第三章：模型架構與方法論

本章旨在詳細闡述我們為解決發音錯誤檢測問題所提出的模型架構與核心方法。我們將首先簡要回顧傳統硬對齊方法的侷限性，隨後引出並深入剖析作為本研究核心的「軟對齊」典範，包含其背後的注意力機制、數學模型，以及理論優勢。

3.1 回顧硬對齊典範之侷限

傳統的發音錯誤檢測系統，其根基深植於「硬對齊」（Hard Alignment）典範。此典範的核心操作可被形式化地定義為一個對齊函數 A。對於一個給定的音素序列 $P = (p_1, p_2, ..., p_N)$ 和一個聲學特徵序列 $X = (x_1, x_2, ..., x_T)$，該函數將每一個音素 p_i 映射到一個離散且互不重疊的時間區間 $[t_{start}(i), t_{end}(i)]$：

$$A(p_i, X) \rightarrow [t_{start}(i), t_{end}(i)]$$

其中 $1 \le t_{start}(i) \le t_{end}(i) \le T$。後續的發音評分，如 GOP 計算，將完全局限於這個由 X 截取出的子序列 $X_i = (x_{t_{start}(i)}, ..., x_{t_{end}(i)})$ 之中。

此一模型的根本缺陷在於其「剛性」（Rigidity）。它預設了語音中存在清晰、可準確劃分的音素邊界。然而，如緒論所述，真實語音流的時序結構受到語速、口音及協同發音等因素的嚴重影響，使得這種理想化的邊界假設在實踐中極易失效。一個微小的對齊失誤，便會導致後續的評分模組在錯誤的聲學證據上進行判斷，造成不可逆的誤差傳播（Error Propagation），這構成了該典範難以逾越的性能瓶頸。

3.2 軟對齊：一個基於注意力機制的典範

為克服上述侷限，我們引入「軟對齊」（Soft Alignment）思想。其核心是放棄尋找離散的、絕對的時間邊界，轉而為每一個目標音素 p_i 計算一個在整個聲學特徵序列 X 上的連續權重分佈 $a_i = (a_{i1}, a_{i2}, ..., a_{iT})$。此分佈中的每一個權重 a_{ij} 代表了第 j 個聲學幀 x_j 對於判斷音素 p_i 發音品質的「重要性」或「相關度」。

我們藉由**注意力機制（Attention Mechanism）**來實現這一目標。此機制借鑒了人類認知過程中聚焦於重要資訊的能力，其運作依賴於三個核心組件：查詢（Query）、鍵（Key）和值（Value）。

1. **查詢 (Query, Q)**: 代表了當前任務的「焦點」。在我們的模型中，每一個待評估的目標音素 p_i 都會被轉換為一個查詢向量 q_i。這個向量可由音素的嵌入（Embedding）或透過上下文編碼器生成，它攜帶了「我正在尋找音素 p_i 的聲學實現」的語義資訊。

2. 鍵 **(Key, K)**: 代表了被查詢序列中各個元素的「索引」或「標籤」。整個聲學特徵序列 X 經過一個線性變換後，生成鍵矩陣 $K = (k_1, k_2, ..., k_T)$。每一個鍵向量 k_j 都是對應聲學幀 x_j 的一種抽象表示，用於與查詢向量 q_i 進行相似度匹配。
3. 值 **(Value, V)**: 代表了被查詢序列中各個元素的「內容」或「實質」。與鍵相似，聲學特徵序列 X 也會經過另一個線性變換生成值矩陣 $V = (v_1, v_2, ..., v_T)$。值向量 v_j 是 x_j 的另一種表示，它包含了用於最終計算的豐富聲學資訊。

透過這三個組件的相互作用，模型能夠為每一個音素查詢 q_i 動態地計算出其在整個語音序列上的注意力分佈。

3.3 注意力機制之數學模型

本研究採用縮放點積注意力（Scaled Dot-Product Attention）作為實現軟對齊的核心演算法。對於任意一個音素查詢 q_i，其對應的上下文感知特徵向量 c_i 的計算過程可分解為以下三個步驟：

步驟一：計算相似度分數（Similarity Score）

模型首先計算查詢向量 q_i 與所有鍵向量 k_j $(j = 1, ..., T)$ 的點積，以衡量它們之間的相似度。為了防止因向量維度過高導致點積結果過大，從而使後續 Softmax 函數的梯度變得過於稀疏，我們對點積結果進行縮放。相似度分數 e_{ij} 的計算公式如下：

$$e_{ij} = \frac{q_i \cdot k_j^T}{\sqrt{d_k}}$$

其中，d_k 是鍵向量 k_j 的維度。此步驟的輸出是一個未經正規化的相關性分數向量 $e_i = (e_{i1}, e_{i2}, ..., e_{iT})$。

步驟二：計算注意力權重（Attention Weights）

接著，應用 Softmax 函數將原始的相似度分數向量 e_i 轉換為一個機率分佈，即注意力權重分佈 α_i。這一步是實現軟對齊的關鍵。

$$\alpha_{ij} = \text{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{t=1}^{T} \exp(e_{it})}$$

所得到的權重 α_{ij} 均介於 0 和 1 之間，且其總和為 1 $(\sum_{j=1}^{T} \alpha_{ij} = 1)$。這組權重分佈 α_i 精確地量化了在評估音素 p_i 時，模型對每一聲學幀 x_j 的「關注程度」。

步驟三：生成上下文向量（Context Vector）

最後，將計算出的注意力權重 a_{ij} 作為加權係數，對所有的值向量 v_j 進行加權求和，從而生成最終的上下文向量 c_i。

$$c_i = \sum_{j=1}^{T} a_{ij} v_j$$

這個上下文向量 c_i 是整個聲學序列 X 中與音素 p_i 相關資訊的動態融合與濃縮。它不再依賴於任何預先劃定的僵硬邊界，而是由模型根據數據本身自主學習到的、最優的特徵組合。

3.4 相較於硬對齊之理論優勢

基於上述模型，軟對齊典範在理論上展現出超越硬對齊的顯著潛力：

1. **對時間變異的自適應性 (Adaptability to Temporal Variation)**：硬對齊對每一幀的處理是二元的（屬於或不屬於該音素）。而軟對齊的權重分佈是連續的。對於發音短促的爆破音，其權重分佈可能呈現尖銳的單峰；對於時長較長的元音，其分佈則可能更為平坦寬廣。這種靈活性使其能夠自然地適應不同音素及個體的時長變化。
2. **對協同發音的魯棒性 (Robustness to Coarticulation)**：協同發音意味著音素間的聲學特徵會相互滲透，尤其是在音素過渡區。硬對齊會粗暴地切斷這些過渡區，導致資訊損失。而注意力機制則沒有邊界限制，它可以學會為目標音素的核心區域賦予高權重，同時也為其前後的關鍵過渡幀分配適當的次要權重，從而捕捉到更完整的發音樣貌。
3. **端到端可微分性 (End-to-End Differentiability)**：整個注意力計算過程（矩陣乘法、Softmax）都是可微分的。這意味著軟對齊的過程可以與最終的發音評分任務（例如，一個接在 c_i 後的分類器）聯合進行端到端訓練。模型可以在最小化發音分類誤差的過程中，反向傳播梯度並自主地學習到最優的「對齊」策略。相比之下，多數硬對齊演算法（如 Viterbi）是不可微的，只能作為一個獨立的、固定的預處理步驟，阻礙了全局最優化。
4. **減緩誤差傳播 (Mitigation of Error Propagation)**：硬對齊的單點故障風險極高，一旦邊界劃定錯誤，結果便無可挽回。軟對齊則更為穩健。即使模型的注意力焦點（權重峰值）稍有偏移，由於其依然會考慮到周邊幀的資訊，所生成的上下文向量 c_i 仍能保有大部分的有效聲學證據，從而使系統對微小的時序擾動更具韌性。

綜上所述，基於注意力機制的軟對齊框架，以其內在的靈活性、數據驅動的自適應性以及端到端優化的潛力，為發音錯誤檢測任務提供了一個理論上更為強大和魯棒的解決方案。