HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

Wei-Ning Hsu　Benjamin Bolte　Yao-Hung Hubert Tsai　Kushal Lakhotia
Ruslan Salakhutdinov　Abdelrahman Mohamed

2025 年 2 月 2 日

概要

音声表現学習のための自己教師ありアプローチは、次の 3 つのユニークな課題に直面しています：

各入力発話に複数の音声単位が存在すること、事前学習フェーズにおいて音声単位の辞書が存在しないこと、音声単位の長さが可変であり、明示的なセグメンテーションがないこと。これらの問題に対処するために、我々は Hidden-Unit BERT (HuBERT) アプローチを提案します。このアプローチは、自己教師あり音声表現学習のために設計されており、BERT のような予測損失を使用して目標ラベルを整合させます。我々の手法の重要な要素は、連続的な音声領域に対して予測損失を適用し、モデルがアコースティックモデルと言語モデルを結合したモデルとして学習することを可能にする点です。

HuBERT は、割り当てられたクラスタラベルの「内部品質」ではなく、「一貫性」を重視する未監督クラスタリングステップの信頼性に依存します。単純な K-means クラスタリングを使用し、2 回のクラスタリングを繰り返すことで、100 クラスの目標ラベルを生成します。この手法で、10 分、1 時間、10 時間、960 時間の微調整セットを用いた LibriSpeech（960 時間）および Libri-light（60,000 時間）のベンチマークにおいて、最先端の wav2vec 2.0 の性能に匹敵するか、それを上回る結果を達成しました。

1 億パラメータモデルを使用すると、HuBERT は話者適応およびテストセット全体の評価サブセットにおいて、最大で 19 %および 13 %の相対的な単語誤り率（WER）の削減を実現しました。

1　INTRODUCTION

多くの研究プログラムにおける「北極星」（究極の目標）は、赤ちゃんが母語を学ぶように、リスニングやインタラクションを通じて音声や音響の表現を学習することです。高精度な音声表現には、発話内容の分離された側面に加えて、話し方に関する非言語的な情報（例：話者の特定、感情、ためらい、中断など）が含まれます。さらに、完全な状況理解に到達するためには、発話信号に交差して重なる構造化されたノイズ（例：笑い声、咳、唇を鳴らす音、背景の車のエンジン音、鳥のさえずり、食べ物が焼ける音など）をモデル化する必要があります。

このような高精度な表現の必要性が、音声や音響の自己教師あり学習における研究を推進しました。この場合、設計された前処理タスクの学習プロセスを駆動する目標は入力信号自体から引き出されます。音声表現の自己教師あり学習の前処理タスクの例としては、近接する特徴と時間的に離れた特徴を区別するタスク [1] – [3]、音響特徴の次ステップ予測 [4]、マスクされた文脈を与えられた場合の音響特徴のマスク予測 [5], [6] などがあります。

さらに、自己教師あり学習手法は学習中に言語的リソースに依存しないため、普遍的な表現を学習することが可能です。ラベルやアノテーション、テキストのみのデータは、入力信号に含まれる豊富な情報を無視してしまうためです。ラベル付きデータの

大量収集に頼らずに音声表現を学習することは、新しい言語やドメインをカバーする工業用途や製品にとって極めて重要です。これらのシナリオを網羅する大規模なラベル付きデータセットを収集するのに必要な時間は、現在の急速に進む AI 産業における実際のボトルネックとなっています。特に市場投入までのスピードが製品の成功に重要な役割を果たしています。また、話し言葉のみの方言や言語をカバーするような、より包括的なアプリケーションを構築することも、言語的リソースへの依存を減らすことによる大きな利点です。これらの言語や方言は非標準的な正書法ルールを持つため、利用可能なリソースが非常に少ないか、全く存在しないことが多いのです。

疑似ラベル付け（Pseudo-labeling、PL）は、自己訓練とも呼ばれる半教師あり学習手法の一つであり、1990 年代半ばから成功した応用がなされてきた支配的なアプローチです [7] – [10]。PL は、特定の下流タスクで「教師モデル」を訓練するために、ある程度のラベル付きデータからスタートします。次に、教師モデルを使用してラベルのないデータに疑似ラベルを生成します。その後、教師ラベル付きデータとラベル付きデータを組み合わせて、標準的な交差エントロピー [9] 損失や、教師生成ラベルのノイズを考慮するためのコントラスト損失 [11] を用いて学生モデルを訓練します。疑似ラベル付けのプロセスは、教師ラベルの品質を向上させるために [12]、繰り返し実行される場合があります。