這篇論文的實驗部分旨在比較基於 Logit 的發音品質分數（GOP）與傳統基於機率的 GOP 分數在錯誤發音檢測中的表現。實驗主要針對兩種 L2 英語語音資料集進行，其中之一就是您提到的 **SpeechOcean762 資料集**。

以下是針對 SpeechOcean762 資料集的實驗解釋，以及是否存在對比基準（baseline）的詳細說明：

一、 SpeechOcean762 資料集概述

SpeechOcean762 是一個用於發音評估的開源語料庫，被選用作為 L2 英語語音資料集之一。

1. **資料來源與規模：** 該資料集包含 5,000 個英語語句，由 250 位以普通話為母語的人士錄製，其中包括 125 位成人和 125 位兒童。
2. **註釋標準：** 每個語句都由五位專家在句子、單詞和音素層面進行註釋。
3. **錯誤發音標註：** 在此資料集中，有 3,401 個音素被標註為錯誤發音。由於 SpeechOcean762 包含人類標註的音素準確度分數，這使得研究可以評估 GOP 分數與人類判斷之間的相關性。

這項研究著重於非母語兒童的語音，因為其 L1 遷移和不一致的音素實現會導致聲學變異性高，對發音評估構成特別的挑戰。

二、 實驗中的對比基準（Baseline）

是的，實驗中存在明確的對比基準（Baseline）。 該研究的核心在於將新提出的基於 Logit 的 GOP 方法與**傳統的、基於 Softmax 後驗機率的方法**進行比較。

1. 核心對比基準：GOPDNN

傳統的 GOP 演算法（在深度神經網路中應用時稱為 **GOPDNN**）構成了主要的對比基準。

- **GOPDNN 的定義：** GOPDNN 是從後驗機率（posterior probabilities）導出的，具體做法是取音素片段上平均 Softmax 輸出的負對數。
- **基準的必要性：** 傳統的 Softmax-based GOP 雖然廣泛採用，但存在著過度自信（overconfidence）和梯度飽和等限制，這些問題降低了其在微妙的錯誤發音檢測中的有效性。
- **研究目的：** 因此，研究的主要問題（RQ）就是：與傳統的 Softmax-based GOP 分數相比，基於 Logit 的 GOP 分數在多大程度上增強了錯誤發音檢測並改善了與人類評分員分數的相關性。

2. 實驗中比較的算法與指標

論文在 SpeechOcean762 資料集上比較了以下幾種 GOP 分數，全部與 GOPDNN 基準進行對比：

方法類型	算法名稱	描述
傳統基準 (Probability-based)	**GOPDNN**	基於平均 Softmax 後驗機率的傳統方法。
基於 Logit 的新方法	**GOPMaxLogit**	捕捉目標音素在對齊幀中的模型峰值信心。
	GOPMargin	量化目標音素相對於最強競爭音素的平均優勢（邊際）。
	GOPVarLogit	測量模型對目標音素預測信心的變異性。
混合方法	**GOPCombined**	結合了 GOPMargin 和 GOPDNN，旨在平衡 Logit 和機率特徵的優勢。

3. 衡量算法優劣的指標

為了了解算法的好壞，研究使用了多種評估指標，並特別利用 SpeechOcean762 的人類註釋進行相關性分析：

- **分類性能指標：** 準確度（Accuracy）、精確度（Precision）、召回率（Recall）、F1 分數和馬修斯相關係數（MCC）。
- **與人類感知相關性指標：** 由於 SpeechOcean762 包含人類註釋的音素準確度評分，研究使用了**皮爾遜相關係數（PCC）** 和 均方誤差（MSE）來量化 GOP 分數對人類評分預測的準確性。

三、 SpeechOcean762 上的基準與結果對比

在 SpeechOcean762 資料集的測試結果（如表 2 所示）中，GOPDNN 作為基準，在某些分類指標上表現強勁，但與人類感知相關性較低：

- **GOPDNN（基準）的表現：** GOPDNN 獲得了最高的準確度（0.947）、精確度（0.333）和 MCC（0.367），顯示其在區分正確和錯誤發音音素方面是有效的。然而，其 PCC 分數（低信心：0.278；高信心：0.295）明顯低於其他基於 Logit 的方法。這表明雖然 GOPDNN 在分類錯誤方面表現良好，但它與人類標註的音素分數一致性較差，對於主觀發音評估而言可靠性較低。
- **GOPMaxLogit（新方法）的表現：** **GOPMaxLogit** 在與人類評分相關性方面表現最好，達到了最高的 PCC 分數（低信心：0.442；高信心：0.456），優於所有其他 GOP 指標

（包括 GOPDNN 基準）。這表明最大 Logit 值能更好地以符合人類感知的方式捕捉發音品質。

總結來說，GOPDNN 基準雖然在分類準確性上表現出色，但 Logit-based 的 GOPMaxLogit 方法在與人類評分員判斷（PCC）的相關性上顯著優於傳統基準。這項對比證實了基於 Logit 的方法（特別是 GOPMaxLogit）在提升發音評估可靠性方面的價值。