Leveraging Allophony in Self-Supervised Speech Models for Atypical Pronunciation Assessment

Kwanghee Choi, Eunjung Yeo, Kalvin Chang, Shinji Watanabe, David Mortensen
Carnegie Mellon University, USA
{kwanghec,swatanab,dmortens}@andrew.cmu.edu

Abstract

Allophony refers to the variation in the phonetic realization of a phoneme based on its phonetic environment. Modeling allophones is crucial for atypical pronunciation assessment, which involves distinguishing atypical from typical pronunciations. However, recent phoneme classifier-based approaches often simplify this by treating various realizations as a single phoneme, bypassing the complexity of modeling allophonic variation. Motivated by the acoustic modeling capabilities of frozen self-supervised speech model (S3M) features, we propose MixGoP, a novel approach that leverages Gaussian mixture models to model phoneme distributions with multiple subclusters. Our experiments show that MixGoP achieves state-of-the-art performance across four out of five datasets, including dysarthric and non-native speech. Our analysis further suggests that S3M features capture allophonic variation more effectively than MFCCs and Mel spectrograms, highlighting the benefits of integrating MixGoP with S3M features.[1]

1 Introduction

A phoneme can be phonetically realized differently depending on its environment, a phenomenon known as *allophony* in phonology (Twaddell, 1952; Ladefoged, 1965; Collins et al., 2019). For instance, the English phoneme /t/ exhibits various allophonic realizations: [tʰ] (aspirated stop) in *tap*, [t] (unaspirated stop) in *stop*, [ɾ] (flap) in *butter*, and [ʔ] (glottal stop) in *kitten*. Accurately capturing these variations is crucial, as it reflects the full spectrum of phonetic realizations within a phoneme. It is particularly important for atypical pronunciation assessment (Twaddell, 1952; Jokisch et al., 2009; Vidal et al., 2019), as it has to distinguish atypical (out-of-distribution; OOD) from atypical (in-distribution) pronunciations (Yeo et al., 2023a).

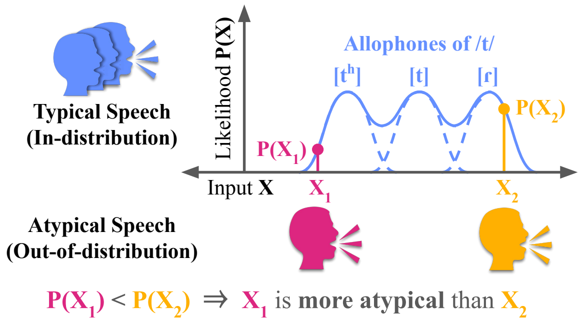


Figure 1: *Summary of our method, MixGoP. We model the likelihood of each phoneme using a Gaussian mixture, trained on typical speech (in-distribution), to capture allophonic variations. We then evaluate on atypical speech (out-of-distribution). The y-axis represents the log-likelihood of a phoneme, where lower values indicate greater atypicality.*

Before the era of deep neural networks (DNNs), allophones were modeled for speech recognition (Sagayama, 1989; Lee et al., 1990; Young et al., 1994). However, DNN-based approaches (Hu et al., 2015b; Yeo et al., 2023a) depend on phoneme classifiers that treat speech segments from a single phoneme as a single cluster, avoiding the complexity of modeling allophones. This is partly due to DNN's strong classification capabilities, which rely on trained hidden features to model individual phonemes well.

In recent years, self-supervised speech models (S3Ms) have shifted the landscape of acoustic modeling. Unlike DNNs, S3Ms leverage their frozen features directly, without requiring additional training (Feng et al., 2023; Chang et al., 2024). Their effectiveness motivates us to revisit modeling allophones via Gaussian Mixture Models (GMMs) (Bilmes et al., 1998; Young et al., 1994). Consequently, we propose MixGoP, a GMM-based approach that models each phoneme as a set of allophonic subclusters (see Figure 1). By integrating GMMs with S3M features, we aim to directly cap-

[1] The full codebase is available at https://github.com/juice500ml/acoustic-units-for-ood

利用音位學在自監督語音模型中進行異常發音評估

黃熙喬，英晶兒，凱文 · 張，信治 · 渡邊，大衛 · 莫滕森
美國卡內基美隆大學
{kwanghec,swatanab,dmortens}@andrew.cmu.edu u

摘要

音位學指的是基於音節的音韻環境，其音韻實現的變化。對音位學的建模對異常發音評估至關重要，這涉及區分異常發音和典型發音。然而，最近的基於音節分類器的方針通常通過將各種實現視為單一音節來簡化這一點，跳過了建模音位學變化的複雜性。受到凍結的自監督語音模型（S3M）特徵的聲學建模能力的啟發，我們提出了一種新方法，稱為 MixGoP，該方法利用高斯混合模型來建模具有多個子簇的音節分佈。我們的實驗表明，MixGoP 在五個數據集中的四個上取得了最優性能，包括失語症和非母語者的語音。我們的分析還表明，S3M 特徵比 MFCC 和梅爾頻譜圖更能有效地捕捉音位學變化，這突出了將 MixGoP 與 S3M 特徵整合的優勢。[1]

1 緒論

一個音素可能會根據其環境而以不同的方式發音，這種現象在語音學中稱為 allophony（Twaddell, 1952; Ladefoged, 1965; Collins et al., 2019）。例如，英語音素 /t/ 會有各種 allophone 實現：[tʰ]（送氣塞音）在 tap 中，[t]（不送氣塞音）在 stop 中，[ɾ]（ap）在 butter 中，以及 [ʔ] （喉塞音）在 kitten 中。準確捕捉這些變化是至關重要的，因為它反映了音素內部所有發音變體的全貌。對於異常發音評估（Twaddell, 1952; Jokisch et al., 2009; Vidal et al., 2019）來說尤其重要，因為它必須區分異常（分佈外；OOD）和異常（分佈內）的發音（Yeo et al., 2023a）。

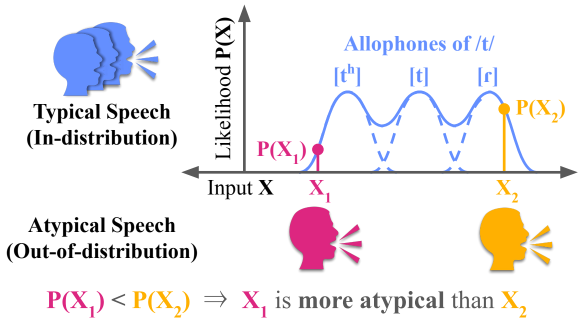


圖 1：我們的方法 MixGoP 的總結。我們使用高斯混合模型對每個音素進行建模，該模型在典型語音（分佈內）上進行訓練，以捕捉 allophone 變化。我們在異常語音（分佈外）上進行評估。y 軸表示音素的對數似然值，其中較低的值表示更爲異常。

在深度神經網絡（DNN）時代之前，allophone 被用於語音識別（Sagayama，1989；Lee et al.，1990；Young etal.，1994）。然而，基於 DNN 的方法（Hu etal.，2015b ；Yeo et al.，2023a）則依賴於將單個音素的語音片段視為單一簇的音素分類器，以避免建模 allophone 的複雜性。這部分是因為 DNN 的強大分類能力，它依賴於訓練後的隱藏特徵來很好地建模單個音素。

近幾年來，自監督語音模型（S3Ms）已改變了聲學建模的格局。與 DNNs 不同，S3Ms 直接利用其凍結的特徵，無需進一步訓練（Feng 等人，2023；Chang 等人，2024）。他們的有效性激發我們重新考慮通過高斯混合模型（GMMs）建模音位變體（Bilmes 等人，1998；Young 等人，1994）。因此，我們提出了一種基於 GMM 的方法 MixGoP，將每個音素建模為一組音位子群集（見圖 1）。通過整合 GMMs 與 S3M 特徵，我們旨在直接 ……

[1] 完整的代碼庫可在以下位置獲取：https://github.com/juice500ml/acoustic-units-for-ood

ture the allophonic variations. We evaluate MixGoP with S3Ms in atypical pronunciation assessment with dysarthric and non-native speech.

Furthermore, we analyze the S3M features on how well they capture allophonic variation compared to Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms. While previous work has shown that S3Ms encode phonetic (Pasad et al., 2021; Wells et al., 2022; Abdullah et al., 2023; Choi et al., 2024b) and phonemic information (Martin et al., 2023; Choi et al., 2024a), a detailed investigation of how they capture allophony remains underexplored.

In summary, the contributions of our study are:

- MixGoP, a novel pronunciation-assessment approach that considers allophonic variation.
- Achieving state-of-the-art performance in four out of five dysarthric and nonnative datasets.
- Analysis of the utility of S3M features on MixGoP for capturing allophonic variations.

2 Method

In this section, we briefly review the conventional approach to pronunciation assessment, Goodness of Pronunciation (GoP) (Witt and Young, 2000). We highlight its limitations: (i) modeling a phoneme as a single cluster, and (ii) assuming atypical speech are in-distribution with respect to typical speech. We then introduce our method, MixGoP, which addresses these limitations by (i) modeling allophonic variation through a mixture distribution and (ii) relaxing in-distribution assumptions by removing the softmax function.

2.1 What is Goodness of Pronunciation?

GoP is a phoneme-level pronunciation score[2] that measures how much the acoustic output of atypical (dysarthric or nonnative) speech deviates from that of typical speech (healthy or native). GoP is measured by how likely a speech segment ($\mathbf{s}$) is to be the intended phoneme (p). Given the phoneme classifier $P_\theta(p|\mathbf{s})$ with trainable parameters θ, GoP is measured as the log phoneme posterior,[3]

$$\mathsf{GoP}_p(s) = \log P_\theta(p|\mathbf{s}). \tag{1}$$

2.2 Limitations of GoP

Conventional phoneme classifiers used in GoP assume a *single cluster* for each phoneme. This is because logits $f_\theta(\mathbf{s})$ are often modeled with a speech encoder Enc and a subsequent fully-connected (FC) layer with weights $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}|\times F}$ (Xu et al., 2021; Yeo et al., 2023a):

$$f_\theta(\mathbf{s}) = \mathbf{W} \cdot \mathsf{Enc}(\mathbf{s}) \tag{2}$$

where $\mathsf{Enc}(\mathbf{s}) \in \mathbb{R}^F$, $|\mathcal{V}|$ denoting the vocabulary size (total number of phonemes), and F the output dimension of the encoder. If we consider a frozen encoder, the trainable parameter $\theta = \{\mathbf{W}\}$. Here, the weights $\mathbf{W}$ can be understood as a codebook, containing a F-dim centroid for each phoneme. It requires a *unimodal* (single peak) clustering of hidden features $\mathsf{Enc}(\mathbf{s})$ for each phoneme. This limits the ability to capture allophonic variation, as allophones are represented as distinct acoustic subclusters within each phoneme.

Another limitation comes from the assumption that observed speech segments are *in-distribution* with respect to the training data. This comes from the phoneme classifier P_θ formulation,

$$P_\theta(p|\mathbf{s}) = \mathtt{softmax}(f_\theta(\mathbf{s}))[p]. \tag{3}$$

With the phoneme classifier relying on the $\mathtt{softmax}$ function, which models a categorical distribution, $\mathbf{s}$ is expected to be within phoneme distribution found in typical speech. However, this assumption is less suitable for atypical speech, which often exhibits substantial acoustic differences from typical speech (Yeo et al., 2023a; Korzekwa et al., 2021).

2.3 MixGoP: Modeling multiple subclusters within a single phoneme

To address the two limitations presented in Section 2.2, we introduce MixGoP, a mixture distribution-based GoP.

First, to overcome the unimodal assumption, MixGoP replaces phoneme classifier $P_\theta(p|\mathbf{s})$ in eq. (1) with a Gaussian mixture model (GMM).

[2] While Witt and Young (2000) uses the term "phone"-level pronunciation scores, we use the term "phoneme" to emphasize that the unit includes allophones. Note that Witt and Young (2000) also suggests that their use of "phone" roughly corresponds to a phoneme.

[3] GoP is traditionally calculated as the average log probability of phoneme p over the corresponding time frames: $\frac{1}{|F|}\sum_{f\in\mathbf{s}} \log P_\theta(p|f)$, where f refers to frames within the utterance $\mathbf{s}$. In our study, we simplify this by considering the GoP as the log probability of the phoneme as a whole, rather than averaging over frames.

本研究的目的是利用音位變體來評估非典型發音。我們使用 Mix-GoP 和 S3Ms 在非流利和母語非母語的發音評估中進行評估。

此外，我們分析 S3M 特徵在捕捉音位變體方面的效果，與梅爾頻率倒頻譜系數（MFCCs）和梅爾頻譜相比。雖然先前的研究已經顯示 S3Ms 編碼聲學（Pasad 等人，2021 ；Wells 等人，2022 ；Abdullah 等人，2023 ；Choi 等人，2024b ）和音韻信息（Martin 等人，2023 ；Choi 等人，2024a ），但對於它們如何捕捉音位變體的詳細研究仍然相對不足。

總結來說，我們研究的貢獻是：

MixGoP，一種考慮音位變體的創新發音評估方法。

在五個非流利和母語非母語數據集中，我們實現了四個的先進性能。

分析 S3M 特徵在 Mix-GoP 中捕捉音位變體的效用。

2 方法

在本節中，我們簡要回顧傳統的發音評估方法，發音品質（GoP）（Witt 和 Young，2000）。我們強調其限制：(i) 將音素建模為單一簇，以及 (ii) 假設非典型語言與典型語言的分布相同。然後我們介紹我們的方法，MixGoP，它通過 (i) 使用混合分佈建模音位變化以及 (ii) 通過移除 softmax 函數來放鬆分布假設來解決這些限制。

2.1 什麼是發音品質？

GoP 是一個音素級別的發音分數，它衡量非典型（運動失調或非母語）語言的聲學輸出與典型（健康或母語）語言的差異程度。GoP 是通過衡量語音片段（$\mathbf{s}$）成為預期音素（p）的可能性來衡量的。給定具有可訓練參數的音素分類器 $P_\theta(p|\mathbf{s})$ θ，GoP 是作為對數音素後驗來衡量的。

$$\mathsf{GoP}_p(s) = \log P_\theta(p|\mathbf{s}). \tag{1}$$

2.2 GoP 的局限性

在 GoP 中使用的傳統音素分類器假設每個音素有一個 單一聚類 。這是因為 logits $f_\theta(\mathbf{s})$ 通常使用語音編碼器 Enc 和後續的完全連接 (FC) 層 $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}|\times F}$ (Xu 等人 , 2021;Yeo 等人 ,2023a):

$$f_\theta(\mathbf{s}) = \mathbf{W} \cdot \mathsf{Enc}(\mathbf{s}) \tag{2}$$

其中 $\mathrm{Enc}(\mathbf{s}) \in \mathbb{R}^F$, $|\mathcal{V}|$ 表示詞典大小（總音素數量），而 F 表示編碼器的輸出維度。如果我們考慮一個凍結的編碼器，可訓練參數 $\theta = \{\mathbf{W}\}$。這裡的權重 $\mathbf{W}$ 可以理解為一個碼本，包含每個音素的 F- 維中心點。這需要對每個音素進行單峰 （單一峰值）隱藏特征的聚類。這限制了捕捉音位變化的能力，因為音位變體被表示為每個音素內的獨立聲學子聚類。

另一個局限性來自假設觀察到的語音片段與訓練數據在分布上。這來自音素分類器的 P_θ 公式。

$$P_\theta(p|\mathbf{s}) = \mathtt{softmax}(f_\theta(\mathbf{s}))[p]. \tag{3}$$

音素分類器依賴於 $\mathtt{softmax}$ 函數，該函數模擬一個分類分布，$\mathbf{s}$ 預期將在典型語音中找到的音素分佈內。然而，這個假設對於不典型語音來說不太適合，因為不典型語音通常與典型語音有顯著的聲學差異（Yeo 等人 , 2023a; Korzekwa 等人 ,2021）。

2.3 MixGoP：在單一音素內建模多個子聚類

為了解決第二章 2.2 節提出的兩個限制，我們提出了一種基於混合分佈的 GoP，稱為 MixGoP。

首先，為了克服單峰假設，MixGoP 將等式（1）中的音素分類器 P_θ （$p|\mathbf{s}$）用高斯混合模型（GMM）取代。

[2] 雖然 Witt 和 Young (2000) 使用「phone」級別的發音分數，我們使用「phoneme」來強調該單位包括音位。注意，Witt 和 Young (2000) 也建議他們對「phone」的使用大約對應於音素。

GoP 傳統上計算為音素 p 在相應時間框架上的平均對數概率：$\frac{1}{|F|}\sum_{f\in\mathbf{s}} \log P_\theta(p|f)$，其中 f 指的是語句 $\mathbf{s}$ 中的框架。在我們的研究中，我們將其簡化為考慮音素的對數概率作為整體，而不是對框架進行平均。

GMM is a weighted sum of Gaussian distributions that can directly model the phoneme likelihood $P_\theta(\mathbf{s}|p)$ (distribution of speech segment $\mathbf{s}$ for each individual phoneme p). Accordingly, we formulate the phoneme likelihood as follows:

$$P_\theta(\mathbf{s}|p) = \sum_{c=1}^{C} \pi_p^c \mathcal{N}(\mathsf{Enc}(\mathbf{s})|\boldsymbol{\mu}_p^c, \boldsymbol{\Sigma}_p^c) \qquad (4)$$

where $\mathcal{N}$ denotes the multivariate Gaussian distribution, $\boldsymbol{\mu}_p^c \in \mathbb{R}^F$ and $\boldsymbol{\Sigma}_p^c \in \mathbb{R}^{F\times F}$ is the mean vector (centroid) and covariance matrix, and $\pi_p^c \in [0,1]$ is the mixing coefficient. Here, the trainable parameter $\theta = \{\boldsymbol{\mu}_p^c, \boldsymbol{\Sigma}_p^c, \pi_p^c\}_{c\in[C],p\in\mathcal{V}}$. Then, we can newly define our MixGoP score as:

$$\mathtt{MixGoP}_p(s) = \log P_\theta(\mathbf{s}|p). \qquad (5)$$

Our MixGoP score differs from the original GoP score in eq. (1) by replacing the phoneme posterior $P_\theta(p|\mathbf{s})$ with the phoneme likelihood $P_\theta(\mathbf{s}|p)$. By doing so, we are also removing the influence of phoneme prior $P(p)$, which is known to be effective in practice (Yeo et al., 2023a).

Second, MixGoP removes the `softmax` function of eq. (3) by directly using the log-likelihood in eqs. (4) and (5). It relaxes the assumption of phonemes in atypical speech being in-distribution. The quadratic term inside each Gaussian:

$$-\frac{1}{2}(\mathsf{Enc}(\mathbf{s}) - \boldsymbol{\mu}_p^c)^T(\Sigma_p^c)^{-1}(\mathsf{Enc}(\mathbf{s}) - \boldsymbol{\mu}_p^c) \qquad (6)$$

directly relates to the Mahalanobis distance, which is commonly used for OOD detection (Lee et al., 2018). By avoiding the `softmax`, MixGoP is likely to be more robust in handling OOD speech.

In summary, we train a total of $|\mathcal{V}|$ GMMs (one for each phoneme) where each GMM is composed of C subclusters, *e.g.*, $C = 32$. C is kept constant across all phonemes, as it is known that sufficiently large number of Gaussian mixtures can approximate any probability density (Nguyen et al., 2020). Experiments on the influence of C on downstream performance can be found in Appendix C.2. We use the k-means algorithm to determine the initial cluster centers and the expectation-maximization (EM) algorithm to optimize the parameters of the Gaussian mixtures, using `scikit-learn 1.4.1` (Pedregosa et al., 2011). By considering allophony in modeling, MixGoP is expected to better reflect the distribution of each phoneme.

3 Experiments

3.1 Datasets

We use five datasets: three dysarthric speech datasets (UASpeech (Kim et al., 2008), TORGO (Rudzicz et al., 2012), and SSNCE (TA et al., 2016)) and two non-native speech datasets (speechocean762 (Zhang et al., 2021) and L2-ARCTIC (Zhao et al., 2018)). In this paper, we use healthy or native speech as the training sets, and dysarthric and non-native speech as the test sets, in line with the OOD literature (Hendrycks and Gimpel, 2017). Refer to Appendix A for more details.

3.2 Feature extraction

For our experiments, we compare various speech feature extractors $\mathsf{Enc}(\mathbf{s})$ (eqs. (2) and (4)).

Traditional acoustic features. We use the Mel-Frequency Cepstral Coefficients (MFCCs) and Mel spectrograms as baselines, using the default hyperparameters of librosa (McFee et al., 2015).

TDNN-F features. We compare with a factorized time-delay neural network (TDNN-F) model (Povey et al., 2018) for the speechocean762 dataset, as TDNN-F features have been often used as baselines (Zhang et al., 2021; Gong et al., 2022; Chao et al., 2022; Do et al., 2023).

S3M features. We employ two frozen S3Ms: XLS-R-300M (Babu et al., 2022) and WavLM-Large (Chen et al., 2022). XLS-R (shorthand for XLS-R-300M), trained cross-lingually, has demonstrated strong performance in ASR for low-resource languages (Babu et al., 2022) and dysarthric speech assessment (Yeo et al., 2023a). We also employ WavLM (shorthand for WavLM-Large), a state-of-the-art model for various tasks, including phoneme recognition (Feng et al., 2023; Yang et al., 2021).

As different layers of S3Ms are known to encode different information (Pasad et al., 2021, 2023), we use features from each layer. Specifically, we extract convolutional features (denoted as layer index 0) and all consecutive Transformer features (denoted as layer indices 1 through 24).

Feature segmentation. We segment the features according to the start and end timestamps of each phoneme. Refer to the detailed time-alignment process in Appendix A. Then, we apply center

GMM 是一個加權的高斯分佈總和，可以直接模擬音素似然 $P_\theta(\mathbf{s}|p)$ （語音片段 s 的分布，對於每個個體音素 p ）。因此，我們將音素似然表達如下：

$$P_\theta(\mathbf{s}|p) = \sum_{c=1}^{C} \pi_p^c \mathcal{N}(\mathsf{Enc}(\mathbf{s})|\boldsymbol{\mu}_p^c, \boldsymbol{\Sigma}_p^c) \qquad (4)$$

其中 $\mathcal{N}$ 表示多維高斯分佈，$\boldsymbol{\mu}_p^c \in \mathbb{R}^F$ 和 $\boldsymbol{\Sigma}_p^c \in \mathbb{R}^{F\times F}$ 是均值向量（中心）和協方差矩陣，而 $\pi_p^c \in [0,1]$ 是混合係數。在此，可訓練的參數 $\theta = \{\boldsymbol{\mu}_p^c, \boldsymbol{\Sigma}_p^c, \pi_p^c\}_{c\in[C]}$，$p\in\mathcal{V}$。然後，我們可以重新定義我們的 MixGoP 得分如下：

$$\mathtt{MixGoP}_p(s) = \log P_\theta(\mathbf{s}|p). \qquad (5)$$

我們的 MixGoP 得分與等式（1）中的原始 GoP 得分不同，我們將音素後驗概率 P_θ （p|s）用音素似然 P_θ （s|p）來替換。這樣做同時也去除了音素先驗概率 P （p）的影響，這在實踐中已被證明是有效的（Yeo 等人，2023a）。

其次，MixGoP 通過直接使用等式（4）和（5）中的對數似然來移除等式（3）中的 softmax 函數。它放鬆了異常語音中音素分佈的假設。每個高斯內部的二次項：

$$-\frac{1}{2}(\mathsf{Enc}(\mathbf{s}) - \boldsymbol{\mu}_p^c)^T(\Sigma_p^c)^{-1}(\mathsf{Enc}(\mathbf{s}) - \boldsymbol{\mu}_p^c) \qquad (6)$$

直接相關於馬哈拉諾比斯距離，這是常用于 OOD 檢測的（Lee 等人，2018）。通過避免 softmax，MixGoP 可能會在處理 OOD 語音時更具魯棒性。

總結來說，我們共訓練了 $|\mathcal{V}|$ 個 GMM（每個音素一個），其中每個 GMM 由 C 個子簇組成，例如，$C = 32$。C 在所有音素中保持不變，因為已知足夠大的高斯混合可以逼近任何機率密度（Nguyen 等人，2020）。關於 C 對下游性能影響的實驗可以在附錄 C2 中找到。我們使用 k-means 演算法確定初始簇中心，並使用期望最大化（EM）演算法優化高斯混合的參數，使用 scikit-learn 14.1（Pedregosa 等人，2011）。考慮到在建模中利用音位，MixGoP 預期將更好地反映每個音素的分佈。

3 实驗

3.1 資料集

我們使用五個資料集：三個失語症語音資料集（UASpeech（金等，2008）、TORGO（魯茲奇克等，2012）和 SSNCE（TA 等，2016））以及兩個非母語語音資料集（speechocean762（張等，2021）和 L2-ARCTIC（趙等，2018））。在本文中，我們使用健康或母語語音作為訓練集，失語症和非母語語音作為測試集，與 OOD 文獻（亨德里斯克和吉姆佩爾，2017）相一致。詳情請參考附錄 A。

3.2 特徵提取

為了我們的實驗，我們比較了各種語音特徵提取器 Enc(s)（等式 (2) 和 (4)).

傳統聲學特徵。我們使用梅爾頻率倒頻譜係數（MFCCs）和梅爾頻譜圖作為基準，使用 librosa 的默認超參數（麥菲等，2015）。

TDNN-F 特徵。我們將與一個因數化的時間延遲神經網路 (TDNN-F) 模型進行比較（Povey 等人，2018），該模型在 speechocean762 資料集上已被經常用作基準（張等人，2021；龔等人，2022；趙等人，2022；杜等人，2023）。

S3M 特徵。我們使用兩個凍結的 S3M：XLS-R-300M（Babu 等人，2022）和 WavLM-Large（Chen 等人，2022）。XLS-R（XLS-R-300M 的簡稱），跨語言訓練，在低資源語言的 ASR 和失語症語音評估中表現出色（Babu 等人，2022；Yeo 等人，2023a）。我們還使用 WavLM（WavLM-Large 的簡稱），這是一個在多種任務中表現最優的模型，包括音素識別（Feng 等人，2023；楊等人，2021；陳等人，2021）。

由於 S3M 的不同層次已知可以編碼不同的信息（Pasad 等人，2021，2023），我們使用每個層次的特徵。具體來說，我們提取卷積特徵（表示為層次索引 0）和所有連續的 Transformer 特徵（表示為層次索引 1 通過 24）。

特徵分割。我們根據每個音素的起始和結束時間戳對特徵進行分割。參考附錄 A 的詳細時間對齊過程。然後，我們應用中心

pooling to extract one feature per segment.[4]

3.3 Baselines

We verify the effectiveness of our MixGoP by comparing it against various baselines (Yeo et al., 2023a; Sun et al., 2022; Shahin et al., 2024; Schölkopf et al., 2001). We evaluate on all the speech features listed in Section 3.2 across all the methods for fair comparison. These baselines are categorized into two groups: (i) phoneme classifier-based and (ii) OOD detector-based approaches.

Phoneme classifier-based approaches encompass conventional GoP formulations, which assume a unimodal distribution and in-distribution of phonemes, as discussed in Section 2.2. We employ four popular GoP formulations[5]: GMM-GoP (Witt and Young, 2000), NN-GoP (Hu et al., 2015b), DNN-GoP (Hu et al., 2015b), and MaxLogit-GoP (Yeo et al., 2023a). Note that all formulations use the same underlying phoneme classifier $P_\theta(p|\mathbf{s})$. They only differ by how to calculate the GoP scores. Refer to Yeo et al. (2023a) for more details.

OOD detector-based approaches calculate GoP by measuring how likely an input is to be an outlier. In other words, they can quantify the level of atypicalness. Our MixGoP is one of these approaches, as MixGoP models the likelihood $P_\theta(\mathbf{s}|p)$ with typical speech (eq. (4)) and identifies outliers (atypical speech) based on their likelihood (eq. (5)). We additionally test three baselines: k-nearest neighbors (kNN) (Sun et al., 2022), one-class support vector machine (oSVM) (Schölkopf et al., 2001), and phoneme-specific oSVM (p-oSVM) (Shahin and Ahmed, 2019). While kNN has been utilized for OOD detection (Sun et al., 2022), it has not previously been applied to dysarthric or non-native speech. Conversely, oSVM and p-oSVM have been applied to the evaluation of both disordered and non-native speech (Shahin and Ahmed, 2019; Shahin et al., 2024).

3.4 Training details

Phoneme classifier-based. The phoneme classifier in eq. (1) is trained on features from Section 3.2 with a single learnable FC layer (eq. (2)) with the default settings of Adam optimizer (Kingma and Ba, 2015) for a maximum of 500 iterations.

OOD detector-based. For kNN, we construct a kNN model for each phoneme using the features from Section 3.2. Then, we use the maximum Euclidean distance between the test data feature and the nearest 10% training data feature as the GoP score, following Sun et al. (2022). For oSVM, all phonemes is modeled with a single oSVM model (Shahin et al., 2024), while p-oSVM modeled each phoneme as a separate oSVM model. All oSVM models are trained with features using the default hyperparameters of `scikit-learn 1.4.1` (Pedregosa et al., 2011). Radial basis function was used for both oSVM and p-oSVM. We use the distance from the hyperplane as the GoP score.

MixGoP. In our MixGoP framework, we apply random subsampling of 512 features per phoneme. Empirical analysis indicates that subsampling does not necessarily degrade performance (See Section 5.2). The number of subclusters for each phoneme-wise Gaussian mixture is set to 32. A detailed investigation into the effect of the number of subclusters on GoP performance is discussed in Appendix C.2.

3.5 Evaluation

As described in Section 3.2, we segment the spoken utterance $\mathbf{x}$ phoneme-wise: $\mathbf{x} = \{(p_1, \mathbf{s_1}), (p_2, \mathbf{s_2}), \cdots, (p_N, \mathbf{s_N})\}$, where p_i is the phoneme label, $\mathbf{s}_i$ is the observed speech segment, and N is the total number of phonemes within the utterance. Following Yeo et al. (2023a), we define the pronunciation score of an utterance $\mathbf{x}$ as:

$$\mathtt{Pronunciation}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\mathtt{GoP}_p(s), \quad (7)$$

where the definition of the GoP is different per each method. That is, the GoP scores are averaged across the utterance.

Similar to Yeo et al. (2023a), we evaluate performance using the Kendall-tau correlation coefficient between the utterance-level pronunciation scores and the ground truth dysfluency/disfluency scores provided by the dataset. While Yeo et al. (2023a) used additional training data, our setting only uses the aformentioned datasets.

Unlike other datasets, L2-ARCTIC contains only phoneme-wise mispronunciation detection labels

[4] We chose center pooling over average pooling to reduce the impact of inaccurate phoneme alignments on GoP calculations. Both pooling methods are known to encode similar amounts of phonetic information for S3Ms (Pasad et al., 2024; Choi et al., 2024b).

[5] GoP formulations were named after the underlying models (Zhang et al., 2021; Yeo et al., 2023a), leading to names like GMM-GoP. However, the GMM-GoP scoring method (eq. (1)) does not necessarily rely on GMM.

透過池化來提取每個段落的特徵。[4]

3.3 基準

我們通過與各種基準進行比較來驗證我們的 MixGoP 的有效性（ Yeo 等人，2023a ；Sun 等人，2022 ；Shahin 等人，2024 ；Schölkopf 等人，2001）。我們在所有方法上評估了第 3.2 节中列出的所有語音特徵，以進行公平比較。這些基準被分為兩組：（ i ）基於音素分類器的和（ ii ）基於 OOD 检測器的方法。

基於音素分類器的方針**涵蓋了傳統的 GoP 公式，這些公式假設音素的單峰分佈和分佈內的音素，如第 2.2 节所討論的。我們採用了四種流行的 GoP 公式**：GMM-GoP （ Witt 和 Young，[5]2000 ），NN-GoP （ Hu 等人，2015b ），DNN-GoP （ Hu 等人，2015b ），和 MaxLogit-GoP （ Yeo 等人，2023a ）。注意，所有公式都使用相同的基礎音素分類器（ sP_θ ） $p|$。**它們只差在計算 GoP 分數的方式上。參考** Yeo 等人 （ 2023a ）以獲得更多細節。

基於 OOD 检測器的方針**通過測量輸入成為異常值的可能性來計算 GoP。換句話說，它們可以量化不典型性的程度。我們的 MixGoP 是這些方法之一，因為 MixGoP 模擬了典型語音（等式**（ 4P_θ ））的似然性，並根據其似然性（等式（ 5 ））識別異常值（不典型語音）。此外，我們還測試了三個基準：k-nearest neighbors （ kNN ）（ $|p$Sun 等人，2022 ），one-class support vector machine （ oSVM ）（ Schölkopf 等人，2001 ），和 phoneme-specific oSVM （ p-oSVM ）（ Shahin 和 Ahmed，2019 ）。雖然 kNN 已經被用於 OOD 檢測（ Sun 等人，2022 ），但它以前沒有被應用於失語症或非母語語音。相反，oSVM 和 p-oSVM 已經被用於評估失調和非母語語音（ Shahin 和 Ahmed，2019 ；Shahin 等人，2024 ）。

3.4 訓練細節

基於音素分類器的。 等式（ 1）中的音素分類器是基於第 3.2 節的特徵進行訓練的。使用一個單一的學習性全連接層（等式（ 2 ））並使用 Adam 優化器的默認設置（ Kingma andBa,2015 ）進行最多 500 次迭代。

基於 OOD 檢測器的。 對於 kNN，我們使用第 3.2 節的特徵為每個音素構建一個 kNN 模型。然後，我們使用測試數據特徵與最近的 10% 訓練數據特徵之間的最大歐氏距離作為 GoP 得分，參考 Sun et al. （ 2022 ）。對於 oSVM，所有音素都使用一個單一的 oSVM 模型（ Shahinet al., 2024 ）進行建模，而 p-oSVM 則將每個音素建模為一個分離的 oSVM 模型。所有 oSVM 模型都使用默認超參數使用 scikit-learn 1.4.1 的特徵進行訓練（ scikit-learn 1.4.1 （ Pedregosaetal., 2011 ））。對於 oSVM 和 p-oSVM 都使用徑向基函數。我們使用超平面的距離作為 GoP 得分。

在我們的 MixGoP 框架中，我們對每個音素應用隨機子抽樣，每個音素 512 個特徵。經驗分析表明，子抽樣不會必然降低性能（見第 5. 2 節）。每個音素向量的高斯混合子群數設為 32。對於子群數對 GoP 性能的影響進行了詳細的討論，詳見附錄 C.2。

3.5 評估

如第 3.2 節所述，我們將語音輸入 x 按音素進行分節：x$(p_1$, s1), $(p_2$, s2), $\cdots$, $(p_N$, $\mathbf{s_N})\}$，其中 p_i 是音素標籤，s$_i$ 是觀察到的語音片段，N 是輸入中的音素總數。遵循 Yeo 等人（ 2023a ）的方法，我們定義語音輸入 x 的發音分數為：

$$\mathtt{Pronunciation}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\mathtt{GoP}_p(s), \quad (7)$$

其中 GoP 的定義對每種方法都不同。即，GoP 分數是對整個輸入進行平均的。

與 Yeo 等人（ 2023a ）相似，我們使用語音輸入級發音分數與數據集提供的地面實際失調 / 非失調分數之間的 Kendall-tau 相關係數來評估性能。雖然 Yeo 等人（ 2023a ）使用了額外的訓練數據，但我們的設定僅使用上述數據集。

與其他數據集不同，L2-ARCTIC 只包含音素級誤發音檢測標籤。

我們選擇中心池化而非平均池化，以減少不準確的音素對齊對 GoP 計算的影響。這兩種池化方法都已知對 S3Ms （ Pasad etal., 2024;Choi et al., 2024b ）編碼相似的語音信息量。

GoP 公式化名稱來自於其基礎模型（ Zhang et al., 2021; Yeo et al., 2023a ），導致名稱如 GMM-GoP。然而，GMM-GoP 評分方法（等式（ 1 ））並不一定依賴於 GMM。

Table 1: *Kendall-tau correlation coefficient between the pronunciation scores and the dysfluency/disfluency (absolute value). Bigger is better. For S3Ms, the best performance across layers is displayed.*

	Dataset	Feature	Phoneme classifier-based				Out-of-distribution detector-based			
			GMM-GoP	NN-GoP	DNN-GoP	MaxLogit-GoP	kNN	oSVM	p-oSVM	MixGoP (Proposed)
Dysarthric speech	UASpeech	MFCC	0.428	0.410	0.361	0.430	0.418	0.107	0.105	0.182
		MelSpec	0.209	0.172	0.242	0.214	0.101	0.099	0.086	0.039
		XLS-R	0.552	0.553	0.548	0.547	0.559	0.354	0.247	0.602
		WavLM	0.568	0.568	0.546	0.558	0.606	0.537	0.327	**0.623**
	TORGO	MFCC	0.406	0.345	0.391	0.406	0.347	0.169	0.105	0.282
		MelSpec	0.271	0.262	0.150	0.271	0.287	0.211	0.241	0.196
		XLS-R	0.677	0.674	0.641	0.675	0.704	0.586	0.536	**0.713**
		WavLM	0.682	0.681	0.633	0.681	0.703	0.671	0.621	0.707
	SSNCE	MFCC	0.265	0.254	0.267	0.273	0.045	0.194	0.076	0.082
		MelSpec	0.183	0.161	0.051	0.187	0.154	0.114	0.106	0.174
		XLS-R	0.542	0.542	0.499	0.544	0.503	0.193	0.167	0.541
		WavLM	0.547	0.547	0.486	0.547	0.523	0.358	0.234	**0.553**
Non-native speech	speechocean762	MFCC	0.390	0.375	0.255	0.405	0.322	0.202	0.111	0.126
		MelSpec	0.214	0.064	0.232	0.229	0.111	0.109	0.071	0.004
		TDNN-F	0.400	0.356	0.243	0.360	0.361	0.099	0.001	0.197
		XLS-R	0.533	0.531	0.372	0.536	0.443	0.312	0.157	0.499
		WavLM	0.535	0.533	0.380	0.534	0.432	0.395	0.173	**0.539**
	L2-ARCTIC	MFCC	0.136	0.141	0.119	0.119	0.042	0.004	0.034	0.043
		MelSpec	0.049	0.039	0.032	0.032	0.022	0.003	0.027	0.010
		XLS-R	0.243	**0.312**	0.191	0.191	0.168	0.037	0.067	0.152
		WavLM	0.240	0.269	0.196	0.196	0.189	0.082	0.078	0.182

(0 for correct, and 1 for mispronounced). Therefore, we directly measure the correlation between the predicted phoneme-wise pronunciation scores and the mispronunciation detection labels.

3.6 Results

Table 1 presents the experimental results, where we report the best-performing layer's results for S3Ms. Refer to Appendix C.1 for the performance of individual layers. Table 1 shows that our proposed MixGoP method achieved the state-of-the-art performance across all the datasets except L2-ARCTIC. As for the L2-ARCTIC dataset, the best performance was observed with NN-GoP.

Comparison between features. We observe that all other features generally underperform compared to both S3Ms, further highlighting the general effectiveness of frozen S3Ms (Yang et al., 2021). It aligns with Choi and Yeo (2022), where S3Ms, unlike MFCCs, stores information as a relative distance between the features so that the Mahalanobis distance of MixGoP (eq. (6)) or the Euclidean distance of kNN (Appendix C.4) can be effective. We explore this discussion in detail in Section 4.

Comparison between datasets. We observe that MixGoP tends to be effective on dysarthric datasets, whereas NN-GoP tends to perform well on non-native datasets. We suspect this is due to dysarthric test sets being more OOD than non-native test sets, so accurate likelihood estimation becomes more important. Dysarthric train/test datasets are split by healthy and dysarthric speakers. However, non-native datasets only contain non-native speakers, where we split train/test using utterance-wise or phoneme-wise pronunciation scores. (Details in Appendix A.) Hence, it is likely that the train/test difference of non-native datasets is less severe than that of dysarthric datasets. Inherent acoustic differences between dysarthric and non-native speech may have further widen the dataset differences (Yeo et al., 2023a; Korzekwa et al., 2021).

Comparison within the same groups. For *phoneme classification-based* baselines, performance greatly differs across methods, despite all four methods being based on the same classifier. This supports the importance of selecting an appropriate equation for uncertainty quantification in phoneme class-based methods, aligning with

表 1：發音分數與口吃 / 流利度（絕對值）之肯德尔 tau 相關係數。越大越好。對於 S3Ms，顯示最佳層次性能。

	資料集	功能	基於音素分類器				基於分佈外檢測器			
			GMM-GoP	NN-GoP	DNN-GoP	MaxLogit-GoP	kNN	oSVM	p-oSVM	MixGoP（提案）
Dysarthric speech	UASpeech	MFCC	0.428	0.410	0.361	0.430	0.418	0.107	0.105	0.182
		梅頻譜	0.209	0.172	0.242	0.214	0.101	0.099	0.086	0.039
		XLS-R	0.552	0.553	0.548	0.547	0.559	0.354	0.247	0.602
		WavLM	0.568	0.568	0.546	0.558	0.606	0.537	0.327	**0.623**
	TORGO	MFCC	0.406	0.345	0.391	0.406	0.347	0.169	0.105	0.282
		梅規範	0.271	0.262	0.150	0.271	0.287	0.211	0.241	0.196
		XLS-R	0.677	0.674	0.641	0.675	0.704	0.586	0.536	**0.713**
		WavLM	0.682	0.681	0.633	0.681	0.703	0.671	0.621	0.707
	SSNCE	MFCC	0.265	0.254	0.267	0.273	0.045	0.194	0.076	0.082
		梅頻譜	0.183	0.161	0.051	0.187	0.154	0.114	0.106	0.174
		XLS-R	0.542	0.542	0.499	0.544	0.503	0.193	0.167	0.541
		WavLM	0.547	0.547	0.486	0.547	0.523	0.358	0.234	**0.553**
Non-native speech	speechocean762	MFCC	0.390	0.375	0.255	0.405	0.322	0.202	0.111	0.126
		MelSpec	0.214	0.064	0.232	0.229	0.111	0.109	0.071	0.004
		TDNN-F	0.400	0.356	0.243	0.360	0.361	0.099	0.001	0.197
		XLS-R	0.533	0.531	0.372	0.536	0.443	0.312	0.157	0.499
		WavLM	0.535	0.533	0.380	0.534	0.432	0.395	0.173	**0.539**
	L2-ARCTIC	MFCC	0.136	0.141	0.119	0.119	0.042	0.004	0.034	0.043
		梅規範	0.049	0.039	0.032	0.032	0.022	0.003	0.027	0.010
		XLS-R	0.243	**0.312**	0.191	0.191	0.168	0.037	0.067	0.152
		WavLM	0.240	0.269	0.196	0.196	0.189	0.082	0.078	0.182

(0 代表正確，1 代表發音錯誤)。因此，我們直接測量預測的音素級發音分數與發音錯誤檢測標籤之間的相關性。

3.6 結果

我們在表中呈現實驗結果，其中我們報告了最佳表現層次的 S3Ms 結果。請參考附錄 C.1以了解單一層次的性能。表 1顯示，我們提出的 MixGoP 方法在所有數據集上除了 L2-ARCTIC 外都取得了最優性能。至於 L2-ARCTIC 數據集，最佳性能是 NN-GoP。

特徵之間的比較。我們觀察到所有其他特徵與 S3Ms 相比普遍表現不佳，進一步突顯了凍結 S3Ms （楊等，2021）的普遍有效性。這與 Choi 和 Yeo （ 2022 ）的研究結果一致，其中 S3Ms 與 MFCCs 不同，將信息作為特徵之間的相對距離存儲，因此 MixGoP （等式（ 6 ））的馬哈拉諾比斯距離或 kNN （附錄 C4 ）的歐氏距離可以有效地使用。我們在第 4 節中詳細探討了這個問題。

數據集之間的比較。我們觀察到 MixGoP 在失語症數據集上 tends to be effective，而 NN-GoP 在非母語數據集上 tends to perform well。我們懷疑這是因為失語症測試集比非母語測試集更為 OOD，因此精準的似然估計變得更加重要。失語症訓練 / 測試數據集按健康人和失語症發言人分開。然而，非母語數據集只包含非母語發言人，我們使用語句或音素發音分數進行訓練 / 測試分開。（詳情見附錄 A。）因此，非母語數據集的訓練 / 測試差異可能不如失語症數據集嚴重。失語症和非母語語音之間的內在聲學差異可能進一步擴大了數據集差異（ Yeo 等，2023a ； Korzekwa 等，2021）。

同一組內的比較。對於基於音素分類的基線，儘管所有四種方法都是基於相同的分類器，但性能在各方法之間差異很大。這支持了在音素分類方法中選擇合適的不確定性量化公式的重性，與 ……

the findings of Yeo et al. (2023a). Regarding *OOD detector-based* baselines, SVM-based methods usually perform the worst across all datasets. In contrast, kNN achieves performance comparable to our proposed MixGoP on the TORGO and L2-ARCTIC datasets, while MixGoP delivers the best overall performance.

4 Allophony of S3M features

Section 3 empirically demonstrates that leveraging S3M features with MixGoP helps enhance downstream performance compared to other features, such as MFCCs and Mel spectrograms. This section aims to further verify the suitability of S3Ms for representing individual phonemes with allophonic variations. First, we examine S3M features at the phoneme-level, using the dimensionality reduction technique in Section 4.1. Next, we design a metric to quantify the ability of capturing allophonic variations, comparing S3M features to MFCCs and Mel spectrogram in Section 4.2. For our analyses, we used the healthy speech recordings from the TORGO dataset (Rudzicz et al., 2012), which includes gold-standard phonemic transcriptions and alignments.

4.1 Motivating Observation

S3Ms are trained to reconstruct masked signals using surrounding information. Hence, we hypothesize that this will allow the S3Ms to capture local acoustic characteristics, including allophones from various phonetic environments. To verify such phenomena, we observed the final layer features of WavLM for each phoneme, which have generally shown the best performance across datasets (See Figure 5). Specifically, we use UMAP dimensionality reduction (McInnes et al., 2018) with the cosine distance metric to visualize the features, similar to Choi and Yeo (2022). We also extract the four utterances closest to each of the ten centroids to observe the phonetic environments of each cluster.

Figure 2 demonstrates one example, with the distribution of /ʌ/ (/AH/ in ARPABET) and its environments of the healthy subset of TORGO. We observed multiple clusters for each phoneme, each with phonetically similar environments, which motivates the metric for quantifying allophony ability.

4.2 Quantifying Allophony

Previous studies have found that S3M features model the phoneme distributions with multiple

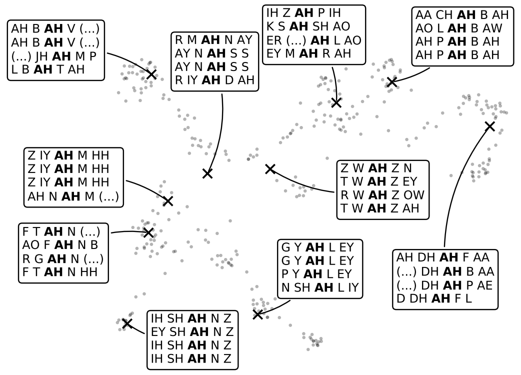


Figure 2: *Visualization of WavLM-Large features of the* /AH/ *phoneme in the TORGO healthy subset. Phonemes are indicated using ARPABET. We observe that* /AH/ *consists of subclusters, each reflecting allophones from different surrounding phonetic environments.*

clusters (Wells et al., 2022; Martin et al., 2023). However, there has been limited analysis on directly quantifying the relationship between the S3M feature subclusters and allophony. To this end, we design a setting that measures the mutual information between the S3M feature subcluster indices and the surrounding phonetic environment of each phoneme, which is an indicator of allophones.

First, to obtain the subclusters within MFCCs, Mel spectrograms, and S3M features, we apply the k-means algorithm with $k = 32$ clusters to the features of each phoneme $V \in \mathcal{V}$. Then, each utterance has the designated k-means cluster index I. For a dataset with a total of $|\mathcal{V}|$ phonemes, we train $|\mathcal{V}|$ different k-means models.

Note that our MixGoP uses k-means clusters as the initializer. Also, we observed few to no EM optimization steps due to high dimensionality (Wang et al., 2015). As a result, the initialized cluster centroids will likely be similar to the final centroids $\boldsymbol{\mu}_p^c$ in eq. (4) for calculating phoneme likelihood $P_\theta(\mathrm{s}|p)$.

We then compare the utterance-wise cluster indices with their allophony. Since the TORGO dataset does not provide phonetic transcriptions, we utilize the surrounding phonetic environment, which is closely linked to allophonic variation. For simplicity, we define the environment E as the natural class of the preceding and following phonemes, similar to phoneme environment clustering (Sagayama, 1989). We use the height, backness, and roundness for the vowels and the place and manner of the consonants for the natural class. For example, each /i/ and /k/ is represented as `close-front-unrounded` and `velar-plosive`,

利用聲調在自監督語音模型中對異常發音進行評估的研究成果，Yeo 等人 (2023a)。關於基於 OOD 检测器的 基線，基於 SVM 的方法通常在所有數據集上表現最差。相比之下，kNN 在 TORGO 和 L2-ARCTIC 數據集上的表現與我們提出的 MixGoP 相當，而 MixGoP 則提供最佳整體表現。

4 聲調的 S3M 特徵

第 3 節從實證上證明，利用 S3M 特徵與 MixGoP 相結合有助於提升下游性能，與其他特徵（如 MFCC 和梅爾頻譜圖）相比。本節旨在進一步驗證 S3M 對於表示具有聲調變化的個別音素的適用性。首先，我們在節 4.1 中使用維度減少技術檢查詞素級別的 S3M 特徵。接著，我們在節 4.2 中設計一個指標來量化捕捉聲調變化的能力，將 S3M 特徵與 MFCC 和梅爾頻譜圖進行比較。為了進行我們的分析，我們使用了來自 TORGO 數據集的健康語音錄音（ Rudzicz 等人，2012 ），該數據集包括金標準的音素轉寫和對齊。

4.1 積極觀察

S3M 是為了使用周圍信息重建掩蓋信號而訓練的。因此，我們假設這將允許 S3M 捕捉局部聲學特徵，包括來自各種語音環境的聲調。為了驗證這種現象，我們觀察了每個詞素的 WavLM 最終層特徵，這些特徵在所有數據集上通常表現最佳（見圖 5 ）。具體來說，我們使用 UMAP 維度減少（ McInnes 等人，2018 ）和余弦距離指標來視覺化特徵，與 Choi 和 Yeo （ 2022 ）的方法相似。我們還提取了每個十個質心最接近的四个發音，以觀察每個簇的語音環境。

圖表 2 展示了一個例子，其中包含 TORGO 健康子集的分布 /ʌ/ （在 ARPABET 中為 /AH/ ）及其環境。我們觀察到每個音素有多個群集，每個群集都有相似的音韻環境，這激發了衡量變音能力的指標。

4.2 量化聲調

先前的研究發現，S3M 特徵可以模擬多個輔音的聲母分布

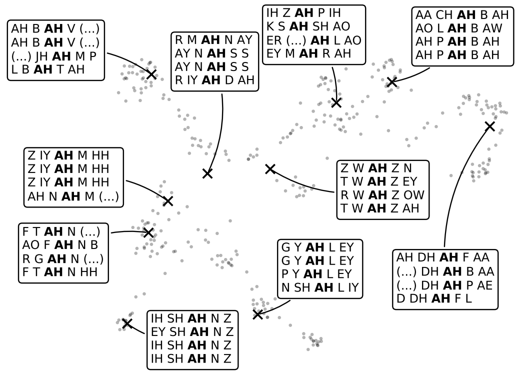


圖 2: WavLM-Large 在 TORGOhealthy 子集中的 /AH/ /AH/ 音素的可視化。音素使用 ARPABET 表示。我們觀察到 /AH/ 由子群組成，每個子群反映來自不同周圍聲學環境的音位變體。

群集（ Wells 等人，2022 ；Martin 等人，2023） 。然而，直接量化 S3M 特征子群集与音位变体之间关系的研究有限。为此，我们设计了一个设置，该设置测量 S3M 特征子群集索引与每个音素周围语音环境之间的互信息，这是音位变体的指标。

首先，為了獲取 MFCCs 、梅爾頻譜圖和 S3M 特徵中的子簇，我們將 k-means 算法應用於每個音素 $V \in \mathcal{V}$ 的特徵，並使用 $k = 32$ 個簇。然後，每個發音都有指定的 k-means 簇索引 I。對於總共有 $|\mathcal{V}|$ 個音素的數據集，我們訓練 $|\mathcal{V}|$ 個不同的 k-means 模型。

注意我們的 MixGoP 使用 k-means 群集作為初始化器。由於維度較高，我們觀察到幾乎沒有 EM 優化步驟（王等，2015 ） 。因此，初始化的群集質心可能與等式（ 4 ）中計算音素可能性的最終質心 $\boldsymbol{\mu}_p^c$ 相似 P_θ （ $\mathrm{s}|p$ ） 。

我們然後將語句級別的群集索引與其輔音變體進行比較。由於 TORGO 數據集沒有提供聲韻轉寫，我們利用周圍的聲韻環境，這與輔音變體密切相關。為了簡單起見，我們定義環境 E 為前後音素的自然類別，與音素環境分群相似（ Sagayama，1989 ） 。我們使用元音的高度、後舌位和圓唇度，以及輔音的發音位置和發音方式作為自然類別。例如，每個 /i/ 和 /k/ 被表示為 close-front-unrounded 和 velar-plosive，

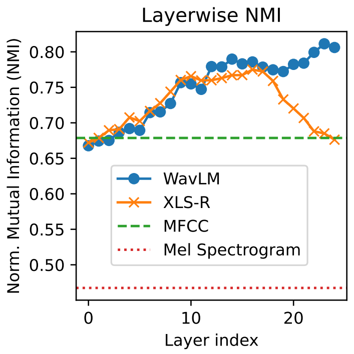


Figure 3: *Normalized Mutual Information $MI(I;E)/H(E)$ between the k-means cluster indices I and the phonetic environment E on the TORGO healthy subset. We show the layerwise NMI for S3Ms and absolute value for MFCC and Mel spectrogram.*

respectively. If the phoneme is word-initial or word-final, we also include them as the environment information. Therefore, with $|\mathcal{C}|$ number of natural classes, the number of all the possible environments is $(|\mathcal{C}|+1)^2$.

To quantify allophonic information within each subclusters, we measure the mutual information $\mathrm{MI}(I;E)$ between the cluster indices I and the environment E. We normalize the value by the environment entropy $H(E)$ so that the resulting value is between 0 and 1. We call the metric *Allophone environment-Normalized Mutual Information* (ANMI) $\mathrm{MI}(I;E)/H(E)$. The actual calculation is nearly identical to Phoneme Normalized Mutual Information (PNMI) (Hsu et al., 2021), except we replaced phoneme V to environment E.

Results. Figure 3 shows the phonetic environment information inside cluster indices for MFCCs, Mel spectrograms, XLS-R, and WavLM. For S3M models, we plot across different layers. We can observe that S3Ms contain more information on the phonetic environment compared to traditional features, implying that S3Ms successfully capture allophony. This finding introduces an interesting implication regarding the effect of varying cluster sizes when applying k-means to S3M features for discrete units (Chang et al., 2024). It is known that different cluster sizes lead to varying levels of granularity, with smaller cluster size capturing phoneme information while larger cluster size capturing speaker information (Sicherman and Adi, 2023). Our results suggest that cluster size in between may capture allophonic variations.

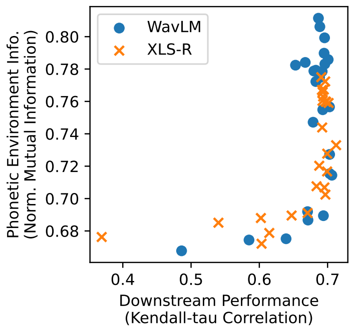


Figure 4: *Comparing phonetic environment information with the downstream task performance.*

5 Analysis

5.1 Does capturing phonetic environment leads to better downstream performance?

It is crucial to examine whether capturing phonetic environments (or allophones) actually improves downstream performance. To assess this, we compare the amount of phonetic environment information inside S3Ms and the actual downstream performance on the pronunciation assessment. In Figure 4, we observe that the downstream performance positively correlates until around NMI $<$ 0.72, where the downstream performance saturates even if the amount of phonetic environment information increases. We suspect this behavior is due to S3Ms capturing more surrounding information, which may not be useful for the pronunciation assessment task. Our hypothesis aligns with the previous empirical observation of Pasad et al. (2023) and Choi et al. (2024b) that S3Ms have non-negligible word-level modeling abilities, which requires a larger temporal receptive field. Moreover, the layerwise trends of Figure 3, *i.e.*, WavLM persistently increasing and XLS-R peaking in the middle, are also similar to previous empirical observations on word-level layerwise information (Pasad et al., 2023).

5.2 Sample efficiency of MixGoP

We randomly subsampled training set samples to check the influence of training data size. To train the GMM for each phoneme, we can either use all the occurrences in the dataset or limit the maximum number of samples. For optimal performance, we searched for the maximum number of samples between 64, 128, 256, 512, or using the full dataset. For example, if we set the maximum as 64, and /a/ and /i/ each have a total of 100 and 50 samples, to train the GMM of /a/, we randomly subsample 64

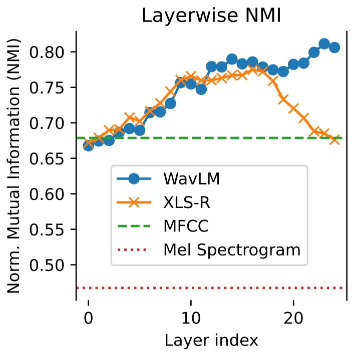


圖 3：標準化互信息 MI(I;E)/H(E) 之間的 k-means 群集索引與 TORGO 健康子集上的聲韻環境 I。我們展示了 S3Ms 的層次 NMI 以及 MFCC 和梅爾頻譜圖的絕對值。

分別。如果該音素是詞首或詞中，我們也將其作為環境信息包含。因此，在自然類別數量為 $|\mathcal{C}|+1$ 的情況下，所有可能的環境數量為 $(^2)$。

為了量化每個子群集內的變音信息，我們測量群集索引 I 與環境 E 之間的互信息 MI。我們將該值按環境熵 I 進行標準化 (E)，使得結果值在 0 到 1 之間。我們將此指標稱為 H 變音環境 - 標準化互信息 (ANMI) MIE(;)(I) E。實際計算幾乎與聲韻素標準化互信息 (PNMI) (/HHsu 等人 ,E2021) 相同，我們只是將聲韻素替換為環境。

結果。圖 3 展示了 MFCC 、梅爾頻譜圖、XLS-R 和 WavLM 的群集索引內部的聲韻環境信息。對於 S3M 模型，我們在不同的層次上進行繪圖。我們可以觀察到，與傳統特徵相比，S3Ms 包含更多關於聲韻環境的信息，這意味著 S3Ms 成功地捕捉到了變音。這一發現引出了一個有趣的含義，關於在應用 k-means 於 S3M 特徵以獲得离散單元時，變化群集大小的影響。Chang 等人 , 2024)。已知不同的群集大小會導致不同級別的粒度，較小的群集大小捕獲音節信息，而較大的群集大小捕獲講者信息。Sicherman 和 Adi,2023。我們的結果表明，中間的群集大小可能捕獲變音變化。

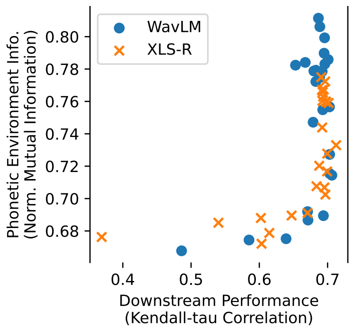


圖 4：比較聲韻環境資訊與下游任務性能。

5 分析

5.1 是否捕獲聲韻環境會導致下游性能提升？

捕獲聲韻環境（或音位）是否會實際提升下游性能，這非常關鍵。為了評估這一點，我們比較了 S3Ms 中的聲韻環境資訊量與對於發音評估的實際下游性能。在圖 4 中，我們發現下游性能與聲韻環境資訊量呈正相關，直到大約 NMI $<$ 0.72，即使聲韻環境資訊量增加，下游性能也會達到飽和狀態。我們懷疑這種行為是由於 S3Ms 捕獲了更多的周邊信息，這可能對發音評估任務並無幫助。我們的假設與 Pasad 等人 (2023) 和 Choi 等人 (2024b) 的先前經驗觀察相一致，他們指出 S3Ms 具有不可忽視的單詞級別建模能力，這需要更大的時空感受野。此外，圖 3 的層次趨勢，即，也就是說，WavLM 持續增加，XLS-R 在中間達到峰值，也與先前對單詞級別層次信息的經驗觀察相類似（ Pasad 等人 ,2023 ）。

5.2 MixGoP 的樣本效率

我們隨機抽樣訓練集樣本以檢查訓練數據大小對其的影響。為了為每個音素訓練 GMM，我們可以使用數據集中的所有出現次數，或者限制最大樣本數量。為了最佳性能，我們在 64 、 128 、 256 、 512 或使用完整數據集之間尋找最大樣本數量。例如，如果我們將最大值設為 64，而 /a/ 和 /i/ 各有 100 和 50 個樣本，為了訓練 /a/ 的 GMM，我們將隨機抽樣 64 個樣本。

Table 2: *Ablation on random subsampling.*

Dataset	64	128	256	512	Full
UASpeech	0.615	0.620	**0.624**	0.623	0.620
TORGO	0.704	0.709	0.712	**0.713**	**0.713**
SSNCE	0.539	0.545	0.549	**0.553**	0.548
speechocean762	0.502	0.516	0.528	**0.539**	0.536
L2-ARCTIC	0.182	0.178	0.181	0.182	**0.197**

samples. On the other hand, since there are only 50 samples for /i/, all 50 samples are used.

Table 2 shows that increasing the number of training samples generally improves performance. However, performance plateaus as the sample size increases, with 512 samples yielding the best results across various datasets, except for L2-ARCTIC. This suggests MixGoP performs well even with a relatively small number of samples (fewer than 100), which is advantageous for dysarthric and nonnative speech, where data is often limited. However, this also indicates that adding more data does not necessarily lead to further improvements, indicating that the model's performance may be constrained by data scalability.

6 Related works

6.1 Phoneme-level Pronunciation Assessment

Witt and Young (2000) first introduced GoP to estimate the log posterior probability of a phoneme using a Hidden Markov Model (HMM). Later improvements replaced HMMs with deep neural networks (Hu et al., 2015a,b; Li et al., 2016) and S3Ms (Xu et al., 2021; Yeo et al., 2023a; Cao et al., 2024). GoP has also been enhanced by considering additional factors, such as HMM transition probabilities (Sudhakara et al., 2019; Shi et al., 2020) and phoneme duration (Shi et al., 2020).

Section 2.2 emphasizes the usefulness of framing the pronunciation assessment of atypical speech as the OOD detection task. Yeo et al. (2023a) also addressed this by not using `softmax` for phoneme classifiers improved performance. However, the use of `softmax` during training introduced in-distribution bias. Cheng et al. (2020) modeled input probability with latent representations, but their method still depended on assessment scores to train the prediction model. Our approach improves by directly modeling phoneme likelihood instead of relying on phoneme classifiers. Furthermore, our approach explicitly accounts for allophonic variation within phonemes.

6.2 S3M Feature Analysis

Previous literature on the phonetics and phonology of S3Ms often compared downstream task performance of different layers (Martin et al., 2023; Pasad et al., 2021, 2023, 2024; Choi et al., 2024a). Linear probes (Martin et al., 2023; Choi et al., 2024a) or canonical correlation analysis (Pasad et al., 2021, 2023, 2024) are often used to measure the amount of information. Our work is complementary as previous works focus on the existence of information, whereas we further investigate on how the information is structured within the S3M features.

Also, discrete speech units from k-means clustering of S3M features have been used as the tokenizer for speech (Chang et al., 2024). Its underlying assumption comes from the feature structure being useful, *i.e.*, similar-sounding segments are close to each other. Choi et al. (2024b) showed that phonetically similar words are close to each other. Also, Baevski et al. (2020); Hsu et al. (2021); Liu et al. (2023) demonstrated that phonemes and S3M cluster indices strongly correlate with each other. Sicherman and Adi (2023); Abdullah et al. (2023) showed that natural classes are also well-clustered. Finally, Wells et al. (2022) showed that the dynamic nature of a single phoneme articulation is captured by a stream of cluster indices. Extending previous works, we focus on the multimodal nature of phonemes and demonstrate that allophones are construct subclusters within the single phoneme.

7 Conclusion

We demonstrated that improved modeling of allophony can enhance performance in OOD detection for the assessment of atypical speech, and that leveraging S3M features can further improve this performance. Specifically, our novel approach, MixGoP, addresses the limitations of uni-modality and in-distribution assumptions by employing Gaussian mixtures, which effectively model allophones and eliminate the need for softmax probabilities. Additionally, we show that utilizing S3M features further enhances OOD detection performance. Our results also confirm that S3M features capture allophonic variation more effectively than traditional features, validating the extension of our approach to include S3Ms. We evaluated eight methods across five dysarthric and nonnative speech datasets, with MixGoP achieving state-of-the-art performance on four of the datasets.

Our work provides a deeper understanding of

表 2: 隨機子樣本抽樣的消除效果。

資料集	64	128	256	512	Full
UASpeech	0.615	0.620	**0.624**	0.623	0.620
TORGO	0.704	0.709	0.712	**0.713**	**0.713**
SSNCE	0.539	0.545	0.549	**0.553**	0.548
speechocean762	0.502	0.516	0.528	**0.539**	0.536
L2-ARCTIC	0.182	0.178	0.181	0.182	**0.197**

樣本。另一方面，由於只有 50 個樣本可用於 /i/，因此所有 50 個樣本都會被使用。

表 2 顯示，增加訓練樣本數量通常會改善性能。然而，當樣本數量增加時，性能會達到平台期，各數據集在 512 個樣本時取得最佳結果，除了 L2-ARCTIC。這表明 MixGoP 即使在相對較少的樣本數量（少於 100 個）下也能表現良好，這對於失語症和非母語者的語音來說是一個優勢，因為這些情況下數據通常有限。然而，這也表明增加更多數據並不一定會導致進一步的改善，這意味著模型的性能可能受到數據可擴展性的限制。

6 相關工作

6.1 聲素級別發音評估

Witt 和 Young 於 (2000) 首次提出使用隱藏馬克夫模型 (HMM) 來估計聲素的對數後驗概率。之後的改進將 HMM 替換為深度神經網絡 (Hu 等人 , 2015a,b; Li 等人 , 2016) 和 S3M (Xu 等人 , 2021; Yeo 等人 , 2023a; Cao 等人 ,2024)。GoP 亦通過考慮其他因素得到提升，例如 HMM 轉移概率 (Sudhakara 等人 , 2019;Shi 等人 , 2020) 和聲素持續時間 (Shi 等人 , 2020)。

第 2.2 號節強調將不典型語音的發音評估作為 OOD 檢測任務的框架是有用的。Yeo 等人 (2023a) 也通過不使用 softmax 來改善聲素分類器的性能。然而，在訓練過程中引入的 softmax 導致了分布偏差。Cheng 等人 (2020) 使用潛在表示來建模輸入概率，但他們的方法仍然依賴於評估分數來訓練預測模型。我們的方法通過直接建模聲素可能性來改善，而不是依賴於聲素分類器。此外，我們的方法明確考慮了聲素內的變體。

6.2 S3M 特徵分析

S3M 的聲學和聲韻學文獻通常比較不同層的下游任務性能 (Martin 等人 , 2023;Pasad 等人 , 2021, 2023, 2024; Choiet al., 2024a)。線性探針 (Martin 等人 , 2023; Choi 等人 , 2024a) 或典型相關分析 (Pasad 等人 , 2021,2023, 2024) 常用於測量信息的量。我們的工作是補充性的，因為先前的作品關注信息的存在，而我們進一步研究了信息在 S3M 特徵中的結構。

此外，已使用 k-means 聚類的 S3M 特徵的離散語音單元作為語音的 token 器（Chang 等人，2024）。其基礎假設來自於特徵結構的有用性，卽，相似的聲段彼此靠近。Choi 等人（2024b）顯示了音韻相似的詞彼此靠近。此外，Baevski 等人（2020）；Hsu 等人（2021）；Liu 等人（2023）證明了音素和 S3M 聚類索引之間有強烈的相關性。Sicherman 和 Adi（2023）；Abdullah 等人（2023）顯示自然類別也很好地分群。最後，Wells 等人（2022）顯示單個音素的發音動態特質被一串聚類索引所捕捉。擴展先前的作品，我們著重於音素的多元模態性，並證明音位變體是單一音素內的構造子聚類。

7 結論

我們證明了改善音位變體的建模可以提升對異常語音評估的 OOD 檢測性能，並且利用 S3M 特徵可以進一步提升這種性能。具體而言，我們的新方法 MixGoP 通過使用高斯混合模型來解決單模態和分佈假設的局限性，有效地模擬音位變體並消除對 softmax 概率的需求。此外，我們還顯示利用 S3M 特徵可以進一步提升 OOD 檢測性能。我們的結果也證實 S3M 特徵比傳統特徵更有效地捕捉音位變體變化，驗證了我們的方法擴展到包括 S3Ms 的有效性。我們在五個失調和非母語語音數據集上評估了八種方法，其中 MixGoP 在四個數據集上取得了最優性能。

我們的工作對以下內容提供了更深入的理解：

how S3M representations can be hierarchically structured, from allophones to phonemes. Further, it sheds new light on the acoustic modeling perspective of speech, expanding the existing k-means-based speech discretization. It shows the possibility of using atypical speech as a benchmark to measure the quality of S3M features, especially regarding OOD robustness.

Limitations

First, a key limitation is the restricted generalizability of our findings across languages. Although we aim for our work to benefit a wide range of atypical speakers, including both dysarthric and non-native speakers, our research primarily focuses on English (four English datasets and one Tamil dataset). This limitation stems from the availability of publicly accessible datasets, but we recognize the need for broader cross-linguistic research in future work to ensure that our findings are applicable across diverse languages.

Additionally, we employed different methods for forced alignment across datasets, as outlined in Appendix A. Time alignments were either provided by the dataset or automatically generated using the Montreal Forced Aligner (McAuliffe et al., 2017). However, we did not verify the quality of these alignments in our study. This introduces the possibility that variations in alignment quality could have impacted the GoP scores, potentially affecting the overall results. While we do not primarily focus on comparing performance across datasets, future work could benefit from verifying alignment quality to ensure more reliable GoP scores, and cross-dataset comparisons.

We also acknowledge that our allophony analysis was primarily based on the TORGO dataset, which provided time alignments that were manually annotated by linguists. Extending this analysis to other datasets with similarly verified time alignments would further support the generalizability of our findings.

Finally, the method used to calculate utterance-level (Equation (7)) pronunciation scores can be improved. In our current approach, we simply averaged phoneme probabilities across each utterance; however, it is well-known that certain phonemes have a greater impact on overall pronunciation scores. While our initial analysis, as presented in Appendix C.3, provides a preliminary exploration of this issue, further investigation is needed to identify more robust approaches. Expanding upon this analysis could lead to improved techniques that more accurately evaluate atypical speech.

Ethics Statement

The risk of atypical pronunciation assessment research primarily pertains to data handling and the potential for unintended consequences in use.

Firstly, while we used publicly available datasets that have undergone prior ethical review, it is important to recognize that these datasets still contain sensitive information, particularly speakers' voices. Since no additional anonymization processes were applied in this study, we strongly recommend that any replication of this work prioritize the protection of participants' rights and privacy to the greatest extent possible.

Secondly, concerns arise regarding the potential usage of atypical speech assessment scores. These assessments may unintentionally reinforce negative stereotypes or stigmas associated with speech disorders or non-native accents. If the results are interpreted as evaluations of an individual's language ability or intelligence, they could further marginalize dysarthric or non-native speakers. Also, there is a risk in placing too much emphasis on 'correctness' in phoneme-level pronunciation assessment. Focusing heavily on accurate phoneme production prescribes a rigid, normative standard of speech, potentially penalizing linguistic diversity and variation. For both dysarthric and non-native speakers, such an emphasis might overshadow more functional measures of communication success, which may be more meaningful in real-world contexts. Despite these concerns, which warrant careful consideration, we want to emphasize that our work is intended to have a significant positive impact from an ethical perspective.

Finally, we note that ChatGPT was employed for grammatical refinement and to improve the clarity of English usage in the manuscript. We also state that every sentence generated by ChatGPT was reviewed by the authors.

如何將 S3M 表示法以分层結構進行組織，從音位到音素。此外，它從聲學建模的角度對語音提供了新的洞見，擴展了現有的基於 k-means 的語音分類。它展示了使用非典型語音作為參考標準來衡量 S3M 特徵質量的可能性，特別是關於 OOD 韌性的方面。

限制

首先，我們的研究結果在語言上的泛化能力有限。雖然我們的目標是讓我們的工作對廣泛的異常發音者有益，包括運動性障礙和非母語者，但我們的研究主要關注於英語（四個英語數據集和一個泰米爾語數據集）。這種限制來自於公共可訪問的數據集，但我們認識到未來工作需要進行更廣泛的跨語言研究，以確保我們的結果可以應用於不同的語言。

此外，我們在數據集之間採用了不同的強制對齊方法，詳情見附錄 A。時間對齊是由數據集提供或使用蒙特利爾強制對齊器（McAuliffeetal., 2017）自動生成。然而，我們並未在我們的研究中驗證這些對齊的質量。這引入了對齊質量變化的可能性可能會影響 GoP 分數，從而可能影響整體結果。雖然我們並未主要關注跨數據集的性能比較，但未來的工作可以從驗證對齊質量中受益，以確保更可靠的 GoP 分數和跨數據集比較。

我們也承認，我們的音位分析主要基於 TORGO 數據集，該數據集提供了由語言學家手工注釋的時間對齊。將這種分析擴展到其他具有類似驗證時間對齊的數據集，將進一步支持我們結果的泛化能力。

最後，計算語句級別（方程式（7））發音分數的方法可以進一步改善。在我們當前的方法中，我們只是對每個語句中的音素概率進行平均；然而，人們都知道某些音素對整體發音分數有更大的影響。雖然我們的初步分析，如附錄 C3 中所展示的，對這個問題進行了初步探索，但還需要進一步調查以確定發展更為堅固的方案。擴展這項分析可能會導致改善技術，以更準確地評估異常語音。

倫理聲明

異常語音評估研究的風險主要與數據處理以及使用中可能產生的未預期的後果有關。

首先，雖然我們使用了已經經過先前倫理評查的公共數據集，但重要的是要認識到這些數據集仍然包含敏感信息，特別是講者的聲音。由於本研究未進行額外的匿名化處理，我們強烈建議任何重複這項工作的研究應該盡可能地優先保護參與者的權利和隱私。

其次，關於異常語音評估分數的潛在使用引起了一些關注。這些評估可能無意中加強與語音障礙或非母語口音相關的負面刻板印象或偏見。如果結果被解釋為對個人語言能力的評估或智能評估，它們可能會進一步使結巴或非母語講者邊緣化。同時，在音素級別的語音評估中過度強調 ‘正確性’ 也存在風險。過度重視準確的音素產生規定了一個剛性的、規範的語音標準，可能會懲罰語言的多樣性和變化。對於結巴和非母語講者來說，這種重視可能會蓋過更具功能的溝通成功度量標準，這些標準在實際情境中可能更具意義。儘管存在這些需要謹慎考慮的關注，但我們想強調，我們的工作從倫理角度來看將產生顯著的積極影響。

最後，我們注意到 ChatGPT 被用於語法修飾和改善手稿中英語使用的清晰度。我們還聲明，每個由 ChatGPT 生成的句子都經過了作者的審查。

Acknowledgments

This work used the Bridges2 system at PSC and Delta system at NCSA through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

B. M Abdullah, M. M. Shaik, B. Möbius, and D. Klakow. 2023. An information-theoretic analysis of self-supervised discrete representations of speech. In *Proc. Interspeech*.

A. Babu, C. Wang, A. Tjandra, et al. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech*.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*.

J.A. Bilmes et al. 1998. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*.

M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga. 2020. Fast differentiable sorting and ranking. In *Proc. ICML*.

X. Cao, Z. Fan, T Svendsen, and G. Salvi. 2024. A framework for phoneme-level pronunciation assessment using ctc. In *Proc. Interspeech*.

X. Chang, B. Yan, K. Choi, et al. 2024. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study. In *Proc. ICASSP*.

F-A. Chao, T-H. Lo, T-I. Wu, et al. 2022. 3m: An effective multi-view, multi-granularity, and multi-aspect modeling approach to english pronunciation assessment. In *Proc. APSIPA ASC*.

S. Chen, C. Wang, Z. Chen, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*.

S. Cheng, Z. Liu, L. Li, et al. 2020. ASR-Free Pronunciation Assessment. In *Proc. Interspeech*.

K. Choi, J-W. Jung, and S. Watanabe. 2024a. Understanding probe behaviors through variational bounds of mutual information. In *Proc. ICASSP*.

K. Choi, A. Pasad, T. Nakamura, et al. 2024b. Self-supervised speech representations are more phonetic than semantic. In *Proc. Interspeech*.

K. Choi and E-J. Yeo. 2022. Opening the black box of wav2vec feature encoder. *arXiv preprint arXiv:2210.15386*.

B. Collins, I. M Mees, and P. Carley. 2019. Phoneme, allophone and syllable. In *Practical English Phonetics and Phonology*. Routledge.

H. Do, Y. Kim, and G-G. Lee. 2023. Hierarchical pronunciation assessment with multi-aspect attention. In *Proc. ICASSP*.

P. Enderby. 1980. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*.

T-H. Feng, A. Dong, C-F Yeh, et al. 2023. SUPERB @ SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *Proc. SLT*.

Y. Gong, Z. Chen, I-H. Chu, P. Chang, and J. Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *Proc. ICASSP*.

D. Hendrycks and K. Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. ICLR*.

A. Hernandez, E.J. Yeo, S. Kim, and M. Chung. 2020. Dysarthria detection and severity assessment using rhythm-based metrics. In *Proc. Interspeech*.

W-N. Hsu, B. Bolte, Yao-Hung H. Tsai, et al. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM TASLP*.

W. Hu, Y. Qian, and F. K. Soong. 2015a. An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech. In *SLaTE*.

W. Hu, Y. Qian, F. K Soong, and Y. Wang. 2015b. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*.

O. Jokisch, A. Wagner, R. Sabo, et al. 2009. Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system. In *Proc. of SPECOM*.

H. Kim, M. Hasegawa-Johnson, A. Perlman, et al. 2008. Dysarthric speech database for universal access research. In *Proc. Interspeech*.

H. Kim, K. Martin, M. Hasegawa-Johnson, and A. Perlman. 2010. Frequency of consonant articulation errors in dysarthric speech. *Clinical linguistics & phonetics*.

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

J. Kominek and A. W. Black. 2004. The cmu arctic speech databases. In *Proc. Speech Synthesis Workshop*.

感謝

本研究使用 PSC 的 Bridges2 系統和 NCSA 的 Delta 系統，透過先進資訊基礎設施協調生態系統：服務與支援（ AC-CESS ）計劃的分配 CIS210014，該計劃由美國國家科學基金會提供資助，資助編號為 #2138259, #2138286, #2138307, #2137603, 和 #2138296。

參考文獻

B. M Abdullah, M. M. Shaik, B. Möbius, 和 D. Klakow. 2023. 自監督語音分離表示的資訊論分析。在 *Proc. Interspeech*。

A. Babu, C. Wang, A. Tjandra, 等人。2022. XLS-R：自監督跨語言語音表示學習規模。在 *Proc. Interspeech*。

A. Baevski, Y. Zhou, A. Mohamed, 和 M. Auli. 2020. wav2vec 2.0：語音表示的自監督學習框架。在 *Proc. NeurIPS*。

J.A. Bilmes 等人 1998 年。EM 算法的温和教程及其在高斯混合模型和隱藏馬克夫模型參數估計應用。國際計算機科學研究所。

M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga. 2020. Fast differentiable sorting and ranking. In *Proc. ICML*.

X. Cao 、 Z. Fan 、 T Svendsen 和 G. Salvi。2024 年。使用 ctc 的音素級別發音評估框架。在 *Proc. Interspeech*。

X. Chang 、 B. Yan 、 K. Choi 等。2024 年。探討使用离散語音單元進行語音識別、翻譯和理解的框架：一項比較研究。在 *Proc. ICASSP*。

F-A. Chao 、 T-H. Lo 、 T-I. Wu 等。2022 年。3m：一種有效的多視角、多粒度和多維度的英語發音評估建模方法。在 *ProcAPSIPAASC*。

S. Chen 、 C. Wang 、 Z. Chen 等。2022 年。Wavlm：全堆疊語音處理的大規模自監督預訓練。*IEEEJSTSP*。

S. Cheng 、 Z. Liu 、 L. Li 等。2020 年。無 ASR 語音評估。在 *Proc. Interspeech*。

K. Choi, J-W. Jung, 和 S. Watanabe. 2024a. 通過互信息變分來理解探測行為。在 ICASSP 論文中。

K. Choi, A. Pasad, T. Nakamura 等 . 2024b. 自監督語音表示比語義更具語音性。在 Interspeech 論文中。

K. Choi 和 E-J. Yeo. 2022. 打開 wav2vec 特徵編碼器的黑盒子。arXiv 預印本 arXiv:2210.15386。

B. Collins, I. M Mees 和 P. Carley. 2019. 聲母、音位和音節。在 實用英語音韻學與音位學。Routledge 出版。

H. Do, Y. Kim 和 G-G. Lee. 2023. 基於多維度的階層性發音評估。在 ICASSP 論文中。

P.Enderby. 1980. Frenchaydysarthriaassessment. 英國語言障礙雜誌。

T-H. 风格 , A. 東 , C-F 言 , 等人 . 2023. SUPERB @ SLT 2022: 自監督語音表示學習的泛化與效率挑戰 . 於 SLT 論文集 .

Y. 強 , Z. 陳 , I-H. 朱 , P. 張 , 和 J. 璃士 . 2022. 基於 Transformer 的多維多粒度非母語者發音評估 . 於 ICASSP 論文集 .

D. 赫德瑞克斯和 K. 金佩爾 . 2017. 神經網絡中偵測錯分類和分布外範例的基線 . 於 ICLR 論文集 .

A. 賀雷茲 , E.J. 耶 , S. 金 , 和 M. 鍾 . 2020. 基於節奏的指標進行失語症檢測和嚴重程度評估 . 於 Interspeech 論文集 .

W-N. 許 , B. 博爾特 , Yao-Hung H. 蔡等 . 2021. Hu-BERT: 通過掩蓋預測隱藏單元的自監督語音表示學習 . *IEEE/ACMTASLP*.

W. 胡慶 , Y. 錢 , 和 F. K. 索恩 . 2015a. 改進的基於深度神經網絡的錯誤發音檢測和二語學習者語音診斷方法 . 於 SLaTE 論文集 .

W. Hu, Y. Qian, F. K Soong, 和 Y. Wang. 2015b. 基於深度神經網絡訓練的聲學模型與基於轉移學習的邏輯回歸分類器的誤發音檢測改善。語音通訊。

O. Jokisch, A. Wagner, R. Sabo, 等人。2009. 語言學習系統中對語音和語調評估的多語言語音數據收集。在 SPECOM 會議論文集。

H. Kim, M. Hasegawa-Johnson, A. Perlman, 等人。2008. 用於通用無障礙研究的大舌音語音數據庫。在 Interspeech 會議論文集。

H. Kim, K. Martin, M. Hasegawa-Johnson, 和 A. Perlman。2010. 大舌音語音中輔音發音錯誤的頻率。臨床語言學與聲學。

D. P. Kingma 和 J. Ba。2015. Adam：一種隨機優化方法。在 ICLR 會議論文集。

J. Kominek 和 A. W. Black。2004. CMU Arctic 語音數據庫。在 語音合成工作坊論文集。

D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, and B. Kostek. 2021. Mispronunciation detection in non-native (l2) english with uncertainty modeling. In *Proc. ICASSP*.

P. Ladefoged. 1965. The nature of general phonetic theories. *Monograph Series on Languages and Linguistics*, pages 27–42.

J. Lee, K. Lee, H. Lee, and J. Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proc. NeurIPS*.

K-F. Lee, S. Hayamizu, H-W. Hon, et al. 1990. Allophone clustering for continuous speech recognition. In *Proc. ICASSP*.

K. Li, X. Qian, and H. Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM TASLP*.

O. D. Liu, H. Tang, and S. Goldwater. 2023. Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces. In *Proc. Interspeech*.

L. MacKenzie and D. Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of british english. *Linguistics Vanguard*, 6(s1):20180061.

K. Martin, J. Gauthier, C. Breiss, and R. Levy. 2023. Probing Self-supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration. In *Proc. Interspeech*.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*.

B. McFee, C. Raffel, D. Liang, et al. 2015. librosa: Audio and music signal analysis in python. In *Proc. Python in Science Conference*.

L. McInnes, J. Healy, N. Saul, and L. Großberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*.

B. K. Ng and P. S. Chiew. 2023. L1 influence on stop consonant production: A case study of malaysian mandarin-english bilinguals. *Asian Englishes*.

T. Nguyen, H. D. Nguyen, F. Chamroukhi, and G. J. McLachlan. 2020. Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*.

A. Pasad, C-M. Chien, S. Settle, et al. 2024. What Do Self-Supervised Speech Models Know About Words? *TACL*.

A. Pasad, J-C Chou, and K. Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *Proc. ASRU*, pages 914–921.

A. Pasad, B. Shi, and K. Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *Proc. ICASSP*.

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

D. Povey, G. Cheng, Y. Wang, et al. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech*, pages 3743–3747.

F. Rudzicz, A. K. Namasivayam, and T. Wolff. 2012. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*.

S. Sagayama. 1989. Phoneme environment clustering for speech recognition. In *Proc. ICASSP*.

B. Schölkopf, J. C Platt, J. Shawe-Taylor, et al. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*.

M. Shahin and B. Ahmed. 2019. Anomaly detection based pronunciation verification approach using speech attribute features. *Speech Communication*.

M. Shahin, J. Epps, and B. Ahmed. 2024. Phonological level wav2vec2-based mispronunciation detection and diagnosis method. In *Proc. Interspeech*.

J. Shi, N. Huo, and Q. Jin. 2020. Context-aware goodness of pronunciation for computer-assisted pronunciation training. In *Proc. Interspeech*.

A. Sicherman and Y. Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *Proc. ICASSP*.

S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh. 2019. An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities. In *Proc. Interspeech*.

Y. Sun, Y. Ming, X. Zhu, and Y. Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *Proc. ICML*.

M. C. TA, T. Nagarajan, and P. Vijayalakshmi. 2016. Dysarthric speech corpus in tamil for rehabilitation research. In *Proc. TENCON*.

William F Twaddell. 1952. Phonemes and allophones in speech analysis. *Journal of the Acoustical Society of America*.

J. Vidal, L. Ferrer, and L. Brambilla. 2019. Epadb: A database for development of pronunciation assessment systems. In *Proc. Interspeech*.

Z. Wang, Q. Gu, Y. Ning, and H. Liu. 2015. High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Proc. NeurIPS*.

D. Korzekwa, J. Lorenzo-Trueba, S. Zaporowski, S. Calamaro, T. Drugman, 和 B. Kostek. 2021. 在非母語（第二語言）英語中的錯誤發音檢測與不確定性建模。在 ICASSP 會議論文。

P. Ladefoged. 1965. 一般聲學理論的性質。語言與語言學系列論著，頁面 27–42。

J. Lee, K. Lee, H. Lee, 和 J. Shin. 2018. 一個簡單的統一框架用於檢測異常分佈樣本和敵對攻擊。在 NeurIPS 會議論文。

K-F. Lee, S. Hayamizu, H-W. Hon, 等 . 1990. 適應聲學的音位群聚技術應用於連續語音識別。在 ICASSP 會議論文。

K. Li, X. Qian, 和 H. Meng. 2016. 使用多分佈深度神經網絡在第二語言英語語音中檢測和診斷錯誤發音。*IEEE/ACMTASLP* 會議論文。

O. D. Liu, H. Tang, 和 S. Goldwater. 2023. 自監督預測編碼模型在正交子空間中編碼發言者和聲學信息。在 Interspeech 會議論文。

L. MacKenzie 和 D. Turton. 2020. 評估現有強制對齊軟體對各種英式英語的準確性。 語言學先鋒，6(s1)：20180061。

K. Martin, J. Gauthier, C. Breiss 和 R. Levy. 2023. 探測自監督語音模型中的音素和音節信息：一個關於輸氣的案例研究。在 Interspeech 論文集。

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner 和 M. Sonderegger. 2017. 蒙特利爾強制對齊器：使用 Kaldi 的可訓練文本 - 語音對齊。在 Interspeech 論文集。

B. McFee, C. Raffel, D. Liang 等。2015. librosa：Python 中的音頻和音樂信號分析。在 Python 在科學會議論文集。

L. McInnes, J. Healy, N. Saul 和 L. Großberger. 2018. Umap：均勻流形近似和投影。開源軟件雜誌。

B. K. Ng 和 P. S. Chiew. 2023. L1 對塞擦音發音的影響：馬來西亞華語 - 英語雙語者的案例研究。亞洲英語。

T. Nguyen, H. D. Nguyen, F. Chamroukhi, 和 G. J. McLachlan. 2020. 使用有限混合連續密度函數進行逼近，該函數在無窮大時消失。*Cogent Mathematics& Statistics*。

A. Pasad, C-M. Chien, S. Settle, 等人。2024. 自监督語音模型對單詞的知識是什麼？*TACL*。

A. Pasad, J-C Chou, 和 K. Livescu。2021. 自监督語音表示模型層次分析。在 *Proc.ASRU*，頁面 914–921。

A. Pasad, B. Shi, 和 K. Livescu。2023. 自监督語音模型比較層次分析。在 *Proc.ICASSP*。

F. Pedregosa, G. Varoquaux, A. Gramfort, 等人。2011. Scikit-learn：Python 中的機器學習。*Journalof MachineLearning Research*。

D. Povey, G. Cheng, Y. Wang, 等人。2018. 深度神經網絡的半正交低秩矩陣分解。在 *Proc. Interspeech*，頁面 3743– 3747。

F. Rudzicz, A. K.Namasivayam, 和 T.Wolff. 2012. 異常發音者聲學和發音聲學語言數據庫 torgo. 語言資源與評估 .

S. Sagayama. 1989. 語音環境聚類應用於語音識別。在 ICASSP 論文集中。

B. Schölkopf, J. C Platt, J. Shawe-Taylor, et al. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*.

M. Shahin 和 B. Ahmed. 2019. 基於語音屬性特徵的異常檢測發音驗證方法。語音通訊 .

M. Shahin, J. Epps, 和 B. Ahmed. 2024. 基於語音學水平 wav2vec2 的錯誤發音檢測和診斷方法。在 Interspeech 論文集中。

J. Shi, N. Huo, 和 Q. Jin. 2020. 電腦輔助發音練習的語音正確性。在 Interspeech 論文集中。

A. Sicherman 和 Y. Adi. 2023. 分析离散自監督語音表示應用於語音建模。在 ICASSP 論文集中。

S. Sudhakara, M. K. Ramanathi, C. Yarra, 和 P. K. Ghosh. 2019. 考慮 HMM 穿梭概率的 DNN-HMM 系統對發音評估的改進良好發音度 (gop) 度量。在 Interspeech 論文中。

Y. Sun, Y. Ming, X. Zhu, 和 Y. Li. 2022. 使用深度最近鄰域進行分佈外檢測。在 ICML 論文中。

M. C. TA, T. Nagarajan, 和 P. Vijayalakshmi. 2016. 用於復健研究的泰米爾語失語症語音語料庫。在 TENCON 論文中。

William F Twaddell. 1952. 語音分析中的音素和音位。在聲學學會紀要期刊中。

J. Vidal, L. Ferrer, 和 L. Brambilla. 2019. Epadb：用於發音評估系統開發的數據庫。在 Interspeech 論文中。

Z. Wang, Q. Gu, Y. Ning, 和 H. Liu. 2015. 高維 EM 算法：統計優化與漸近正態性。在 NeurIPS 論文中。

D. Wells, H. Tang, and K. Richmond. 2022. Phonetic Analysis of Self-supervised Representations of English Speech. In *Proc. Interspeech*.

S. M. Witt and S. J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108.

X. Xu, Y. Kang, S. Cao, et al. 2021. Explore wav2vec 2.0 for Mispronunciation Detection. In *Proc. Interspeech*.

S-W. Yang, P-H. Chi, Y-S. Chuang, et al. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech*, pages 1194–1198.

E-J. Yeo, K. Choi, S. Kim, and M. Chung. 2023a. Speech Intelligibility Assessment of Dysarthric Speech by using Goodness of Pronunciation with Uncertainty Quantification. In *Proc. Interspeech*.

E-J. Yeo, H. Ryu, J. Lee, et al. 2023b. Comparison of l2 korean pronunciation error patterns from five l1 backgrounds by using automatic phonetic transcription. In *Proc. ICPhS*.

S. J. Young, J. J. Odell, and P. C. Woodland. 1994. Tree-based state tying for high accuracy modelling. In *Proc. HLT*.

J. Zhang, Z. Zhang, Y. Wang, et al. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech*.

G. Zhao, S. Sonsaat, A. Silpachai, et al. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*.

A Datasets

In our study, we utilize read speech at the sentence level. While the TORGO and SSNCE datasets include both word- and sentence-level data, we focused exclusively on sentences to ensure consistency, as the non-native datasets contain only sentence-level speech. Although the UASpeech dataset consists solely of word-level materials, we included it for comparison with prior work that applied GoP to dysarthric speech (Yeo et al., 2023a).

In contrast to non-native speech datasets, which provide utterance-level scores, dysarthric speech datasets offer intelligibility scores at the speaker level. Therefore, we applied these speaker-level scores to each utterance in the dysarthric datasets.

All the datasets are publicly available, with licenses that allow academic use. We used the datasets for exclusively academic purposes.

A.1 Dysarthric Speech Datasets

UASpeech (Kim et al., 2008) comprises recordings from 25 English speakers, of whom 14 have dysarthria and 11 are healthy. The severity of dysarthria was assessed using the Frenchay Dysarthria Assessment (FDA) (Enderby, 1980), categorizing four speakers as having high-intelligibility, three as mid-intelligibility, three as low-intelligibility, and four as very low-intelligibility. This study analyzes common and uncommon words selected for their diverse phonetic sequences, which are crucial for evaluating the pronunciation of phonemes in varied contexts. Although the dataset includes recordings from an 8-microphone array, we utilize only the fifth microphone for computational efficiency. In total, 6,589 utterances from healthy speakers and 8,370 utterances from dysarthric speakers are used. For time alignment information, we apply the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017).

TORGO (Rudzicz et al., 2012) consists of recordings from 15 English speakers, including 8 with dysarthria and 7 healthy individuals. The severity was determined using FDA scores, classifying two speakers as mild, one as mild-to-moderate, one as moderate-to-severe, and four as severe. To balance the classes, mild-to-moderate and moderate-to-severe speakers were merged into a single moderate category. A total of 156 healthy and 413 dysarthric utterances are used. Similar to UASpeech, we apply MFA for time alignment information. Then, we integrate alignments that were

D. Wells, H. Tang, 和 K. Richmond. 2022. 英語語音自我監督表示的語音分析。在 Interspeech 會議論文。

S. M. Witt 和 S. J. Young. 2000. 電子語言學習中的音素級別發音評分與評估。*SpeechCommunication*，30:95–108。

X. Xu, Y. Kang, S. Cao 等。2021. 探索 wav2vec 2.0 在錯誤發音檢測上的應用。在 Interspeech 會議論文。

S-W. Yang, P-H. Chi, Y-S. Chuang 等。2021. SUPERB：語音處理通用性能評估基準。在 Interspeech 會議論文，頁面 1194–1198。

E-J. Yeo, K. Choi, S. Kim 和 M. Chung。2023a. 使用發音良好度與不確定性量化評估失語症語音的語音可懂度。在 Interspeech 會議論文。

E-J. Yeo, H. Ryu, J. Lee 等。2023b. 使用自動語音轉寫比較來自五個母語背景的韓語第二語言發音錯誤模式。在 ICPhS 會議論文。

S. J. Young, J. J. Odell, 和 P. C. Woodland. 1994. 基於樹狀結構的狀態綁定以實現高準確建模。在 HLT 論文中。

J. Zhang, Z. Zhang, Y. Wang 等。2021. speechocean762：一個開源的英語非母語者語音語料庫，用於發音評估。在 Interspeech 論文中。

G. Zhao, S. Sonsaat, A. Silpachai 等。2018. L2- arctic：一個英語非母語者語音語料庫。在 Interspeech 論文中。

A 資料集

在我們的研究中，我們使用句子的讀詞。雖然 TORGO 和 SSNCE 資料集包括單詞和句子級別的數據，但我們專注於句子，以確保一致性，因為非母語者資料集只包含句子級別的語音。雖然 UASpeech 資料集只包含單詞級別的材料，但我們為了與先前將 GoP 應用於運動性失語症語音的研 究進行比較，而將其包括在內（Yeo 等人，2023a）。

與提供發音句子的非母語者語音資料集相比，運動性失語症語音資料集在講者級別提供可懂度分數。因此，我們將這些講者級別的分數應用於運動性失語症資料集中的每個發音。

所有數據集均公開提供，並附帶允許學術使用的授權。我們僅用於學術目的。

A.1 癱頸語音數據集

UASpeech （Kim 等人，2008）包含 25 位英語講者的錄音，其中 14 位有癱頸，11 位健康。癱頸的嚴重程度使用法國癱頸評估（FDA）進行評估（Enderby，1980），將四名講者的聽力分為高可懂，三名為中可懂，三名為低可懂，四名為非常低可懂。此研究分析了為了其多樣的音韻序列而選擇的常見和罕見的單詞，這對評估不同情境下音素的發音至關重要。雖然數據集包括來自 8 個麥克風陣列的錄音，但我們僅使用第 5 個麥克風以確保計算效率。總共使用了健康講者的 6,589 個發音和癱頸講者的 8,370 個發音。對於時間對齊信息，我們使用蒙特利爾強制對齊器（MFA）（McAuliffe 等人，2017）。

TORGO （Rudzicz 等人，2012）包括 15 位英語講者的錄音，其中 8 位有癱頸，7 位健康。嚴重程度使用 FDA 得分確定，將兩名講者分為輕度，一名為輕度至中度，一名為中度至重度，四名為重度。為平衡類別，將輕度至中度和中度至重度講者合併為一個中度類別。總共使用 156 個健康和 413 個癱頸發音。與 UASpeech 相似，我們使用 MFA 進行時間對齊信息。然後，我們整合對齊信息。

manually adjusted by two linguists, as outlined by Hernandez et al. (2020). These manually refined alignments will be made publicly accessible in our repository.[1]

SSNCE (TA et al., 2016) comprises Tamil speech recordings from 20 dysarthric and 10 healthy speakers. Severity is categorized based on intelligibility scores rated on a 7-point Likert scale by two speech pathologists, resulting in 7 mild (scores 1-2), 10 moderate (scores 3-4), and 3 severe (scores 5-6) speakers. Each speaker recorded 260 distinct sentences, totaling 2,600 healthy and 5,200 dysarthric utterances. We use the time stamps provided with the datasets.

A.2 Nonnative Speech Datasets

speechocean762 (Zhang et al., 2021) consists of 5000 utterances from 250 Mandarin-speaking nonnative children and adult speakers. English proficiency levels' ratio maintains a 2:1:1 ratio for good, medium, and poor, ensuring representation across different proficiencies. Our analysis focuses on total scores at the sentence level, pronunciation quality graded on a scale from 0 to 10. We use scores 9 and 10 as training data and the rest as test data. We employ forced alignments generated using the Kaldi recipe, following the experimental setup of Zhang et al. (2021). This allows for a direct comparison of S3M and Kaldi features, as the quality of phoneme alignment can affect the evaluation results (MacKenzie and Turton, 2020).

L2-ARCTIC (Zhao et al., 2018) is a non-native English corpus comprising recordings from 24 speakers, whose first languages (L1) include Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. Each speaker contributes approximately one hour of read speech derived from CMU ARCTIC prompts (Kominek and Black, 2004). For this study, we use 150 utterances per speaker, where manual annotations were available for phonememic errors such as substitutions, deletions, and additions. We use the existing data split as the train/test split, where we further exclude mispronounced utterances in the training data. Unlike other datasets, L2-ARCTIC only contains phoneme-wise mispronunciation detection labels (0/1). Therefore, we used the GoP scores and the label at the phoneme level, not at the utterance level. We use the time stamps provided with the datasets.

Table 3: *Ablation on the number of clusters for Gaussian mixtures.*

Dataset	4	8	16	32	64
UASpeech	0.613	0.620	0.621	**0.623**	0.622
TORGO	0.704	0.703	0.706	**0.713**	N/A
SSNCE	0.544	0.548	**0.553**	**0.553**	N/A
speechocean762	**0.545**	0.543	0.544	0.539	0.537
L2-ARCTIC	0.175	0.178	0.176	**0.182**	0.179

B Computational cost

Extracting S3M features for all the datasets takes less than one day for each S3M on a single NVIDIA V100 GPU. Running the experiments on all the datasets takes less than a day on a 64-CPU 128GB-memory machine for the default hyperparameter settings.

C Additional analyses and discussions

C.1 Layerwise analysis of downstream performance

It is known that different layers of S3Ms tend to encode different information (Pasad et al., 2021, 2024, 2023; Choi et al., 2024b). Hence, we further compare the layerwise trend of XLS-R and WavLM in Figure 5 to provide a guideline for which layer to use for each downstream task. WavLM tends to have decent or even the best performance in the final layer, while XLS-R features often suffer a great decrease in performance in the final layer. This indicates that the choice of which layer to use is more critical when using XLS-R compared to WavLM, whereas the last layer of WavLM generally shows decent performance.

C.2 Impact of the number of subclusters

We explored the optimal number of subclusters for the downstream task. We conducted a grid search with cluster sizes of 4, 8, 16, 32, and 64, while keeping the best model and layer index from Section 3.6 fixed. For the TORGO and SSNCE datasets, we did not test 64 clusters due to insufficient phoneme samples to train the Gaussian mixtures.

As shown in Table 3, increasing the number of clusters often results in marginal improvements in downstream performance. We attribute this to the fact that better distribution modeling tends to enhance performance. This reinforces our hypothesis that the number of subclusters does not need to exactly match the number of allophones, as sufficiently large number of Gaussian mixtures can

由兩位語言學家人工調整，詳情見 Hernandez 等人所述（2020）。這些人工精細化的對齊將在我們的存儲庫中公開提供。[1]

SSNCE （TA 等人，2016）包含來自 20 位失語症和 10 位健康講者的泰米爾語音錄。嚴重程度根據兩位語言治療師評定的 7 點李克特量表上的可懂度得分進行分類，結果有 7 位輕度（得分 1-2）、10 位中度（得分 3-4）和 3 位重度（得分 5-6）講者。每位講者錄製了 260 個不同的句子，總計 2,600 個健康和 5,200 個失語症發音。我們使用數據集提供的時間標記。

A.2 非母語語音數據集

speechocean762 （張等人，2021）包含 250 位漢語講者的非母語兒童和成人講者的 5,000 個發音。英語流利度比保持為 2:1:1，分別為良好、中度和差，確保了不同流利度的代表性。我們的分析聚焦於句子級別的總分，發音質量評分範圍為 0 到 10。我們使用得分 9 和 10 作為訓練數據，其餘作為測試數據。我們採用 Kaldi 配方生成的強制對齊，遵循張等人的實驗設置（2021）。這允許直接比較 S3M 和 Kaldi 特徵，因為音素對齊的質量會影響評估結果（MacKenzie 和 Turton，2020）。

L2-ARCTIC （趙等人，2018）是一個非母語英語語料庫，包含 24 位講者的錄音，他們的第一語言（L1）包括印地語、韓語、漢語、西班牙語、阿拉伯語和越南語。每位講者貢獻了大約一小時的讀音，來自 CMU ARCTIC 提示（Kominek 和 Black，2004{{v22}}）為了這項研究，我們使用每位講者的 150 個發音，其中對於如替換、刪除和增加等音節錯誤的語音學標記是手動可用的。我們使用現有的數據分割作為訓練 / 測試分割，並在訓練數據中排除錯誤發音的發音。與其他數據集不同，L2-ARCTIC 只包含音節級別的錯誤發音檢測標籤（0/1）。因此，我們使用 GoP 得分和音節級別的標籤，而不是發音級別的標籤。我們使用數據集提供的時間標記。

表 3：高斯混合模型中群數的消除效果。

資料集	4	8	16	32	64
UASpeech	0.613	0.620	0.621	**0.623**	0.622
TORGO	0.704	0.703	0.706	**0.713**	N/A
SSNCE	0.544	0.548	**0.553**	**0.553**	N/A
speechocean762	**0.545**	0.543	0.544	0.539	0.537
L2-ARCTIC	0.175	0.178	0.176	**0.182**	0.179

B 計算成本

為所有數據集提取 S3M 特徵所需時間不到一天，每個 S3M 在單一 NVIDIA V100 GPU 上。在 64-CPU 128GB- 記憶體機器上運行實驗，在默認超參數設置下，所需時間不到一天。

C Additional analyses and discussions

C.1 下游性能的層次分析

已知 S3M 的不同層次往往會編碼不同的信息（Pasad 等人，2021；Choi 等人，2024b；2023；2024；2023；Choi 等人，2024b；2024b）。因此，我們進一步比較了 XLS-R 和 WavLM 的層次趨勢，見圖 5，以提供針對每個下游任務應該使用哪個層次的指南。WavLM 在最後一層往往表現得相當甚至最佳，而 XLS-R 的特徵在最後一層往往會出現很大的性能下降。這表明，與 WavLM 相比，使用 XLS-R 時選擇哪個層次更加關鍵，而 WavLM 的最後一層通常表現相當不錯。

C.2 子簇數量的影響

我們探索了下遊任務的最優子簇數量。我們進行了網格搜索，簇大小為 4 、 8 、 16 、 32 和 64，同時保持第 3.6 節中固定的最佳模型和層次索引。由於 TORGO 和 SSNCE 數據集的音素樣本不足以訓練高斯混合模型，因此我們未對 64 個簇進行測試。

如圖 3 所示，增加子群組的數量往往只會帶來微小的下游性能提升。我們認為這是因為更好的分佈建模往往會提升性能。這也強化了我們的假設，即子群組的數量不一定要與輔音數量完全匹配，因為足夠大的高斯混合可以

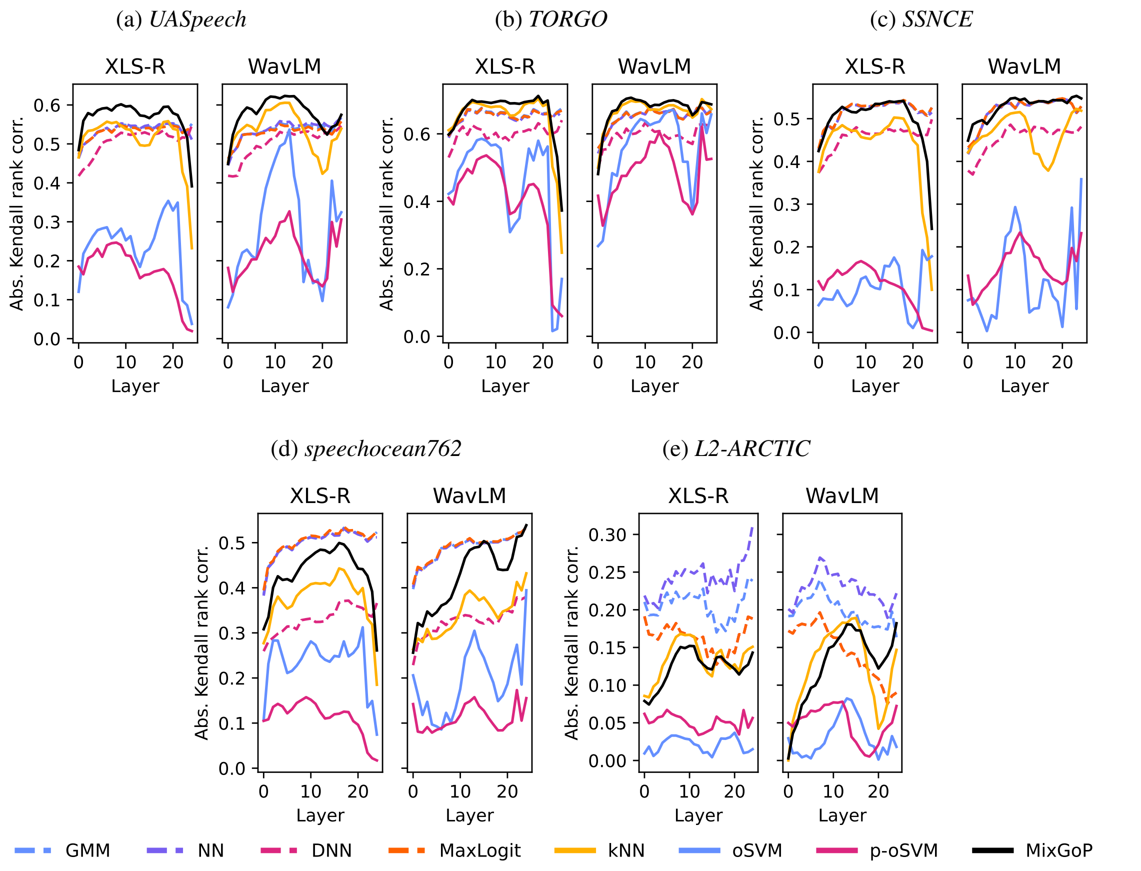

(a) *UASpeech* (b) *TORGO* (c) *SSNCE* (d) *speechocean762* (e) *L2-ARCTIC*

Figure 5: *Kendall-tau correlation coefficient when features are extracted from different layers of S3M models.*

approximate any probability density (Nguyen et al., 2020). The best performance was consistently achieved with a cluster size of 32 across datasets, with the exception of the speechocean762 dataset. For speechocean762, the highest performance was attained with the smallest cluster size, 4, although performances with cluster sizes of 8 and 16 were also similar.

To further analyze the behaviors of different clusters, we performed additional layerwise analysis on TORGO, the smallest dataset, in Figure 6. We can clearly observe that a larger number of clusters leads to better downstream performance across different layers. Especially on the best-performing layer (layer index 19 of XLS-R and 6 for WavLM), it shows bigger differences per different number of clusters. We can also observe that the number of clusters 16 and 32 is the most similar, potentially indicating the performance saturation with respect to the number of clusters.

C.3 Learnable phoneme-wise attention

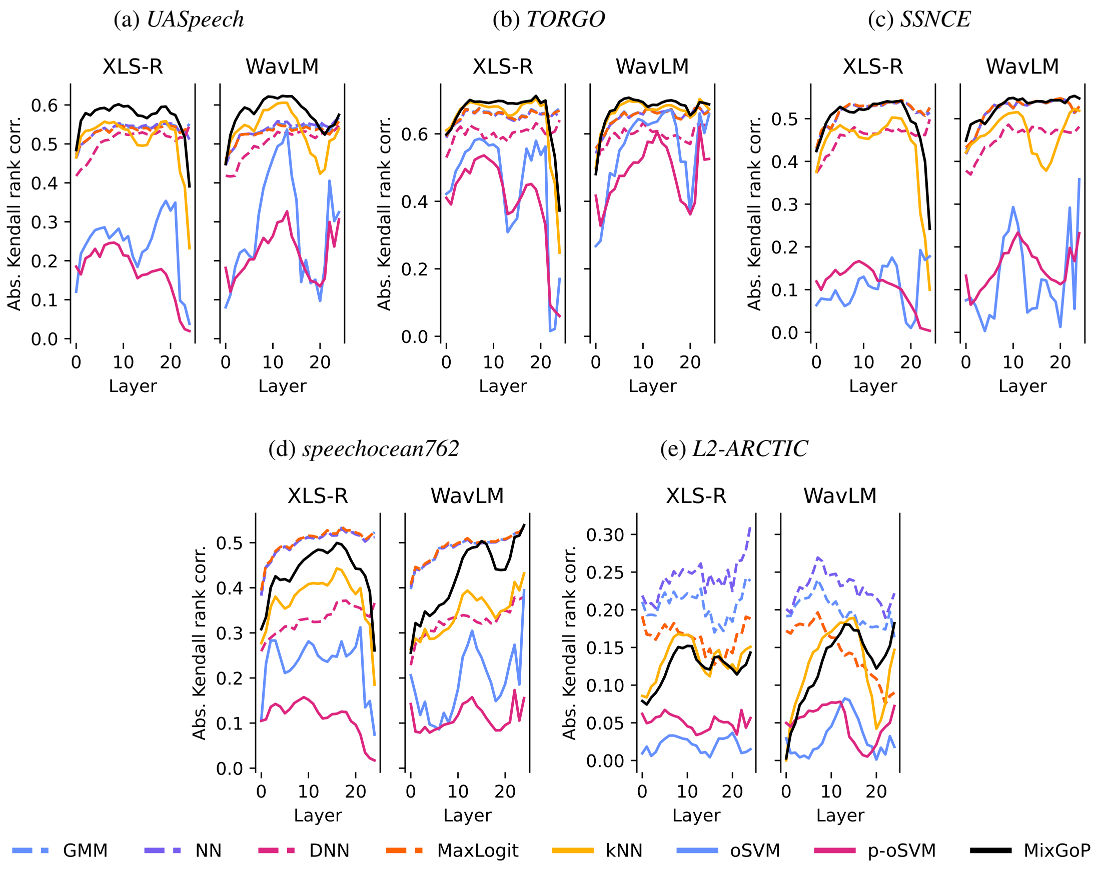
In dysarthric speech, pronunciation scores for certain phonemes are known to exert more influence on speech intelligibility (Yeo et al., 2023a). Relevant factors include their place and manner of articulation, and articulatory complexity (Kim et al., 2010). Non-native speakers, on the other hand, make different pronunciation mistakes based on their native language (Ng and Chiew, 2023; Yeo et al., 2023b).

To model the phoneme-wise importance, we design a learnable attention module $\alpha \in \mathbb{R}^{|\mathcal{V}|}$ to satisfy two conditions: (i) bounded attention weights ($0 \leq \alpha[p] \leq 1$ for any phoneme p), and (ii) the weights sum up to one ($\sum_{i=1}^{N} \alpha[p_i] = 1$):

$$\alpha[p_i] = \frac{e^{\mathbf{w}_{p_i}}}{\sum_{i=1}^{N} e^{\mathbf{w}_{p_i}}}, \qquad (8)$$

where the phoneme-wise logits $\mathbf{w} \in \mathbb{R}^{|\mathcal{V}|}$ is a learnable parameter with the vocabulary size. The formulation is nearly the same as the `softmax` function with the minor difference: if the same phoneme occurs multiple times within the utterance, it shares the same weights.

We can extend our MixGoP by combining the Gaussian mixtures (Equation (4)) and the attention


(a) *UASpeech* (b) *TORGO* (c) *SSNCE* (d) *speechocean762* (e) *L2-ARCTIC*

.

近似的任何概率密度（Nguyen 等人，2020）。在所有數據集上，最佳性能始終在 32 個簇大小時達到，除了 speechocean762 數據集。對於 speechocean762，最佳性能是在最小的簇大小 4 時達到的，雖然 8 和 16 個簇大小的性能也相似。

為了進一步分析不同簇的行為，我們在最小的數據集 TORGO 上進行了額外的分層分析，見圖 6。我們可以清楚地看到，更多的簇數導致了不同層次上的下游性能更好。特別是在最佳性能的層次（XLS-R 的第 19 層和 WavLM 的第 6 層），它顯示了隨著簇數量的不同而產生的更大差異。我們還可以觀察到，簇數 16 和 32 最為相似，這可能表明在簇數量方面性能已達到飽和。

C.3 可學習的音素級別注意力

在失語症語音中，某些音素的發音分數對語音可懂度的影響較大。對語音可懂度（Yeo 等人，2023a）。相關因素包括他們的發音位置和方式，以及發音複雜度（Kim 等人，2010）。另一方面，非母語者會根據他們的母語做出不同的發音錯誤（Ng 和 Chiew，2023；Yeo 等人，2023b）。

為了模擬音素的重要性，我們設計了一個可學習的注意力模塊 $\alpha \in \mathbb{R}^{|\mathcal{V}|}$ 來滿足兩個條件：(i) 總結束於一個音素 $0 \leq \alpha[p] \leq 1$ 的範圍內的注意力權重 (p)，以及 (ii) 權重總和為一 $(\sum_{i=1}^{N} \alpha[p_i] = 1)$：

$$\alpha[p_i] = \frac{e^{\mathbf{w}_{p_i}}}{\sum_{i=1}^{N} e^{\mathbf{w}_{p_i}}}, \qquad (8)$$

其中，音素級別的 logits $\mathbf{w} \in \mathbb{R}^{|\mathcal{V}|}$ 是一個具有詞典大小的可學習參數。該公式幾乎與 softmax 函數相同，唯一的差別是：如果同一音素在語句中多次出現，它將共享相同的權重。

我們可以通過結合高斯混合（公式 (4)）和注意力來擴展我們的 MixGoP。

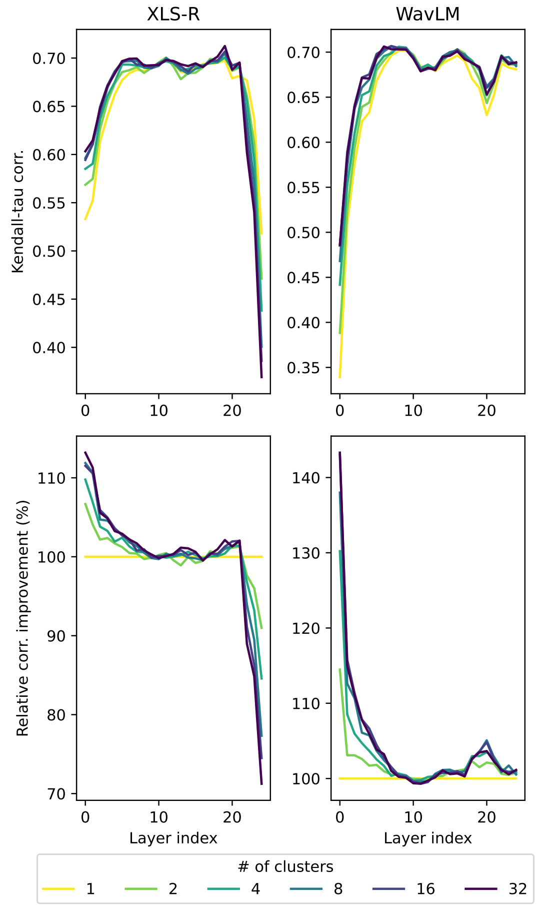


Figure 6: *Downstream performances with varying number of clusters and S3M layer index. Relative performance improvement is obtained by dividing the Kendall-tau correlation with that of cluster size 1. Except for the later layers of XLS-R with extreme performance degradation, having bigger number of clusters yield better performances.*

module (Equation (8)):

$$\texttt{MixGoPAttn}(\mathbf{x}) := \sum_{i=1}^{N} \alpha[p_i] \cdot \log P_\theta(\mathbf{s}|p). \quad (9)$$

We improve upon the phoneme-wise GoP of Equation (7) by (i) replacing the phoneme classifier $P_\theta(p|\mathbf{s})$ by the phoneme density estimator $P_\theta(\mathbf{s}|p)$ and (ii) replacing the uniform importance $1/N$ with the learnable weight $\alpha[p_i]$.

We train the phoneme-wise attention module by directly maximizing the Spearman's rank correlation coefficient between $\texttt{MixGoPAttn}(\mathbf{x})$ and the pronunciation score y (degree of dysfluency/disfluency for $\mathbf{x}$):

$$\mathcal{L} = -\texttt{spearman}(\text{MixGoPAttn}(\mathbf{x}), y), \quad (10)$$

where we freeze the Gaussian mixture models and only train the logits $\mathbf{w}$. The sorting operation within `spearman` is made differentiable by soft sorting (Blondel et al., 2020), following Blondel et al. (2020)'s implementation.

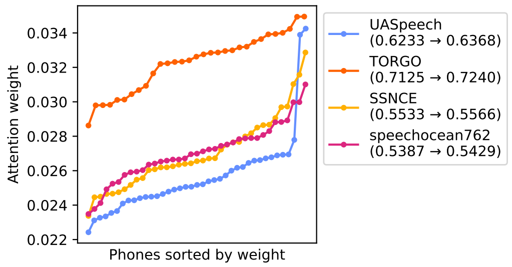


Figure 7: *Learned attention scores and the performance improvement on UASpeech, TORGO, SSNCE, and speechocean762.*

We applied 5-fold cross-validation to obtain accurate evaluation results. The phoneme attention module is trained on four datasets, excluding L2-ARCTIC as the dataset does not provide utterance-level scores. Note that this setting differs from the experiments demonstrated in Table 1, as we use the test set to train the attention module.

As demonstrated in Figure 7, applying the attention module resulted in small yet consistent performance improvements across all datasets. Furthermore, the attention weight differences between the least and most influential phonemes varied by up to 1.5 times, indicating variability in their contributions. This highlights the importance of considering the differing influence of each phoneme when calculating utterance-level pronunciation scores, which is crucial for optimizing pronunciation assessment performance.

However, we could not identify a consistent pattern across datasets regarding which phonemes consistently received higher or lower attention scores. This suggests that the variation in attention weights may be influenced by various factors, such as speakers, and phoneme distribution. Further investigation into the underlying mechanisms driving these variations is necessary to gain a deeper understanding of the differing impact of phonemes in pronunciation assessment.[6]

C.4 Comparing MixGoP and kNN

Both kNN and our proposed MixGoP heavily depend on the distances induced by S3M features. Also, kNN is often the closest competitor to MixGoP, as shown in Table 1, indicating their similar-

[6] Full list of attention weights can be found in `https://github.com/juice500ml/acoustic-units-for-ood`

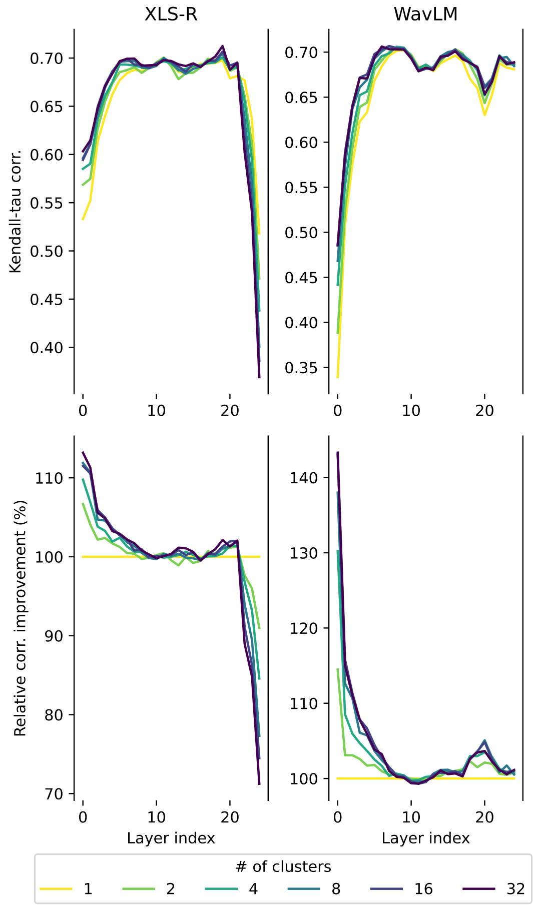


圖 6：隨著聚類數量和 S3M 層次索引的變化，下游性能。相對性能提升通過將 Kendall-tau 相關係數除以聚類大小爲 1 的相關係數來獲得。除了 XLS-R 的後期層次性能極端下降外，擁有更多的聚類數量會帶來更好的性能。

模組（方程式（8））：

$$\texttt{MixGoPAttn}(\mathbf{x}) := \sum_{i=1}^{N} \alpha[p_i] \cdot \log P_\theta(\mathbf{s}|p). \quad (9)$$

我們通過以下方式改進公式（7）的音素級別 GoP：（i）將音素分類器 P_θ（$p|\mathbf{s}$）用音素密度估計器 P_θ（$\mathbf{s}|p$）替換；（ii）將均勻重要性 $1/N$ 用可學習的權重 $\alpha[p_i]$ 替換。

我們通過直接最大化 MixGoPAttn($\mathbf{x}$) 與發音分數 y（$\mathbf{x}$ 的發音不正確 / 發音不清程度）之間的 Spearman 排名相關係數來訓練音素級別注意力模組：

$$\mathcal{L} = -\texttt{spearman}(\text{MixGoPAttn}(\mathbf{x}), y), \quad (10)$$

其中我們凍結高斯混合模型，只訓練 logits $\mathbf{w}$。排序操作

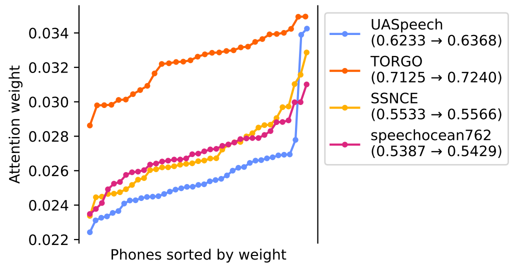


圖 7：學習到的注意力分數以及 UASpeech 、 TORGO 、 SSNCE 和 *speechocean762* 上的性能提升。

利用聲調在自監督語音模型中進行異常發音評估

我們使用五折交叉驗證來獲得精準的評估結果。該聲母注意力模塊在四個數據集上進行訓練，排除 L2-ARCTIC，因為該數據集不提供語句級別的得分。注意，這種設定與表 1中展示的實驗不同，我們使用測試集來訓練注意力模塊。

如圖 7 所示，應用注意力模塊在所有數據集上均產生了微小但一致的性能提升。此外，最不影響和最影響的聲母之間的注意力權重差異可達 1.5 倍，這說明了它們貢獻的變化。這強調了在計算語句級別發音得分時考慮每個聲母的不同影響的重要性，這對優化發音評估性能至關重要。

然而，我們在數據集中未能發現關於哪些聲母持續地收到較高或較低注意力得分的統一模式。這表明注意力權重的變化可能受到各種因素的影響，例如發音者和聲母分佈。對於推動這些變化的潛在機制進行進一步研究是必要的，以獲得對發音評估中聲母不同影響的更深理解。

C.4 比較 MixGoP 和 kNN

kNN 和我們提出的 MixGoP 都嚴重依賴於 S3M 特徵導致的距離。同時，kNN 經常是 MixGoP 最接近的競爭對手，如表 1所示，這表明它們的相似性。

完整的注意力權重清單可見於 https://github.com/juice500ml/acoustic-units-for-ood ities. However, there are multiple differences in their implementation details. kNN relies on Euclidean distance and selects the maximum distance among nearest neighbors. MixGoP employs Mahalanobis distance and measures the distances from the centroids obtained by the EM algorithm. We leave the study on the effectiveness of kNN and MixGoP's key components for future work.

然而，他們在實作細節上存在多種差異。kNN預設使用歐幾里得距離並選擇最近鄰居中的最大距離。MixGoP預設使用馬哈拉諾比斯距離並測量由EM算法得到的中心點距離。我們將關於kNN和MixGoP的關鍵組成部分之有效性的研究留待未來進行。