Self-supervised Flemish Dutch Speech Models vs. Fine-tuned English Models in Robustness to Noise and Accent Variation

Assignee Research

June 28, 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: How do self-supervised speech models pre-trained on low-resource Flemish Dutch compare to fine-tuned English models in terms of robustness to noise and accent variation on standard ASR test sets?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

12 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.7/10.

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.	✓	0.16
The baseline HuBERT Large model consists of 7 convolutional feature extraction layers and 24 Transformer layers, with a	✓	0.29
The K-means quantiser that generated the clusters has 500 centroids learned on LibriSpeech train-clean-100h.	✓	0.25
The bottleneck dimension is set to 1024, which performed best in [10].	✓	0.20
The proposed method improves a state-of-the-art HuBERT Large model on a downstream accented speech recognition task.	✓	0.28

References

- http://arxiv.org/abs/2407.13782v1
- http://arxiv.org/abs/2309.13994v1
- http://arxiv.org/abs/2109.14357v1