Advanced Personal Voice Activity Detection through Attention Score module with Conformer Block and FiLM Layers

通過 Conformer 模塊與特徵線性調製層結合自注意力分數損失進行個人語音活動偵測

Ruei-Xian Chang[1], En-Lun Yu[2], Berlin Chen[2], Shih-Chieh Huang[3] and Jeih-Weih Hung[1]

[1]National Chi Nan University, [2]National Taiwan Normal University, [3]Realtek Semiconductor Corp.

s110352016@ncnu.edu.tw, enlunyu@ntnu.edu.tw, berlin@ntnu.edu.tw,

eric.sc.huang@realtek.com, jwhung@ncnu.edu.tw

摘要

近年來，深度學習已成為全球熱門話題，作為人工智慧的一個重要分支，越來越多學者致力於推動這一領域的發展，語音處理技術在科技的進步下也有了大幅度的進展。語音活動檢測(voice activity detection, VAD)是一種重要的語音預處理步驟，通常位於其他語音處理任務之模型的上游，藉由分隔出待處理的語音段落，可以有效降低所串接之語音處理模型的運算複雜度。在本論文中，我們針對語者特定之個人式 VAD (personal VAD, PVAD) 的模型加以開發改進，在提出的新 PVAD 架構中，我們參照了 AS-pVAD 模型當中的自注意力分數模塊(AS block)、將其整合特徵線性調製層(FiLM layers)以更有效率地整合語者資訊、同時將一般 PVAD 之 LSTM 模型架構以 Conformer block 加以取代。經過這樣的模型改良，實驗結果顯示相較於基礎模型架構，目標語者平均精確度增加將近 5%，在其他各項評估分數也有顯著的提升。

Abstract

Deep learning's growing prominence in artificial intelligence has significantly impacted speech processing. Voice activity detection (VAD) acts as a crucial pre-processing step, isolating speech segments for further processing by downstream models, thereby reducing their computational burden. This paper addresses advancements in speaker-specific personal VAD (PVAD) models. Our proposed PVAD architecture leverages the self-attention score module (AS block) from the AS-pVAD model. To more effectively integrate speaker information, we incorporate feature-wise linear modulation layers (FiLM layers) into the architecture. Additionally, we replace the standard Long Short-Term Memory (LSTM) architecture with a Conformer block for enhanced performance.

Experimental results demonstrate a significant improvement in target speaker accuracy (nearly 5% increase) and other evaluation metrics compared to the baseline model. These findings highlight the effectiveness of the proposed modifications in speaker-specific VAD tasks.

關鍵字：語音活動檢測、自注意力分數模塊、特徵線性調製層

Keywords: voice activity detection, attentive score block, feature-wise transform layers, conformer

1. 緒論 (Introduction)

語音活動偵測(Voice Activity Detection, VAD)(Florian Eyben et al., 2013)是一種常見的語音預處理系統，其檢測語音活動是否為人聲及非人聲，因其過濾非語音的特性，常被作為各種語音相關任務、如自動語音辨識(Automatic Speech Recognition, ASR)(Takenori Yoshimura et al., 2020)、語者驗證(Speaker Verification, SV)(Li Wan et al., 2018)和語者分離(speaker separation)(Quan Wang et al., 2018)等架構的前端，以降低系統的資源消耗。

在實際應用場景中，由於通常存在語者干擾與人聲雜訊 (babble noise) 等具有類似語音成分卻不須被關注的語音，導致 VAD 系統在進行偵測時有較高的誤判率，進而使後端的處理系統對於誤判為語音的訊號進行多餘的處理，產生消耗計算資源（如 CPU、內存和電池）、降低處理速度的問題，這對於資源有限的語音處理系統影響尤其明顯。因此，一個優異的 VAD 系統，應該在實際應用場景中精確地區隔語音與非語音片段、同時，在目標語者與語音接收或處理裝置進行對話時，才觸發下游高功率元件進行下一步處理。為

此，文獻(Shaojin Ding et al., 2020)對音框為單位(frame-wise)的 VAD 的架構衍伸出個人語音活動偵測 (Personal Voice Activity Detection, PVAD)的系統，這裡稱為 PVAD 1.0。相較於 VAD，PVAD 僅識別目標語者的語音、忽略非目標語者與純雜訊的片段、更加關注在於目標語者的語音活動，進一步減少後端計算資源的消耗。PVAD性能的提升，有助於個人化設備和服務（如智能家具和藍芽音控設備及智慧型 3C 產品）的發展與應用。

近年來，有諸多的 PVAD 相關研究成果已被發表，包含提出原始 PVAD 1.0 的研究團隊，也提出了進階版，稱作 PVAD 2.0 (Shaojin Ding et al.,2022)，另外，一個基於注意力得分(attention score)機制的輕量級 PVAD (Attentive Score Personal Voice Activity Detection, AS-pVAD)(Fenting Liu et al.,2024)其效能相當優異，AS-pVAD 內部包含了高效提取語者嵌碼(speaker embedding)的模型，除了以低參數量模型解決現有諸多 PVAD 模型對於語者嵌碼系統的外部需求外，AS-pVAD 藉由兩階段偵測的方式，使整體架構對於未註冊語者(unenrolled speaker) 的語句也能發揮一般的 VAD 之效能。此外，AS-pVAD 模型中也採用了一種創新的計算注意力得分損失函數(Attentive Score)之模組，使整體模型更關注於目標語者相關的聲學特徵，使整體 PVAD 架構對於目標語者之嵌碼內容能學習地更加完整。

儘管 AS-pVAD 的效能優異，我們認為它仍然有幾項待改進的部分：

1. 注意力模組效能會隨語者嵌碼(speaker embdding)的表示能力而有顯著影響，因此，使用表示能力較差的語者嵌碼可能會使模型的學習表現不佳，注意力模組對整體 PVAD 的貢獻幅度有限。

2. 注意力模組僅透過串接來結合聲學特徵及語者嵌碼，我們認為在這兩部分應有更具效率的方式來進行整合。

在本文中，我們提出新的一個 PVAD 架構，為上述原始 AS-pVAD 的潛在問題提出改善的方案，實驗結果表示，我們新提出的 PVAD 架構，能夠在相似的模型參數量時，取得更大效益的結果。

以下是本文之內容安排：第二節簡要回顧了 PVAD 1.0 (Shaojin Ding et al., 2020)及 PVAD 2.0 (Shaojin Ding et al., 2022) 模型及提出的新方法流程，第三節包含了實驗設置與實作細節，第四節包含了實驗結果與討論，最後，第五節則為本文結論。

2. 提出方法 (Proposed Method)

2.1 回顧個人語音偵測與個人語音偵測 2.0

PVAD 1.0 法(Shaojin Ding et al.,2020)先從預訓練的語者驗證模型提取目標語者嵌碼(Speaker Embedding)，再將語音從特徵擷取器(Feature Extractor)提取的對數梅爾器組能量音訊特徵(log Mel-filterbank energies)進行串接，輸入個人語音偵測主架構，最後送進輸入線性層，得到以音框為單位(frame-wise)的目標語者決策機率：

$$\hat{F}_t = [F_t; e^{target}], \tag{1}$$

$$p = PVAD(\hat{F}_t), \tag{2}$$

其中F_t為音框 t 之音訊特徵，e^{target}為目標語者嵌碼，p為 PVAD 輸出之該音框對於到目標語者之機率。

然而，在上述的 PVAD 1.0 架構中，語音特徵透過串接特徵來訓練模型可能不是最佳效益的方法，其進階版 PVAD 2.0 (Shaojin Ding et al.,2022)提出使用特徵線性調製層(Feature-wise Linear Modulation layer, FILM layer)代替串接進行特徵融合。因此，PVAD 2.0 求取目標語者的機率過程如下：

$$\overline{F_t} = LSTM(F_t), \tag{3}$$

$$\widetilde{F_t} = \gamma(e^{target}) \cdot \overline{F_t} + \beta(e^{target}), \tag{4}$$

$$p = L(\widetilde{F_t}), \tag{5}$$

其中，$LSTM(\cdot)$為長短期記憶層操作層、其對於音訊特徵加以編碼而成$\overline{F_t}$、$\gamma(\cdot)$和$\beta(\cdot)$分別是作用於目標語者嵌碼e^{target}之 FiLM layer 的縮放和偏移向量，$L(.)$為一線性層(linear layer)。

2.2 注意力分數模塊(AS)

如前所述，文獻(Fenting Liu et al.,2024)在 PVAD 的框架下提出了注意力分數模塊(Attentive Score Block)，此模塊對於目標語者嵌碼及音訊特徵的串接$\hat{F}_t = [F_t; e^{target}]$（如式

(1)所示），透過兩個卷積層提取兩者的相似度分數：

$$M_t = AS(\hat{F}_t), \quad (6)$$

其中，$AS(.)$為上述兩個卷積層之運算，此相似度分數M_t和$\hat{F}_t$進行逐元素相乘，獲得加權特徵$F_{AS,t}$：

$$F_{AS,t} = \hat{F}_t \odot M_t。 \quad (7)$$

上述之使用串聯法結合目標語者嵌碼e^{target}及音訊特徵F_t來求取相似度(注意力)分數M_t，可能有幾項潛在缺點：

1. 維度特徵增加；
2. 串聯內容不同的資訊模態使得模型學習能力降低。

2.3 與注意力分數模塊結合特徵線性調製層 (FiLM-AT-PVAD)

基於上述之原始注意力分數模塊在輸入上的缺點，本論文提出一個進階版的注意力分數模塊，其流程圖如圖一(a)表示，它首先使用目標語者嵌碼e^{target}，求取一特徵線性調製層(FiLM Layer)，其縮放與偏移向量分別以γ和β表示，此 FiLM 層運作在音訊特徵F_t，得到新特徵

$$\tilde{\mathbf{F}}_t = \text{FiLM}(\mathbf{F}_t) = \gamma(\mathbf{e}^{\text{target}}) \cdot \mathbf{F}_t + \beta(\mathbf{e}^{\text{target}}), \quad (8)$$

$\tilde{\mathbf{F}}_t$ 相當於整合了目標語者嵌碼e^{target}與音訊特徵$\mathbf{F}_t$兩方的資訊，接著，$\tilde{\mathbf{F}}_t$ 用於求取相似度(注意力)分數M_t、進而將M_t與音訊特徵F_t點乘而得到輸出加權特徵$F_{AS,t}$，此$F_{AS,t}$通過一個由數層 LSTM 與全連接層(fully connected layer, FC)構成的分類層(classification layer)，求取各輸入音框對應之目標語者的機率。此 PVAD 整體流程圖如圖一(b)所示，為了敘述簡易起見，我們以將其命名為 **FiLM-AT-PVAD**。其中，我們預設音訊特徵為透過 Feature Extractor 模組產生的對數梅爾器組能量音訊特徵F_t，而分類層中的 LSTM 層數為 2。

針對上述所新提出的 **FiLM-AT-PVAD** 架構，我們同時對其加以變化，探索是否能達到更佳的 PVAD 效能，以下，是我們研發出的四種變形：

1. **FiLM-AT-PVAD$_{\text{L1, L1}}$:**
   在 **Feature extractor 模組後多加一層 LSTM，而分類層中的 LSTM 層數由 2 降為 1**，其目的是**藉由 LSTM 層對於音訊特徵加以編碼優化、卻簡化分類層的參數量**，探索在不明顯更動原 FiLM-AT-PVAD 的複雜度前提下，觀察其效能是否能有所提升，我們將其命名為 FiLM-AT-PVAD$_{\text{L1, L1}}$, 其下標 L1, L1 分別代表注意力模組前端為 1 層 LSTM 與後端為 1 層 LSTM。
2. **FiLM-AT-PVAD$_{\text{L1, L2}}$:**
   在 Feature extractor 模組後多加一層 LSTM，而分類層中的 LSTM 層數維持 2，其目的是**探討額外增添 LSTM 的預處理對於音訊特徵加以優化後**，是否能帶來改善。我們將其命名為 FiLM-AT-PVAD$_{\text{L1, L2}}$, 其下標 L1, L2 分別代表注意力模組前端為 1 層 LSTM 與後端為 2 層 LSTM。
3. **FiLM-AT-PVAD$_{\text{C1, L1}}$**
   在 **Feature extractor 模組後多加一層 Conformer 模塊，而分類層中的 LSTM 層數由 2 降為 1**，在此 Conformer 模塊(Anmol Gulati et al.,2020)作為音訊特徵的編碼器，Conformer 模塊為自注意力與卷積神經網路的結合，可進行全局自注意力交互學習，處理時間序列任務，其中自注意力機制幫助模型捕捉時間序列資料中的長期依賴關係，而卷積層捕捉局部特徵，前饋模組進一步變換特徵，增加非線性處理能力以增強模型的整體表達能力，有效的捕捉相對偏移的局部相關性。由於 Conformer 層較 LSTM 層精細與複雜，預期它的參與會提升 PVAD 之效能，但會增加運算複雜度。我們將其命名為 FiLM-AT-PVAD$_{\text{C1, L1}}$, 其下標 C1, L1 分別代表注意力模組前端為 1 層 Conformer 與後端為 1 層 LSTM。
4. **FiLM-AT-PVAD$_{\text{C1, L2}}$**
   在 **Feature extractor 模組後多加一層 Conformer 模塊，而分類層中的 LSTM 層數維持為 2**，因此我們將其命名為 FiLM-AT-PVAD$_{\text{C1, L2}}$, 其下標 C1, L2 分別代表注意力模組前端為 1 層 Conformer 與後端為 2 層 LSTM。

這四種 FilM-AT-PVAD 之變形的流程圖分別如圖二(a)(b)與圖三(a)(b)所示。另外，根據上述的命名法，由於原始 FiLM-AT-PVAD 在 Feature extractor 模組後並未增加

任何模塊，而分類層的 LSTM 層數為 2，因此我們將其命名為 FiLM-AT-PVAD$_{\emptyset, L2}$，其下標$\emptyset$, L2 分別代表注意力模組前端無額外模組、後端為 2 層 LSTM。

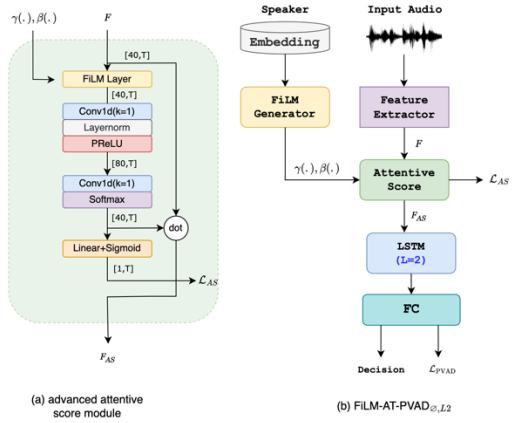

圖 1：(a)進階式注意力分數模塊 (b)原始 FiLM-AT-PVAD (FiLM-AT-PVAD$\emptyset$, L2)架構

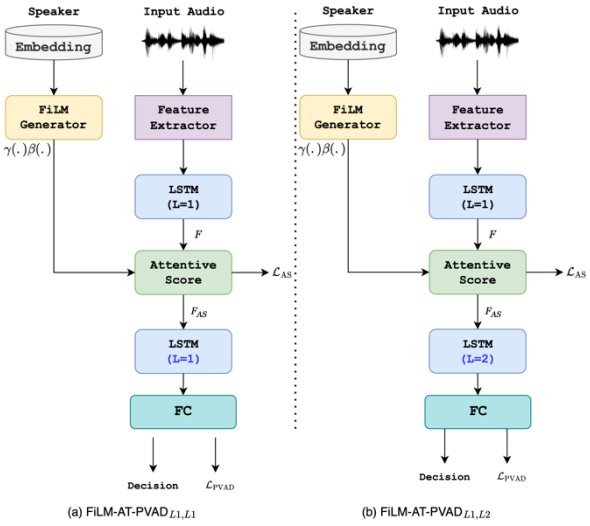

圖 2：(a) FiLM-AT-PVAD$_{L1, L1}$ 架構 (b) FiLM-AT-PVAD$_{L1, L2}$ 架構，二者皆使用一層的 LSTM 作為音訊特徵的編碼器，不同在於後端之 LSTM 層數

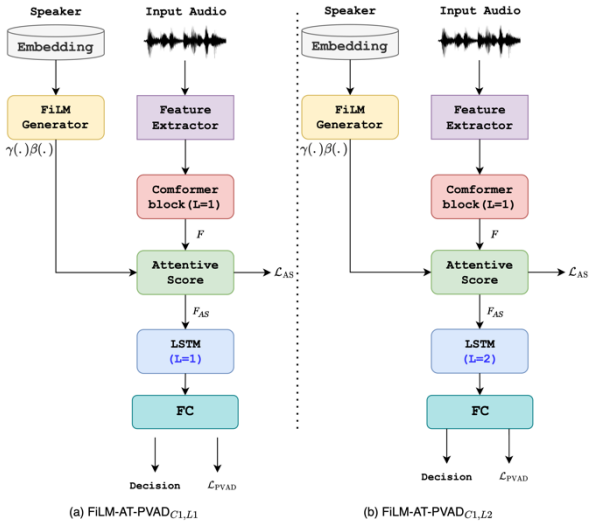

圖 3：(a) FiLM-AT-PVAD$_{C1, L1}$ 架構 (b) FiLM-AT-PVAD$_{C1, L2}$ 架構，二者皆使用一層的 Conformer 作為音訊特徵的編碼器，不同在於後端之 LSTM 層數

2.4 損失函數(Loss function)

為了進行二元分類的PVAD任務，我們使用二元交叉熵(Binary Cross Entropy, BCE)(Zhilu Zhang et al., 2018)作為注意力分數模塊的損失函數之一，其計算方法如下：

$$L_{PVAD} = \frac{1}{T}\sum_{t=0}^{T-1} BCE(y_t, p_t) \quad (9)$$

上式中，$y_t \in \{0,1\}$和$p_t \in \{0,1\}$分別代表音框 t 的真實標籤值(目標語者為 1，其他為 0)和 PVAD 的預測標籤值，T是總時間框數量。

此外，基於(Fenting Liu et al.,2024)，我們引入了注意力分數模塊損失函數L_{AS}來專門訓練注意力分數模塊。注意力分數模塊損失通過計算真實標籤y_t 與估計的加權分數$\widetilde{m}_t$之間的均方根誤差(Mean Squared Error)損失，促使注意力分數模塊專注於學習跨多模態信息的相似性。

$$\widetilde{m}_t = \text{Sigmoid}(\text{L}(M_t)) \quad (10)$$

$$L_{AS} = \frac{1}{T}\sum_{t=0}^{T-1}(y_t - \widetilde{m}_t)^2 \quad (11)$$

公式(10)中的M_t是在公式(6)中顯示的相似度分數，Sigmoid (·)和 L(·)分別表示 Sigmoid 層和線性層操作。最終，將上述兩個損失相加，得到訓練整個 PVAD 模型的總損失：

$$L_{total} = L_{PVAD} + L_{AS}。 \quad (12)$$

3. 實驗設置 Experiment setup

3.1 實驗數據集

我們採用了 LibriSpeech 訓練集(Vassil Panayotov et al., 2018)來評估各種 PVAD 的效能。該訓練集包含三個子集，總計 960 小時來自 2338 位不同說話者的語音數據：訓練集 train-clean-100 和 train-clean-360 提供了總共 460 小時的乾淨語音，而 train-other-500 提供了 500 小時的雜訊語音。同樣地，LibriSpeech 的測試集包含乾淨和雜訊語音，總計 10 小時來自 73 位語者的語音。為了進行實驗，數據集必須包含多個語者之語音的串接。因此我們首先利用均勻分布、選擇一到三作為語者的數目，並將它們的語音連接起來，之後隨機選擇其中一位語者作為目標語者。在這個過程，我們以 0.2 的比率加入未包含目標語者語音的多語者語音，避免模型過度學習目標語者的特徵。

關於說話者嵌碼，我們為每位說話者隨機挑選語音並輸入預訓練的說話者驗證模型，以生成窗級的 d 向量。這些 d 向量經 L2 正規化並平均，進而產生作為對應整段語句(utterance-based)之目標語者嵌碼的之 d 向量 e^{target}。

此外，為了防止學習過擬合並使模型具有強健性，我們使用了 MTR 的數據增強技術(Chanwoo Kim et al., 2017)，功能是引入了具有不同房間模擬脈衝響應的隨機噪聲源，這種數據增強技術能有效地提高模型對不同噪聲和混響條件的模擬，讓模型能夠在更廣泛的現實環境中表現良好。

3.2 實作細節

我們對語音數據及提取 40 維的梅爾濾波器能量作為原始聲學特徵，其框架大小為 25 毫秒，間隔為 10 毫秒。對於我們所實驗的各種 PVAD 模型的參數設定，詳細介紹如下：

(1) AT-PVAD 基礎模型：
此為使用原始 AS-pVAD 之 AS 模組(以語者嵌碼與音訊特徵為輸入，如式(6)(7))，其輸出再經由兩層 LSTM 與兩層線性層組成的分類層，這兩層 LSTM 各有 64 個單元(unit)。

(2) FiLM-AT-PVAD$_{\emptyset, L2}$：注意力模塊後端的兩層 LSTM 各有 64 個單元(unit)。

(3) FiLM-AT-PVAD$_{L1, L1}$：
注意力模塊前端的一層 LSTM 包含 40 個單元，而後端的一層 LSTM 包含 64 層單元。

(4) FiLM-AT-PVAD$_{L1, L2}$：
注意力模塊前端的一層 LSTM 包含 40 個單元，而後端的兩層 LSTM 各有 64 個單元。

(5) FiLM-AT-PVAD$_{C1, L1}$
注意力模塊前端的一層 Conformer 模塊是由兩個前饋模塊、一個自注意力模塊及一個卷積模塊所組成，輸出 40 個單元，而後端的一層 LSTM 包含 64 個單元。

(6) FiLM-AT-PVAD$_{C1, L2}$
注意力模塊前端的一層 Conformer 模塊是由兩個前饋模塊、一個自注意力模塊及一個卷積模塊所組成，輸出 40 個單元，而後端的兩層 LSTM 各有 64 個單元。

對於整體實驗，我們使用 PyTorch (Adam Paszke et al., 2019)實現了所有模型。在模型訓練中，我們最初使用 Adam 優化器(Diederik P. Kingma et al., 2014)，在第一個訓練周期的學習率為1×10^3，隨後在接下來的訓練周期中將學習率降低到 1×10^5。

4. 實驗結果與討論 Results&Discussions

我們採用了多種不同指標來評估各個 PVAD 模型。其中，平均精度(AP)為目標說話者語音(tss)及非目標語音與非語音(ntss&ns) 精確率-召回率曲線下的面積，平均精度期望值(mAP)則是再對平均精度進行加權平均，數值越高，表示模型在檢測目標語者音框的精確性越高；準確率(Acc.)(%)是正確檢測的音框數與總檢測音框數的比率，反映模型在區分目標語者語音與非目標語者音框方面的準確性。由於我們所採用的數據集所包含的正樣本數少於負樣本，因此我們在使用平均精度期望值時會是一個相當重要的指標，此外，我們還透過考慮模型參數的數量(Par.)來評估模型在資源有限上的適用性。

表 1：各種 PVAD 對應的目標語者平均精度、非目標語者語音平均精度、平均精度期望值、準確率及模型參數量，其中 AT-PVAD baseline、FiLM-AT-PVAD$_{\emptyset, L2}$、FiLM-AT-PVAD$_{L1, L2}$、FiLM-AT-PVAD$_{C1, L2}$ **後端皆有 2 層 LSTM**

model	tts	ntss&ns	mAP	Acc.(%)	Par.(k)
PVAD 1.0	88.4	94.7	92.2	84.34	130.24
PVAD 2.0	90.8	96.0	94.1	86.58	97.60
AT-PVAD 基礎模型	86.8	94.4	91.6	83.79	**84.95**
FiLM-AT-PVAD$_{\emptyset, L2}$	90.3	95.1	93.4	86.08	92.03
FiLM-AT-PVAD$_{L1, L2}$	88.8	94.7	92.5	84.84	105.15
FiLM-AT-PVAD$_{C1, L2}$	**95.9**	**98.3**	**97.5**	**91.99**	131.95

表 2：PVAD 2.0 與兩種較輕量之 PVAD 的評估分數，其中 FiLM-AT-PVAD$_{L1, L1}$、FiLM-AT-PVAD$_{C1, L1}$ **後端皆有 1 層 LSTM**

model	tts	ntss&ns	mAP	Acc.(%)	Par.(k)
PVAD 2.0	90.8	96.0	94.1	86.58	97.60
FiLM-AT-PVAD$_{L1, L1}$	91.2	95.5	94.0	86.74	**71.87**
FiLM-AT-PVAD$_{C1, L1}$	**95.6**	**98.2**	**97.3**	**91.47**	98.67

4.1 比較使用 ASPVAD 提出的方法

表 1 顯示了 PVAD 1.0 (Shaojin Ding et al., 2020)、PVAD2.0 (Shaojin Ding et al., 2022)及 AT-PVAD 基礎模型和提出的三種較複雜之模型（後端有 2 層 LSTM）進行比較，從此表格當中，我們有下列觀察與分析：

1) 三種提出的模型(FiLM-AT-PVAD$_{\emptyset, L2}$、FiLM-AT-PVAD$_{L1, L2}$、FiLM-AT-PVAD$_{C1, L2}$)表現都優於 AT-PVAD 基礎模型，由此可以看出我們所提出之注意力分數模塊（藉由 FiLM 整合語者資訊e^{target}與音訊特徵$\mathbf{F}_t$）優於原始注意力分數模塊（單純串接e^{target}與音訊特徵$\mathbf{F}_t$）。

2) AT-PVAD 基礎模型效果相對較差，而使用同一種（原始）注意力分數模塊之 AS-pVAD 法(Fenting Liu et al.,2024)、根據其文獻之成果，效果極佳，這可能是因為它額外採用了更先進的語者嵌碼提取器（ECAPA-TDNN）和聲學特徵編碼器（TCNN），為注意力分數模塊提供了更優越的輸入特徵。

3) FiLM-AT-PVAD$_{L1, L2}$ 相較於 FiLM-AT-PVAD$_{\emptyset, L2}$ 額外於注意力分數模塊前採用一層 LSTM 作為音訊特徵編碼，效果卻反而變差，然而，當使用一層 Conformer 來作作為音訊特徵編碼的 FiLM-AT-PVAD$_{C1, L2}$，在各項指標上都是最佳的，但由於 Conformer 的複雜性，使其對應的模型參數量最多。

4.2 分析模型簡化後的結果

表 2 顯示了 PVAD2.0 及提出的兩種簡化的模型（FiLM-AT-PVAD$_{L1, L1}$、FiLM-AT-PVAD$_{C1, L1}$）的各項指標，觀察此表、並與表 1 相較，我們觀察到：

1) FiLM-AT-PVAD$_{L1, L1}$ 相對於 FiLM-AT-PVAD$_{L1, L2}$ 少了一層 LSTM，雖然模型變小，但各項指標分數皆提升了許多，其中幾項重要的數據：目標語者平均精度(tts)及準確率(Acc.)超過了 PVAD 2.0(AP:91.2 vs. 90.8, Accuracy:86.74 vs. 比 86.58)，這顯示出多一層 LSTM 於後端的分類器效果不一定比較好，對於注意力分數模塊捕捉到的特徵在經過二層 LSTM 之分類器時可能造成梯度消失的結果。

2) FiLM-AT-PVAD$_{C1, L1}$ 相較於 FiLM-AT-PVAD$_{C1, L2}$ 少了一層 LSTM，效果僅微幅下降，但模型

大小大幅減少（98.67 k vs. 131.949k），這使得它更適合應用在資源有限的設備上，在實際應用中具有顯著優勢。

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

本研究拓展新穎之 AS-pVAD 其注意力分數模塊(AS block)的概念，採用特徵線性調製層(FiLM layer)來整合語者嵌碼與音訊特徵，並提出用 Conformer 模塊來對音訊特徵作編碼，而初步實驗證實我們所提出的方法在實際應用中能夠顯著讓 PVAD 模型效果提升，在未來展望上，希望能透過所提出之 FiLM-AT-PVAD$_{C1, L1}$ 模型作為基礎、設計出表現更佳更輕量的 PVAD 模型。

參考文獻 (References)

Florian Eyben, F. Weninger, S. Squartini, and B. Schuller. 2013. *Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies.* 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Takenori Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe. 2020. *End-to-End Automatic Speech Recognition Integrated With CTC-Based Voice Activity Detection.* 2020 IEEE International Conference on Acoustics, Speech and Signal Processing.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. *Generalized end-to-end loss for speaker verification.* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. *Speaker diarization with LSTM.* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, Ignacio Lopez Moreno. 2020. *Personal VAD: Speaker-Conditioned Voice Activity Detection.* The Speaker and Language Recognition Workshop

Fenting Liu, Feifei Xiong, Yiya Hao, Kechenying Zhou, Chenhui Zhang, Jinwei Feng. 2024. *AS-pVAD: A Frame-Wise Personalized Voice Activity Detection Network with Attentive Score Loss.* ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Shaojin Ding, Rajeev Rikhye, Qiao Liang, Yanzhang He, Quan Wang, Arun Narayanan, Tom O'Malley, Ian McGraw. 2022. *Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition.* Interspeech.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition.* Interspeech.

Jiawei Xiao and Peng Lu. 2024. *A Hybrid Model of Conformer and LSTM for Ocean Wave Height Prediction.* Applied Sciences.

Zhilu Zhang and M. R. Sabuncu. 2018. *Generalized cross entropy loss for training deep neural networks with noisy labels.* Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc.

Vassil Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. *Librispeech: An asr corpus based on public domain audio books.* 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani. 2017. *Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. Proc.* Interspeech 2017, pp. 379–383.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc.

Diederik P. Kingma, Jimmy Ba. 2014. *Adam: A Method for Stochastic Optimization.* arXiv preprint arXiv:1412.6980