我將根據您提供的學術論文《2025_naacl-lTowards Efficient and Multifaceted Computer-assisted Pronunciation Training Leveraging Hierarchical Selective State Space Model and Decoupled Cross-entropy Loss.pdf》來回答您的問題。這份文件詳細描述了HMamba模型的架構、方法論及實驗結果，我會聚焦於模型中的資料流變化，包括輸入輸出維度及各層運算，以助您快速理解核心機制。

現在，讓我們深入分析HMamba模型中的資料流變化和運算：

HMamba模型旨在同時處理自動發音評估 (APA) 和錯誤發音檢測與診斷 (MDD) 任務，其核心是分層結構，依據語言的粒度（如音素、單詞和語句級別）進行處理。

1. 輸入與輸出

- **模型整體輸入**
    - **語音訊號 (u)**：由L2學習者發出的時間序列語音訊號。
    - **參考文字提示 (p)**：包含N長度標準音素序列的參考文字提示。
- **模型整體輸出**
    - **APA 任務的評分 (sg)**：針對不同語言粒度 (g) 預測的一組發音方面分數。粒度包括音素級 ($gphn$)、單詞級 ($gwrd$) 和語句級 ($gutt$)。
    - **MDD 任務的錯誤狀態 (e) 和診斷輸出 (y)**：檢測相對於參考提示的錯誤狀態，並生成學習者實際發出音素的正確診斷輸出。

2. 模型每一層的運算及維度變化

HMamba模型採用「自下而上 (bottom-up)」的分層建模結構，並將APA和MDD模組整合在其中，每個模組包含多個迴歸器（針對APA）和一個分類器（針對MDD）。

階段一：特徵提取 (Feature Extraction)

1. **聲學特徵提取 (Acoustic Feature Extraction)**
    - **輸入**：原始語音訊號 ($\mathbf{u}$) 和參考文字提示 ($\mathbf{p}$)。
    - **運算**：

        - 使用預訓練聲學模型作為對齊器，識別音素邊界（包括靜音）。
        - 提取基於音素發音優劣度 (GOP) 的特徵。
        - 提取韻律特徵，例如音素持續時間 (phone duration) 和均方根能量的統計數據 (statistics of root mean squared energy)。
        - 整合自監督學習 (SSL) 特徵，包括wav2vec 2.0、HuBERT和WavLM。在這些SSL特徵被連接 (concatenation) 之前，會對其應用10%的丟棄率 (dropout rate) 以處理維度差異。
        - 將所有這些聲學特徵（GOP、持續時間、能量、wav2vec 2.0、HuBERT、WavLM）連接成一個綜合向量 $\mathbf{a}_t$。
        - 通過一個線性層進行投影：$\mathbf{x}_t = \mathbf{W}\mathbf{a}_t + \mathbf{b}$，其中 $\mathbf{W}$ 和 $\mathbf{b}$ 是可訓練參數。
    - **輸出**：聲學特徵序列 $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}\}$，與參考文字提示 $\mathbf{p}$ 對齊。
    - **維度變化**：$\mathbf{x}_t$ 的維度被設定為128維。序列長度為N，對應於標準音素序列的長度。
2. **語音學特徵提取 (Phonological Feature Extraction)**
    - **輸入**：聲學特徵序列 $\mathbf{X}$。
    - **運算**：
        - 從參考文字提示 $\mathbf{p}$ 中提取標準音素嵌入 (canonical phoneme embeddings) $\mathbf{E}_{phn}$，其中包含靜音 (SIL) 資訊。
        - 提取絕對位置嵌入 (absolute positional embedding) $\mathbf{E}_{abs}$。
        - 提取相對位置嵌入 (relative position embedding) $\mathbf{E}_{rel}$，使用標記如 [B] (begin)、[I] (internal)、[E] (end) 和 [S] (single-phone word) 來表示單詞中音素的相對位置；對於靜音，則分為 [LS] (long silence) 或 [SS] (short silence)。
        - 將這些嵌入特徵逐點相加 (point-wise added) 到 $\mathbf{X}$ 中。
    - **輸出**：音素級輸入特徵 $\mathbf{H}_{g0\ phn}$。

- **維度變化**：所有嵌入特徵的維度均為128維。因此，$\mathbf{H}g0\ phn$ 的維度仍為128維，序列長度為N。

階段二：分層Mamba模型 (Hierarchical Mamba Modeling)

HMamba模型包含三個層次的Mamba區塊，每個Mamba區塊的隱藏單元數設為128。一個Mamba區塊由雙向Mamba層 (BiMamba) 和前饋網路 (FFN) 組成。

1. **音素級建模 (Phone-level Modeling)**

   - **輸入**：音素級輸入特徵 $\mathbf{H}g0\ phn$。
   - **運算**：
     - 輸入 $\mathbf{H}g0\ phn$ 通過 Lp 個Mamba區塊。Lp 在此研究中設為3。
     - **Mamba區塊內部運算**：
       - 層歸一化 (LayerNorm)。
       - 雙向Mamba (BiMamba) 層：處理輸入 $\mathbf{N}gi$。它涉及線性層 (Linear)、翻轉操作 (Flip)、一維卷積 (Conv1D)（卷積核大小為4）、激活函數 (σ) 和選擇性狀態空間模型 (selective SSM) 算法。雙向Mamba會生成前向 ($\mathbf{S}gi\ \rightarrow$) 和後向 ($\mathbf{S}gi\ \leftarrow$) 序列特徵，並最終通過另一個線性層整合輸出。
       - 殘差連接 (residual connection)：BiMamba 的輸出與其輸入相加。
       - 層歸一化 (LayerNorm)。
       - 前饋網路 (FFN)。
       - 殘差連接 (residual connection)：FFN 的輸出與其輸入相加。
     - **APA 模組 (音素級)**：輸出 $\mathbf{H}gLp\ phn$ 傳播到 APA 模組。其中包含一個迴歸器（一個簡單的前饋網路 FFN），用於預測音素級別的方面分數 $s0\ gphn$ (準確度 accuracy)。
     - **MDD 模組 (音素級)**：輸出 $\mathbf{H}gLp\ phn$ 傳播到 MDD 模組。其中包含一個分類器（一個簡單的前饋網路 FFN）和一個softmax函數，協同學習每個時間步 t 的音素類別 C 的分佈 $\hat{y}$。診斷 yt 可通過對 $\hat{y}$ 應用argmax函數來識別。錯誤狀態 et 則通過比較 yt 和 pt 來直接檢測。

- **輸出**：音素級上下文表示 $\mathbf{H}gLp\ phn$。同時也直接輸出音素級APA分數和MDD診斷結果。
- **維度變化**：Mamba區塊保持輸入維度。因此，$\mathbf{H}gLp\ phn$ 的維度仍為128維，序列長度為N。音素級APA分數為標量輸出；MDD診斷為對應音素類別的輸出。

2. **單詞級建模 (Word-level Modeling)**

- **輸入**：來自音素級建模的 $\mathbf{H}gLp\ phn$，作為單詞級建模的輸入 $\mathbf{H}g0\ wrd$。
- **運算**：
  - 輸入 $\mathbf{H}g0\ wrd$ 通過 Lw 個Mamba區塊。Lw 在此研究中設為1。
  - 之後接一個一維卷積層 (1-D convolution layer) 來捕捉局部依賴性。此卷積層有256個卷積核，每個大小為3。
  - **APA 模組 (單詞級)**：單詞級的表示 $\mathbf{H}gLw\ wrd$ 輸入到單詞級 APA 模組。該模組包含三個迴歸器，分別預測單詞級別的準確度 ($s0\ gwrd$)、重音 ($s1\ gwrd$) 和總分 ($s2\ gwrd$)。
- **輸出**：單詞級表示 $\mathbf{H}gLw\ wrd$。同時也直接輸出單詞級APA分數。
- **維度變化**：
  - Mamba區塊的輸出 $\mathbf{H}'gLw\ wrd$ 維度保持在128維。
  - 經過一維卷積層後，由於卷積核的數量為256，輸出 $\mathbf{H}gLw\ wrd$ 的維度變為256維。序列長度與單詞序列對齊。單詞級APA分數為三個標量輸出。

3. **語句級建模 (Utterance-level Modeling)**

- **輸入**：來自單詞級建模的 $\mathbf{H}gLw\ wrd$，作為語句級建模的輸入 $\mathbf{H}g0\ utt$。
- **運算**：
  - 輸入 $\mathbf{H}g0\ utt$ 通過 Lu 個Mamba區塊。Lu 在此研究中設為1。
  - 之後利用注意力池化層 (attention pooling layer) 聚合隱藏資訊。注意力池化的權重 αi 是根據可學習向量 $\mathbf{w}$、連接的音素級和單詞級分數 $\mathbf{q}$ (即 [$s0\ gphn$, $s0\ gwrd$, $s1\ gwrd$, $s2\ gwrd$]) 以及一個可控溫度超參數 τ 計算的。

        - **APA 模組 (語句級)**：池化後的語句表示 $\mathbf{h}_{gutt}$ 輸入到語句級 APA 模組。該模組包含五個迴歸器，分別預測語句級別的準確度 ($s0\ gutt$)、完整性 ($s1\ gutt$)、流暢度 ($s2\ gutt$)、韻律 ($s3\ gutt$) 和總分 ($s4\ gutt$)。
    - **輸出**：池化後的語句表示 $\mathbf{h}_{gutt}$。同時也直接輸出語句級APA分數。
    - **維度變化**：
        - Mamba區塊的輸出 $\mathbf{H}_{gLu\ utt}$ 維度保持在256維。
        - 注意力池化後的輸出 $\mathbf{h}_{gutt}$ 是一個單一向量，其維度仍為256維（從256維序列池化而來）。語句級APA分數為五個標量輸出。

階段三：優化 (Optimization)

雖然不屬於模型層次的資料流，但優化是決定模型如何學習和調整其內部表示的關鍵。

- **APA 損失 ($\mathcal{L}_{APA}$)**：每個 APA 模組都使用均方誤差 (MSE) 進行優化。總體損失是各粒度級別損失的加權和。
- **MDD 損失 ($\mathcal{L}_{MDD}$)**：MDD 任務使用新型的解耦交叉熵損失 (deXent)，它將原始交叉熵損失分為針對錯誤發音和正確發音的兩個獨立損失，並通過調整權重 α 來平衡精確率和召回率。
- **總體損失 ($\mathcal{L}$)**：模型的總體損失是 APA 損失和 MDD 損失的加權和：$\mathcal{L} = \mathcal{L}_{APA} + \beta \cdot \mathcal{L}_{MDD}$，其中 β 是一個可調超參數。

總結來說：

- **輸入**：原始語音（高維度、時序）和文字提示（符號序列）。
- **聲學與語音學特徵提取**：將原始輸入轉換為128維的對齊音素級特徵序列。
- **音素級Mamba區塊 (Lp=3)**：處理128維序列，輸出保持128維，並同時進行音素級APA和MDD預測。
- **單詞級Mamba區塊 (Lw=1) + 1D卷積**：從音素級輸出繼承128維序列，Mamba區塊後接的1D卷積將維度提升至256維，並進行單詞級APA預測。
- **語句級Mamba區塊 (Lu=1) + 注意力池化**：從單詞級輸出繼承256維序列，Mamba區塊後接注意力池化層將序列壓縮為單一的256維向量，並進行語句級APA預測。

- **輸出**：最終的輸出包括多粒度（音素、單詞、語句）的多方面評估分數和音素級的錯誤發音診斷。

通過理解這種分層處理和維度轉換，您可以更清晰地看到HMamba如何從低級聲學語音特徵逐步構建更高級別的語言表示，並同時執行評估和診斷任務。

為了進一步分析，您可能會想了解 HMamba 在不同語言或口音數據集上的表現如何，因為目前的研究僅限於普通話為母語的L2英語學習者，這可能限制了模型的通用性。