Evaluating Logit-Based GOP Scores for Mispronunciation Detection

Aditya Kamlesh Parikh *(aditya.parikh@ru.nl)*, Cristian Tejedor-García, Catia Cucchiarini, Helmer Strik
Center For Language Studies (CLS), Radboud University Nijmegen, The Netherlands

INTRODUCTION

- **Goodness of Pronunciation (GOP)** is a widely used method in mispronunciation detection, but standard **softmax-based GOP** suffers from **overconfidence** and **poor phoneme separation**.
- These limitations reduce GOP's effectiveness, especially for **non-native or child speech**, where **subtle articulatory deviations** are common.
- This study proposes **logit-based GOP** scores as an alternative, preserving richer model information and improving alignment with human perception.
- **RQ:** To what extent does a logit-based GOP score enhance mispronunciation detection and improve correlation with human rater scores compared to traditional softmax-based GOP?

METHODOLOGY

GOP_{DNN} Standard **GOP using softmax probabilities** across aligned phoneme frames.

$GOP_{MaxLogit}$ Takes the **maximum raw logit value** of the target phoneme.

GOP_{Margin} Computes the **average margin** between the target phoneme and the strongest competitor.

$GOP_{VarLogit}$ Measures the **variance of logits** across frames to assess confidence stability.

$GOP_{Combined}$ Weighted **combination of GOP_{DNN} and GOP_{Margin}** and balances both perspectives.

EXPERIMENTAL SETUP

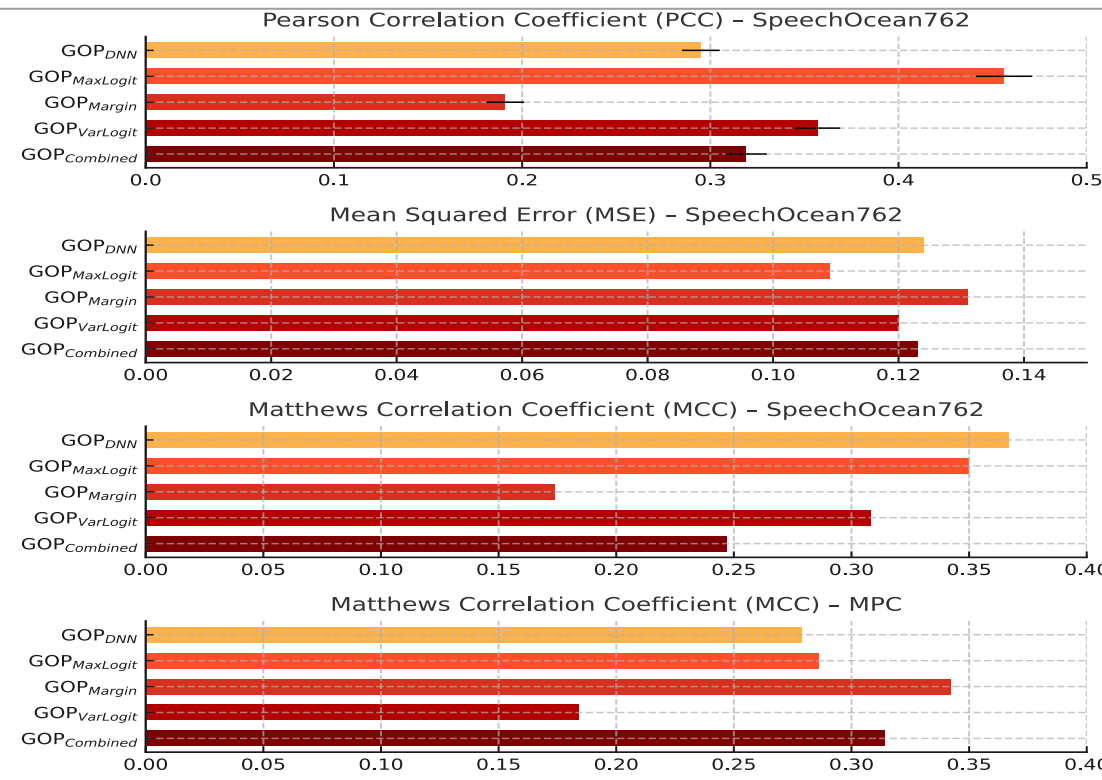
Phoneme Alignment: CTC-segmentation using wav2vec2-xlsr-53-espeak-cv-ft 🤗

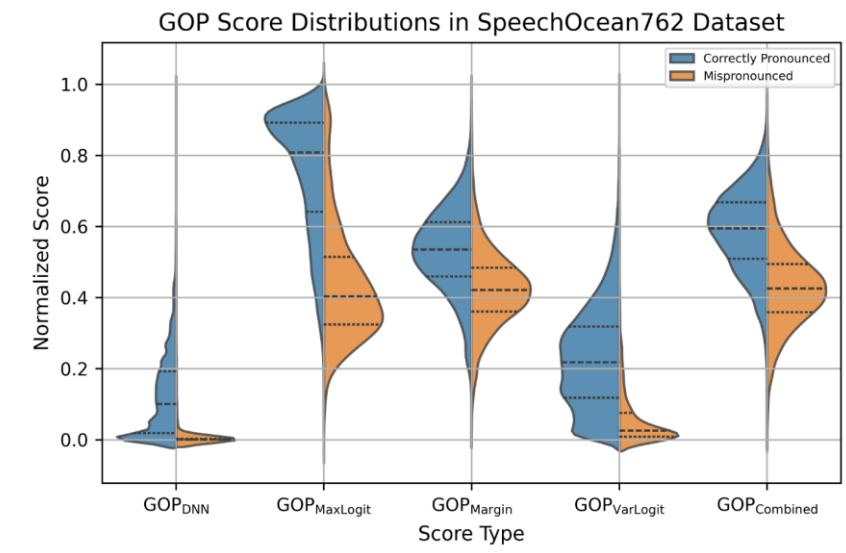
Datasets:

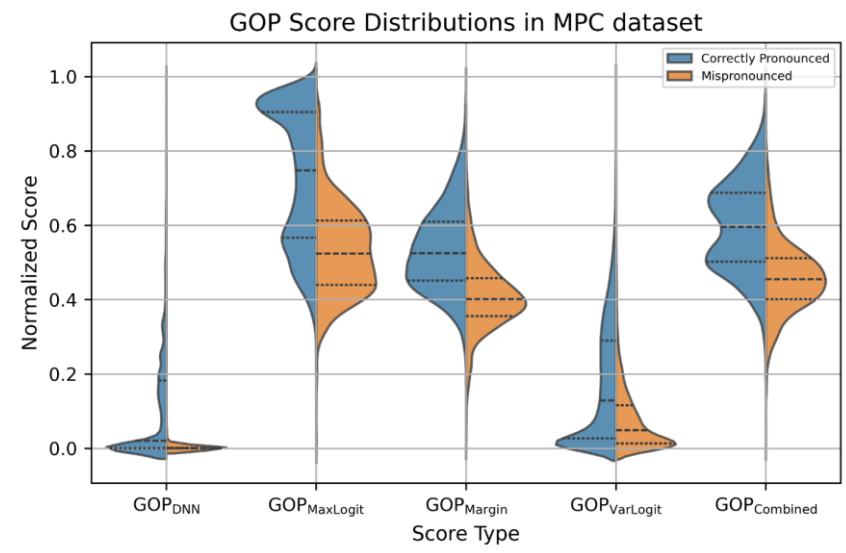
- *My Pronunciation Coach (MPC)* – L1: Dutch, L2: English (Children's speech 👧 with added simulated errors)
- *SpeechOcean 762* – L1: Mandarin, L2: English (Adult + child speech 👧👩‍🏫, human-expert rated)

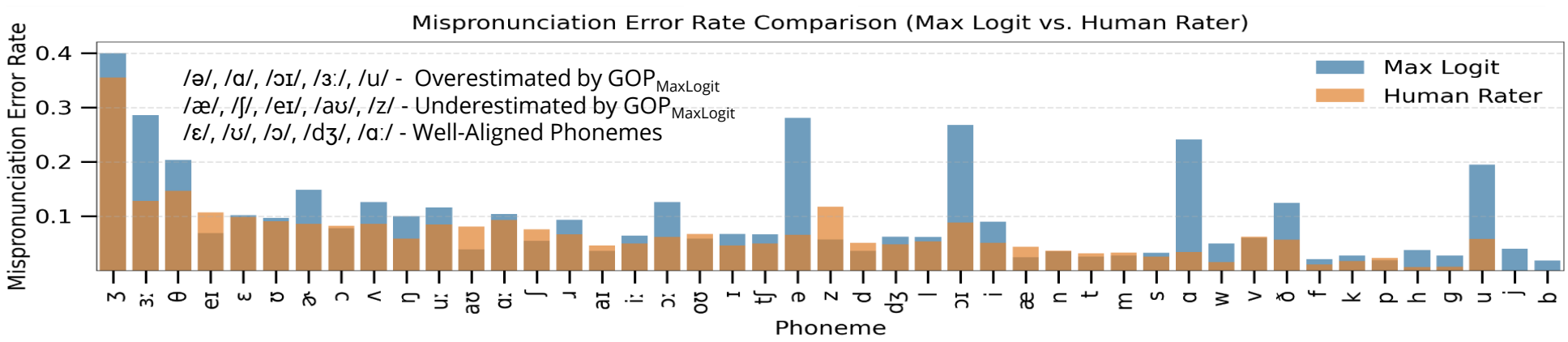
Evaluation Metrics:

- MCC, PCC, MSE; and see the paper for Accuracy, Precision, Recall, F1, AUC.









CONCLUSION AND INSIGHTS

- **$GOP_{MaxLogit}$** closely **aligns with human ratings**, offers **high interpretability**, and **effectively separates correct and mispronounced phonemes**, as shown in distribution plots.
- It achieves the **highest AUCs of 0.736 (MPC)** and **0.754 (SpeechOcean762)**, making it **ideal for perceptual scoring**, but its lower MCC on the MPC may stem from class imbalance or score skew.

Responsible AI for Voice Diagnostics
AiNed Fellowship Grants
NGF.1607.22.013 (RAIVD project)

Radboud Universiteit