜(Mel-Frequency Cepstrum, MFCC)、

時頻圖(spectrogram)等；而後者則求取一個遮罩函數，使此遮罩與原始混合語音相乘後，能近似分離語音，此遮罩函數較著名的選擇包括了理想二值掩蔽（Ideal Binary Mask, IBM）、理想比例遮罩（Ideal Ratio Mask, IRM）(DeLiang Wang et al., 2018)、複數理想比例遮罩(complex ideal ratio mask, cIRM)(Williamson D S., 2015)等等。本研究是針對效果優異的遮罩式語音分離法 Successive Downsampling and Resampling of Multi-Resolution Features (SuDoRM-RF) (Efthymios Tzinis et al., 2020)加以改進，過程中需要訓練出兩個不同的遮罩，用以對混合的語音做相乘，以分離出兩個人聲。

2 SuDoRM-RF

SuDoRM-RF 語音分離模型是採用常見的編碼器(encoder)、分離器(separator)、解碼器(decoder)所組合而成的架構(如圖 1 所示)。這個模型是採用時域分析特徵的語音分離模型，根據文獻(Meta AI, 2023)指出，時域分析(time-domain analysis)的語音分離法相較於短時頻域分析(time-frequency domain analysis)對於效能指標 SI-SDR 的進步一般而言是較顯著的，因為相較於對於固定基底轉換(弦波函數)的時頻分析而言，時域分析可以訓練其轉換的基底，對於語音分離模型的訓練上能有更多的彈性。此外，它還有一項很重要的特點，是它在分離層(separator) 使用了 U- convolutional block 模塊，如圖 2 所示，可以有效的降低運算的複雜度。

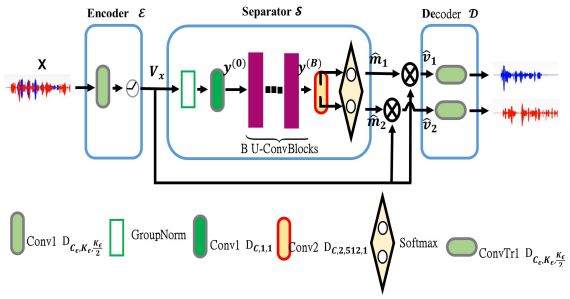

圖 1：SuDoRM-RF 的基本架構

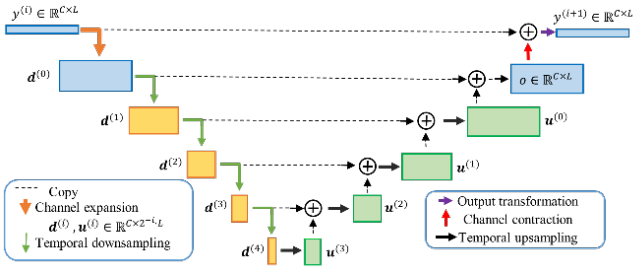

圖 2：U-convolutional block

2.1 U-convolutional block

根據文獻(Efthymios Tzinis et al., 2020)顯示 U-convolutional block 此種語音處理模型所使用的參數相對於 Conv-TasNet (Yi Luo et al., 2019) 、Two-Step TDCN (Efthymios Tzinis et al., 2020) 等模型，使用較少的參數量，這是因為此模型採用了連續下採樣與重採樣的模塊來做訓練，它可以利用這種方式來建立資料與資料間的關聯性，可擷取更多的聲音細節，讓聲音的效果更好。而其中在重採樣的部分，可以通過複製原始數據的方式，來增加採樣點數，透過這種方式，可以在同樣的時間內，收集更多的聲音資訊，讓分離後的聲音更加精確清晰，而且在這過程中不需要增加任何參數。這點可以使我們在訓練過程中使用到較少的模型參數就可以達到很好的效果。

3 提出的新方法 (Proposed Method)

在本研究中，我們以 SuDoRM-RF 為基礎架構、保留 U-convolutional Block 此參數之模型，但參照了著名的語音強化模型 MANNER (Hyun Joon Park et al., 2022) 來改造 SuDoRM-RF，首先，我們採用殘差卷積轉換器模塊 (Residual Conformer Block)來取代原本 SuDoRM-RF 中的 bottleneck 模塊，接著在其做完連續的下採樣以及重採樣後添加了 MANNER 架構中的多視角注意力模塊 (Multi-view Attention block) 進而探究其是否能提升語音分離的效果。

3.1 殘差卷積轉換器模塊 (Residual Conformer Block)

殘差連接(Residual Connection)是深度學習中經常使用到的一種技巧，它的目的是在移除無用資訊的同時還能保留重要資訊，MANNER 法中提出的 Residual Conformer block 除了殘差連接之外還加入了 Conformer 模組，其架構如

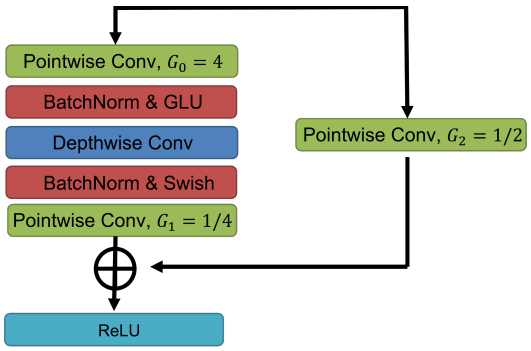

圖 3：殘差連接模塊

圖 3 所示，首先使用常見的 pointwise convolution 來擴展通道大小、以獲得更豐富的語音特徵表示，G_0即是將通道放大 4 倍，本研究所採用之初始值為 256，中間使用到的 GLU 是將原始訊號的一部份做 sigmoid 轉換，再與原始特徵做結合，這麼做的目的是可以保留一部份的原始特徵並抑制原始數據的部分訊息，這可以使模型更好的學習到關鍵特徵以提高模型的精準度。因此通過 GLU 的通道會變成原始數據的一半，最後我們再利用G_1將通道樹變成原始數據的 1/2，變成 1/2 的目的是為了取代原始模型 SuDoRM-RF 內的 bottleneck 模塊，這個模塊是將通道數變為原始通道的一半。

3.2 多視角注意力模塊 (Multi-view Attention block)

注意力機制是近幾年來在深度學習領域中被廣泛運用的一種方法，近幾年許多 AI 機器人中所使用的 Transformer 架構就是依照 attention 這個技術為基礎的。

參照 MANNER 法，其多視角注意力機制是分別從 channel、global 以及 local 三種角度來對輸入特徵施以注意力，如圖 4 所示，其中 channel attention 會對每個通道做平均以及最大池化(Average & Max pooling)過濾通道再加權來增加語音的特徵，global attention 是這是基於 Transformer 中的 self-attention，考慮到分塊輸入中每個分塊的表示，通過自注意力機制提取全局序列信息。

最後 local attention 通過對每個分塊進行卷積操作，捕捉該分塊中的局部序列特徵。可以有效地降低模型的計算成本和內存占用，同時保持較高的準確性和性能。簡單來說 global 跟

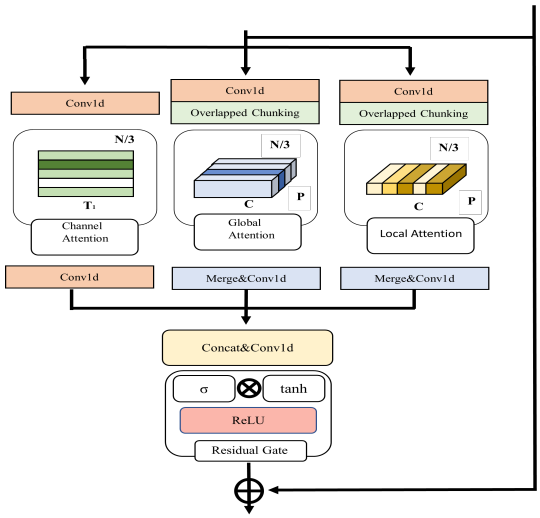

圖 4：多視角注意力模塊

local 則是分別對全局的通道以及特定通道施以注意力機制，以達到最佳的訓練效果。

我們提出的改良式架構，命名為 **E**ffective **S**peaker Separation through **C**onvolutional **M**ulti-View **A**ttention and **SudoN**et (簡稱 ESC MA-SD Net)，其架構圖為圖 5 所示，與圖 1 之原始 SuDoRM-RF 相較，此架構在 U-convolutional blocks 之前與之後分別添加了 Residual Conformer block 與 Multi-view Attention Block，目的在於加強模型對於提取特徵的能力、進而提升語音分離之效果。

4 實驗設置(Experimental Setup)

參照文獻(Efthymios Tzinis et al., 2020)中的 improved sudorm-rf 程式碼作為基礎，並結合了文獻(Hyun Joon Park et al., 2022)程式碼中的 ResCon block 以及 MA block 來完成我們 ESC MA-SD Net 之實驗程式。我們使用的是 Wham! 語音資料庫，它是提取 wsj0-2mix 數據集中的每個雙人混合聲音與獨特的噪聲背景場景配對。訓練資料有 20000 筆測試資料則是有 3000 筆，而進出 U-Convolution block 的通道數我們分別設置了 256 以及 512，U-Convolution block 總共使用了 4 個，深度皆為 5。我們使用語音

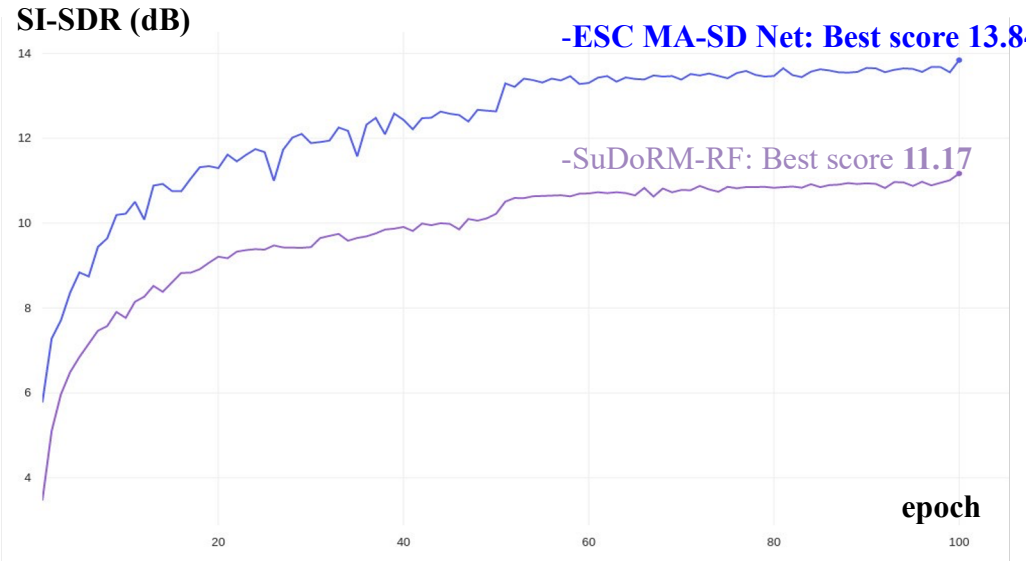

圖 6：ESC MA-SD Net 與 SuDoRM-RF validation 過程比較圖

分離常用的 SI-SDR 指標(Morten Kolbæk et al., 2020)作為實驗的評估依據，該公式為：

$$SISDR = 10log_{10}\frac{||X_T||^2}{||X_E||^2}$$

簡單來說這個評估值是把所得之分離訊號分解為兩個正交訊號：分子項的X_T是與目標訊號相平行的分量（視為目標訊號成分）、分母項X_E則是與目標訊號相垂直的分量（視為干擾成分），若干擾的成分越少，語音的成分越高，這樣 SI-SDR 的分數就會越高，最後本實驗用相同的規格與原本的 SuDoRM-RF 語音分離模型做比較，以觀察改進後的 SuDoRM-RF 模型與原模型的差別。

5 實驗結果與討論 (Experimental Results and Discussions)

我們藉由驗證集之 SI-SDR 值與時頻圖來呈現新方法 ESC MA-SD Net 與基礎之 SuDoRM-RF 的差異，SI-SDR 值如圖 6 所示。該圖橫軸為驗證過程中之 epoch 數、縱軸為各個 epoch 之

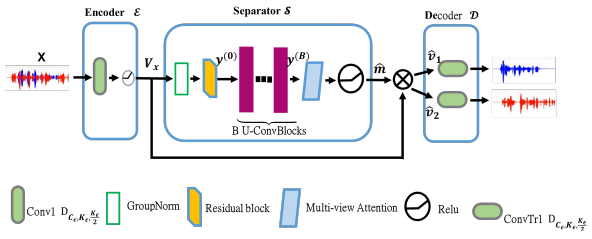

圖 5：ESC MA-SD Net 之流程圖

模型對於驗證資料(validation set)所計算之 SI-SDR 值，分數越高代表分離出來的語音越清晰。因為 SuDoRM-RF 論文中所採用的是數據是以驗證集做比較，為了更公平且貼近 SuDoRM-RF 中的數據，本實驗選擇採以驗證集做比較的方式進行。從圖中可看出，整體驗證過程從一開始到最後第 100 次之 epoch，ESC MA-SD Net 在 SI-SDR 的指標值皆是明顯優於 SuDoRM-RF。因此我們可以推斷得知，原本 SuDoRM-RF 的 1-D convolution 結構之 Bottleneck layer，當改為 MANNER 中的 Residual Conformer Block，雖然同樣可達到 channel 數減半的效果，但 Residual Conformer Block 應可以更充分擷取原始語音訊息，不至於造成訊息損失，之後的 Multi-view Attention Block 也同樣帶來顯著貢獻，因此整體而言使該架構達到更加的語音分離效果。

除了 SI-SDR 指標外，我們從一個混合語句其處理前後的時頻圖為例來觀察新方法的效果，如圖 7 所示，其中包含了原始混合語音、真實之個別語音、及藉由新方法分離後之個別語音所對應的（強度）時頻圖。從圖中可看出，所提方法成功分離了混合語音、所得之語音與真實個別語音在時頻圖上非常相似。此驗

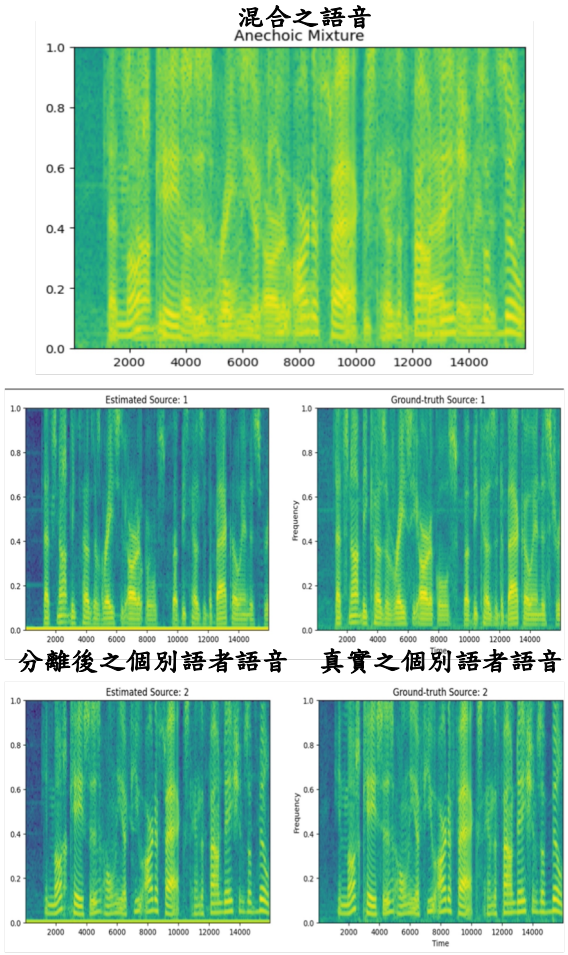

圖 7：ESC MA-SD Net 時頻圖分析結果

證了新方法 ESC MA-SD Net 在此語句上之成功的分離效果。

6 結論與未來期望 (Conclusion and future works)

在本研究中，我們提出使用殘差連接模塊(Residual Conformer Block)以及多注意力模塊(Multi-view Attention block)來改良原始之 SuDoRM-RF 語音分離模型，來減少訓練中可能遺失掉的重要資訊，且可以得到更全面的特徵呈現。而初步實驗結果證實所提方法可提升 SuDoRM-RF 之人聲分離之功效。在未來希望透過其 upsampling 不須額外增加參數的特性，微調 ESC MA-SD Net 的模塊以達到低參數特性、即可與其他模型 (如 TasNet) 同等甚至更佳的結果。

參考文獻 (References)

Amodei, D., Anubhai, R., Battenberg, E., et al. 2016. *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.*

DeLiang Wang, Jitong Chen, 2018. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10) 1702 – 1726* https://ieeexplore.ieee.org/document/8369155

Efthymios Tzinis, Zhepei Wang, Paris Smaragdis. 2020. Sudo rm -rf: Efficient Networks for Universal Audio Source Separation. 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP). https://ieeexplore.ieee.org/document/9231900

Efthymios Tzinis, Shrikant Venkataramani, Zhepei Wang, Cem Subakan, and Paris Smaragdis, "Two-step sound source separation: Training on learned latent targets," *in Proc. ICASSP, 2020.*

Hyun Joon Park, Byung Ha Kang, Wooseok Shin, Jin Sob Kim, Sung Won Han. 2022. MANNER: Multi-view Attention Network for Noise Erasure. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://ieeexplore.ieee.org/document/9747120

Meta AI "Speech Separation on WSJ0-2mix", https://paperswithcode.com/sota/speech-separation-on-wsj0-2mix (online), retrieved on Feb. 14, 2023

Morten Kolbæk, Zheng-Hua Tan, Senior Member, *IEEE*, Søren Holdt Jensen, and Jesper Jensen. (2020). *IEEE/ACM Transactions on Audio, Speech, and Language Processing (28), 825-838.* https://ieeexplore.ieee.org/document/8966946

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(12), 1849-1858. https://doi.org/10.1109/TASLP.2014.2352935

Williamson D S, Wang Y, Wang D L. Complex ratio masking for monaural speech separation[J]. *IEEE/ACM transactions on audio, speech, and language processing, 2015, 24(3): 483-492.* https://ieeexplore.ieee.org/document/7364200

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,* vol. 27, no. 8, pp. 1256–1266, 2019.