Impact of Masking Ratio on Discrete Speech Unit Correction Accuracy for Accent Adaptation

Assignee Research

July 6, 2026

Abstract

Self-supervised pre-trained speech models have strongly improved speech recognition, yet they are still sensitive to domain shifts and accented or atypical speech. Many of these models rely on quantisation or clustering to learn discrete acoustic units. We propose to correct the discovered discrete units for accented speech back to a standard pronunciation in an unsupervised manner. A masked language model is trained on discrete units from a standard accent and iteratively corrects an accented token sequence by masking unexpected cluster sequences and predicting their common variant. Small acc

1 Introduction

This paper examines: Unsupervised Accent Adaptation Through Masked Language Model Correction Of Discrete Self-Supervised Speech Units. Research question: What is the impact of varying the masking ratio in the masked language model on the correction accuracy of discrete speech units for accent adaptation, measured by the reduction in WER on noisy speech datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

14 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The training set consists of 827k utterances (611 hours) of English speech, including North American (350k), British (11	✓	0.34
There are 1k utterances from each accent for validation and testing.	✓	0.17
The baseline HuBERT Large model consists of 7 convolutional feature extraction layers and 24 Transformer layers, with a	✓	0.29
It has been pre-trained on the 60k hour LibriVox corpus with target clusters extracted from the 9th layer of the 2nd ite	✓	0.35
The K-means quantiser that generated the clusters has 500 centroids learned on LibriSpeech train-clean-100h.	✓	0.25
For continually pre-training the HuBERT Large model, we insert Houlsby adapters at every Transformer layer, after the at	✓	0.28
The adapters consist of a feedforward down-projection layer, a ReLU, a feedforward up-projection layer and a layer norma	✓	0.28
The bottleneck dimension is set to 1024, which performed best in [10].	✓	0.20
The target clusters are extracted with the same HuBERT Base 2nd iteration model.	✓	0.19
The proposed method improves a state-of-the-art HuBERT Large model on a downstream accented speech recognition task.	✓	0.29

References

- `http://arxiv.org/abs/2507.09116v3`

- http://arxiv.org/abs/1811.04224v1
- http://arxiv.org/abs/2309.13994v1