A Report on the Paper: *A Robust Accent Classification System Based on Variational Mode Decomposition* by Darshana Subhash, Jyothish Lal G., Premjith B., Vinayakumar Ravi

Presented by: Harishankar Nagar, Vivek Khari

February 2, 2025

Abstract

Accent classification plays a crucial role in enhancing automatic speech recognition (ASR) systems, making them more robust to variations in speech patterns across different geographical regions. This paper presents a comparative analysis of the latest models in accent classification, examining their effectiveness and limitations. We discuss the importance of the task, analyze the performance of state-of-the-art methods, evaluate the metrics used for assessment, and highlight open research challenges and future directions.

1 Introduction

Speech recognition technology has seen significant advancements in recent years, enabling its deployment in a variety of real-world applications such as virtual assistants, customer service automation, and accessibility tools. However, one of the persistent challenges in ASR is handling diverse accents effectively. Accent classification helps improve ASR accuracy by tailoring models to account for linguistic variations. Accents can introduce variations in pronunciation, intonation, and phonetic structures, making standard ASR models less reliable for speakers from different linguistic backgrounds. Therefore, robust accent classification models are essential for ensuring inclusivity and efficiency in speech-driven applications.

2 State-of-the-Art Methods

Recent advancements in deep learning have led to the development of several sophisticated methods for accent classification. Among the most effective techniques is the use of Variational Mode Decomposition (VMD) combined with Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. This approach decomposes speech signals into different modes to enhance signal clarity before extracting crucial frequency-based features for classification.

Another widely used technique is Support Vector Machines (SVM), which is particularly effective for structured feature-based classification. By learning the optimal decision boundaries, SVMs can distinguish accents with high precision. Convolutional Neural Networks (CNNs) have also been extensively used in accent classification, often applied to spectrogram representations of speech data. CNNs are adept at capturing spatial patterns in frequency-time representations, making them suitable for distinguishing accent variations.

Recurrent Neural Networks (RNNs) and their variants, such as Bidirectional Gated Recurrent Units (BiGRUs), are also explored for accent classification due to their ability to model sequential dependencies in speech. Hybrid models, such as CNN-LSTM architectures, leverage both convolutional feature extraction and temporal dependencies for enhanced classification accuracy.

3 Evaluation Metrics

The performance of accent classification models is typically measured using standard classification metrics. Accuracy is one of the primary indicators, representing the proportion of correctly classified instances. Precision and recall provide further insights into model reliability, particularly in classifying minority accents, which may be underrepresented in training datasets. The F1-score, a harmonic mean of precision and recall, is useful in assessing the balance between false positives and false negatives. Confusion matrices offer a detailed breakdown of misclassifications, enabling further analysis of model weaknesses and areas for improvement.

4 Comparison of Methods

Several studies have evaluated the effectiveness of different accent classification approaches. Traditional feature-based methods such as SVMs exhibit strong performance when trained on well-defined feature sets like MFCCs. However, deep learning approaches, particularly CNNs and hybrid models, have demonstrated superior generalization across diverse datasets. The combination of VMD with MFCC extraction has shown to be particularly promising, improving classification accuracy by reducing background noise and enhancing speech clarity.

Method	Accuracy	Precision	Recall	F1-Score
MFCC + MPSA-DenseNet	72.9%	72.4%	72.7%	72.8%
MFCC + 1D CNN-BiGRU-Attention	80.5%	79.8%	78.5%	79.0%
MFCC + SVM	92.5%	91.6%	92.0%	91.8%
MFCC + CNN + LSTM	98.63%	97.0%	96.5%	96.7%
VMD + MFCC + SVM (Proposed)	99.3%	99.1%	99.5%	99.2%

Table 1: Performance comparison of accent classification models.

5 Challenges and Open Problems

Despite advancements in accent classification, several challenges remain. One key issue is the robustness of models when applied to unseen accents or speech recorded in noisy environments. Real-world speech data often contains variations due to background noise, speech rate, and dialectal influences, making it difficult for models trained on clean datasets to generalize effectively. Another challenge is the computational cost associated with deep learning approaches. Training complex neural networks requires significant computational resources, making them impractical for deployment on low-power devices such as mobile phones. Additionally, ethical concerns related to accent bias must be addressed. ASR systems have historically exhibited disparities in performance across different demographic groups, necessitating greater inclusivity in dataset collection and model training.

6 Conclusion and Future Directions

Accent classification is a crucial component in making speech recognition technology more accessible and reliable for users worldwide. This paper has provided a comparative analysis of state-of-the-art methods in accent classification, highlighting the strengths and limitations of different approaches. The findings suggest that incorporating pre-processing techniques such as VMD, coupled with feature extraction methods like MFCC, can significantly improve classification performance. Future research should focus on improving generalization capabilities through transfer learning, developing lightweight models for real-time applications, and addressing ethical concerns to ensure fair and unbiased ASR performance across diverse linguistic groups.