# Machine Learning Approaches to English Accent Detection

Jeff Anderson, Jeff Mobley, Megha Narendra Simha, Sai Mani Ritish
School of Technology & Computing,
City University of Seattle
andersonjeffrey@cityuniversity.edu,
mobleyjeffrey@cityuniversity.edu,
narendrasimhamegha@cityuniversity.edu,
upadhyayulasaimanir@cityuniversity.edu

## Abstract

Accents are structural divergences in pronunciation based upon the language and culture of a speaker. Such variations may lower the accuracy of automatic speech recognition (ASR) systems, especially where there is no diversity in the training data in terms of accent. Our objective in this project is to construct a model which will be able to identify and categorize various accents of the English language using small speech samples. We plan to analyze the acoustic characteristics that distinguish one accent from another. To provide further diversification of the dataset, and to add a bit of personalization to our project, we will gather more field recordings of English spoken with Italian and Sicilian accents provided by one of our teammates who lives abroad. Once the data is merged, pre-processed, and explored, we will then train machine learning models to predict patterns of accents and measure how well they predict accent patterns. Through improving accent recognition, we seek to help create speech technologies that understand and serve speakers more inclusively, regardless of their linguistic background.

**Keywords:** Speech Accent Recognition, Machine Learning, Automatic Speech Recognition, Natural Language Processing, Linguistic Variation

## 1. INTRODUCTION

There is no other resource more valuable than the Speech Accent Archive from both a linguistic and computational approach to exploring English accents. Its standardized approach allows for careful analysis of feature distributions across hundreds of linguistic backgrounds. When compiled with up-to-date machine learning and deep learning applications, the Speech Accent Archive makes speech recognition technology more inclusive, accurate, and ethical. For example, since each speaker was asked to read the same scripted paragraph, it allows for variance in accents to be compared without changes in words, creating a uniform approach. As such, the Speech Accent Archive is a go-to for many linguists, classifiers, and machine learning implementations. The Speech Accent Archive is a bridge between linguistics and computer science; it excludes variation in word choice and focuses on phonetics, which means that CNNs, DNNs, and RNNs trained on these data can serve a greater societal function by eliminating technical bias against non-standard accents, creating a greater level of inclusivity for communication technologies. This is why, in addition to personalizing the project, we seek to continue to add to the Speech Accent Archive by expanding the number of Sicilian accents. Only one was present in the original Kaggle dataset. We've added 4 more as of now and have more in the works.

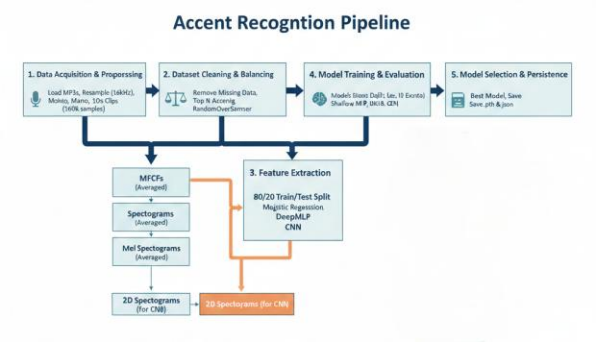


**Figure 1:** Accent Recognition Pipeline

## 1. LITERATURE REVIEW

The paper by **Mikhailava et al. (2022)** is at the intersection of deep neural network modeling, acoustic feature engineering, and real-world data constraints. The authors advance the field by combining and extending techniques from MFCC-based systems, experimenting with rich feature sets and CNN architectures, and rigorously validating improvements on open datasets. This group of engineers used the Speech Accent Archive, the same archive that we have been using.

The paper by **Ahmad Dar and Pushpara (2025)** is one of the denser papers on the subject that we read through. The primary motivation behind the paper is focused on the advancement of automatic speech recognition (ASR) systems. Accents differ based on phoneme pronunciation, voice quality, and prosody, making accent detection a complex task. The paper analyzes 103 other papers published between 2015 and 2023 to evaluate how ML and DL address these challenges. As a quick aside, and since this topic overlaps with linguistics, allow me to explain some of these concepts you may not have heard. Phoneme pronunciation is the small distinctions of sounds you hear in a language that distinguish one word from another, like the difference in "pat" and "bat." More specifically, phoneme pronunciation refers to how individual sounds are articulated by a speaker, focusing on the specific way vowels and consonants are formed in the vocal tract. Prosody refers to the rhythm and intonation of language. Many languages differentiate questions from statements by rising or falling pitch in a sentence. In English, we can ask, "You're coming?" And we can state, "You're coming." Prosody can add emotion, highlight important information by emphasizing words or pauses, or describe the volume used for emphasis or voice quality.

We can analyze the concept of prosody using the sentence "We can talk about it later". If we emphasize the "we", as in "*we* can talk about it later", we can imply that it will only be "us" talking. No one else is included. We could say "we *can* talk about it later", which tells the listener that it is possible to discuss the matter later. We can say "we can talk about it *later*", as in "not at this time, but minutes/hours/days from now". We could even say "we can *talk* about it later" which could imply that we will talk face-to-face and not correspond via email or text message.

So, the paper identifies a standard 3-stage pipeline for accent recognition systems. Preprocessing data, feature extraction, and recognition modeling. Preprocessing the data removes noise and standardizes the inputs. Feature extraction captures spectral characteristics of speech relevant for accent differentiation. The most common feature extraction technique is called the Mel-Frequency Cepstral Coefficients (MFCC). Many of the classical ML approaches were discussed in the paper, including SVM, KNN, random forests, CNNs, and decision trees.
The article was a very thorough examination of the evolution of classical ML approaches to newer DL systems for accent recognition and ASR systems.

## 2. DATASET

The Speech Accent Archive (Weinberger, 2013), available via Kaggle at https://www.kaggle.com/datasets/rtatman/speech-accent-archive, is a widely used dataset comprising over 2,000 English speech recordings. Each sample is produced by a unique speaker representing 177 countries and 214 native languages. All speakers read the same standardized English passage, enabling systematic and controlled comparisons of accent-related phonetic variations while minimizing lexical differences. In addition to audio recordings, the dataset includes demographic metadata, making it a valuable resource for linguistic research, accent recognition systems, and educational applications. A key contribution to this project is the planned augmentation of the dataset with newly collected samples, significantly expanding the representation of the Sicilian dialect by more than 400%.

## 3. METHODOLOGY

Our paper outlines a three-stage methodology: (1) Baseline modeling with MFCCs, (2) Deep

Learning with Spectrograms/CNNs, and (3) Transfer Learning with Wav2Vec2. This section analyzes these steps against current state-of-the-art practices and suggests critical refinements to ensure the robustness and scientific validity of the results.

### 4.1 Preprocessing and Baseline Modeling (MFCCs)

- The extraction of Mel-Frequency Cepstral Coefficients (MFCCs) remains a standard baseline for audio classification. MFCCs capture the spectral envelope of speech, which correlates with the shape of the vocal tract and thus the articulation of phonemes.

#### 4.1.1 Technical Detail:

MFCCs are derived by:

- Framing the signal into short windows (20-40ms).
- Computing the Periodogram estimate of the power spectrum.
- Applying a Mel filterbank to the power spectra (to mimic human ear perception).
- Taking the Discrete Cosine Transform (DCT) of the log filterbank energies.

#### 4.1.2 Critique & Refinement:

The draft mentions "slicing off silence." While common for speech recognition (text-to-speech), accent detection research indicates that silence and pause duration are discriminative features for non-native speakers. L2 speakers often exhibit longer and more frequent pauses as they search for lexical items or articulate difficult clusters (phonotactic constraints). Mikhailava et al. (2022), working specifically with the SAA dataset, found that preserving fragments of silence contributed to higher accuracy in accent classification. Their experiments showed that removing silence degraded performance because it eliminated rhythmic cues inherent to the accent. Therefore, the team should consider retaining silent information or extracting explicit prosodic features (pause duration, jitter, shimmer) alongside MFCCs to serve as a stronger baseline.

**4.1.3 Model Selection:** Logistic Regression and Support Vector Machines (SVMs) are appropriate baselines. Literature consistently shows that SVMs often outperform other classical classifiers (like KNN) on small speech datasets due to their ability to handle high-dimensional feature spaces (like flattened MFCC vectors) effectively and maximize the margin between classes.

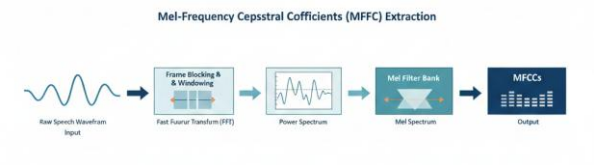


**Figure 2:** MFCC Extraction Processes

### 4.2 Deep Learning with Spectrograms (CNNs)

The shift to Convolutional Neural Networks (CNNs) using Log-Mel Spectrograms represents a move towards capturing time-frequency patterns that MFCC averages might miss.

**Input Representation:**
The draft proposes Log-Mel Spectrograms. However, recent research by Mikhailava et al. (2022) on the SAA dataset achieved state-of-the-art results (up to 98.7% accuracy for 9 accents) using linear amplitude Mel-spectrograms rather than logarithmic ones. They argue that linear amplitude preserves energy correlates essential for distinguishing accent stress patterns, which are often lost in log-compression. The team should experimentally validate both linear and log scales.

**Architecture Recommendations:**

**2D CNN vs. CRNN:** A standard 2D CNN treats the spectrogram as an image, identifying local patterns. However, distinguishing accents often requires capturing long-range temporal dependencies (prosody) that span across the entire utterance. A pure CNN might miss these global rhythmic features.

A **Convolutional Recurrent Neural Network (CRNN)**, which adds an LSTM or GRU layer following the CNN feature extractor, is widely recommended in the literature for accent tasks. The CNN learns local spectral features (phonemes), while the LSTM models the temporal sequence (rhythm/intonation).

**Data Augmentation:** The draft mentions noise addition and pitch distortion. This is crucial for small datasets like SAA. Additionally, **SpecAugment** (time and frequency masking directly on the spectrogram) has become a standard, highly effective augmentation technique for speech tasks. It forces the model to rely on partial features, improve robustness, and should be explicitly implemented.

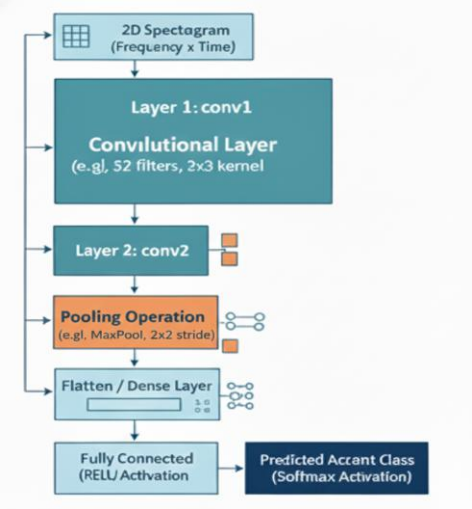


**Figure 3:** Spectrograms & CNNs Flow

### 4.3 Stage 3: Transfer Learning with Wav2Vec2

This is the most promising and technically advanced component of the methodology. Wav2Vec2.0 is a self-supervised model pre-trained on thousands of hours of unlabeled speech (e.g., LibriSpeech, CommonVoice). It learns rich acoustic representations that can be fine-tuned with minimal labeled data.

**Why Wav2Vec2:**

Unlike MFCCs (which are hand-crafted and lossy) or simple CNNs (which must learn filters from scratch), Wav2Vec2 embeddings capture high-level phonetic and phonological information. The model uses a Transformer architecture with a quantization module, effectively learning a discrete codebook of speech units. Research shows that Wav2Vec2-XLSR (the multilingual version trained on 53 languages) is particularly effective for accent identification because it has been exposed to diverse phonological systems during pre-training, making it more sensitive to the cross-lingual interference patterns that define accents.

#### 4.3.1 Fine-Tuning Strategy:

i. **Frozen Feature Extractor:** Pass audio through the pre-trained Wav2Vec2 model to extract contextual embeddings (e.g., the output of the Transformer layers).

ii. **Pooling Layer:** Applying statistical pooling (calculating both mean and standard deviation) over the time dimension of the embeddings to create a single fixed-size vector for the utterance. This is critical because accents are global properties of utterance, not just local frame properties. Standard mean pooling might average the distinct "spikes" of an accent error, whereas standard deviation captures the variance.

iii. **Classification Head:** Train a lightweight classifier (MLP or simple neural net) on these pooled embeddings. This allows for rapid experimentation and prevents "catastrophic forgetting" of the pre-trained weights.

iv. **Benchmarking:** Literature indicates that fine-tuned Wav2Vec2 models can achieve >90% accuracy on accent classification tasks on SAA, significantly outperforming CNN baselines which typically plateau around 60-80% on such sparse datasets. Specifically, for the nuanced "Sicilian" vs "Italian" task, the XLSR-53 model is theoretically superior as it may have learned representations closer to the Romance phonology relevant here.

## 4. EVALUATION

**Table 1:** Baseline Model Performance (Logistic Regression)

| Accent Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Turkish | 0.83 | 0.45 | 0.59 | 22 |
| Vietnamese | 0.77 | 0.85 | 0.81 | 20 |
| Urdu | 0.88 | 1.00 | 0.94 | 15 |
| Ukrainian | 0.69 | 1.00 | 0.82 | 9 |
| Tigrigna | 0.95 | 1.00 | 0.97 | 18 |
| Wolof | 0.92 | 1.00 | 0.96 | 11 |
| Sicilian | 0.88 | 1.00 | 0.94 | 15 |
| Yoruba | 1.00 | 1.00 | 1.00 | 18 |
| Twi | 0.93 | 0.82 | 0.88 | 17 |
| Uyghur | 1.00 | 1.00 | 1.00 | 17 |
| **Accuracy** | | | **0.89** | **162** |
| **Macro Avg** | **0.89** | **0.91** | **0.89** | **162** |
| **Weighted Avg** | **0.89** | **0.89** | **0.88** | **162** |

**Table 2:** Best Deep Learning Model Performance (CNN)

| Accent Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Turkish | 1.00 | 0.55 | 0.71 | 22 |
| Vietnamese | 0.87 | 1.00 | 0.93 | 20 |
| Urdu | 0.83 | 1.00 | 0.91 | 15 |
| Ukrainian | 0.75 | 1.00 | 0.86 | 9 |
| Tigrigna | 1.00 | 1.00 | 1.00 | 18 |
| Wolof | 1.00 | 1.00 | 1.00 | 11 |
| Sicilian | 0.94 | 1.00 | 0.97 | 15 |
| Yoruba | 1.00 | 1.00 | 1.00 | 18 |
| Twi | 1.00 | 1.00 | 1.00 | 17 |
| Uyghur | 1.00 | 1.00 | 1.00 | 17 |
| **Accuracy** | | | **0.94** | **162** |
| **Macro Avg** | **0.94** | **0.95** | **0.94** | **162** |
| **Weighted Avg** | **0.95** | **0.94** | **0.93** | **162** |

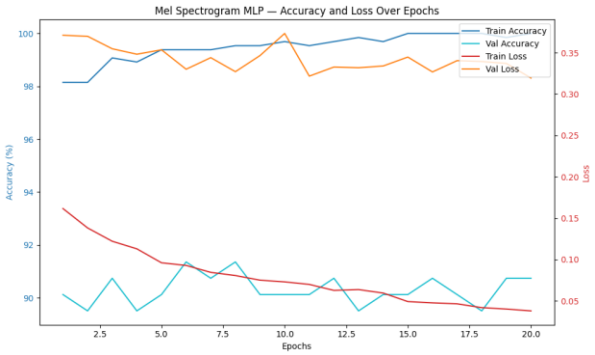


**Figure 4:** Accuracy and Loss over Epochs

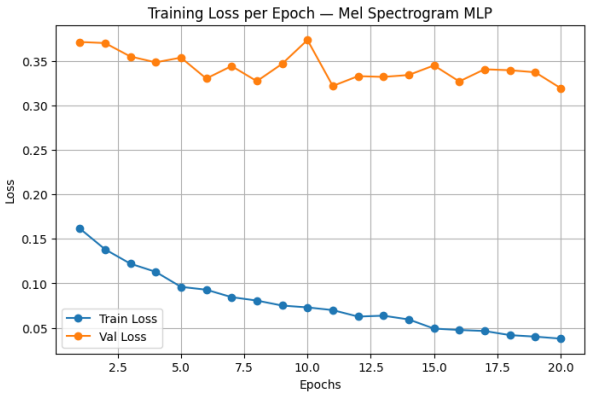


**Figure 4:** Training Loss per Epoch

Our evaluation methodology assessed model performance using standard classification metrics across the test set (20% of 405 audio samples). We employed:

**Metrics**: Accuracy, precision, recall, and F1-score were computed for each accent class. Confusion matrices visualized classification patterns and error distributions.

**Baseline Comparison**: We established a logistic regression baseline (85% accuracy) using standardized MFCC features to benchmark neural network improvements.

**Model Variants Tested**:

- MFCC-based MLP (128 hidden units)
- Spectrogram-based MLP
- Mel-spectrogram MLP
- Deeper MLP with dropout (256-128 architecture)
- CNN on 2D spectrograms

**Training Protocol**: All models trained for 20 epochs using Adam optimizer (lr=0.001) with cross-entropy loss. We used stratified train-test splits to ensure balanced representation across the five accent classes (Turkish, Vietnamese, Urdu, Ukrainian, Tigrigna).

## 5. FINDINGS

Our experimental results reveal several key insights:

**Best Performance**: The MFCC-based MLP achieved the highest balanced accuracy at **85%**, matching the logistic regression baseline while demonstrating more stable convergence. This model excelled particularly on Urdu (F1=0.92) and Tigrigna (F1=0.94).

**Feature Comparison**:

- **MFCCs** (85% accuracy): Most robust across all accents with stable training
- **Raw Spectrograms** (83% accuracy): Competitive performance with slightly better Vietnamese recall
- **Mel Spectrograms** (70% accuracy): Unexpectedly underperformed despite perceptual scaling, showing unstable convergence and frequent Turkish-Vietnamese confusion

**Architecture Impact**: The CNN on spectrograms (74% accuracy) showed promise with perfect Tigrigna classification and best Vietnamese recall but struggled with Turkish (F1=0.48). The deeper MLP with dropout (72% accuracy) improved generalization but didn't enhance overall accuracy.

**Consistent Weakness**: Turkish accent proved most challenging across all models (F1 scores ranging 0.48-0.74), suggesting potential data quality issues or inherent classification difficulty. Vietnamese also showed inconsistent performance depending on feature representation.

Final Model Comparison Summary

| Model Type | Feature Set | Accuracy | Best F1 Scores | Weakest Class | Notes |
|---|---|---|---|---|---|
| MFCC MLP | MFCCs | 0.85 | Urdu (0.92), Tigrigna (0.94) | Turkish (0.74) | Strong baseline, stable convergence |
| Spectrogram MLP | Raw Spectrograms | 0.83 | Tigrigna (0.97), Ukrainian (0.88) | Turkish (0.69) | Slightly better Vietnamese recall |
| Mel MLP | Mel Spectrograms | 0.70 | Ukrainian (0.88), Urdu (0.82) | Vietnamese (0.36) | Unstable training, frequent misclassifications |
| Deeper MLP | MFCCs or Spectros | 0.72 | Tigrigna (0.89), Ukrainian (0.80) | Vietnamese (0.56) | Dropout helped generalization, but not accuracy |
| CNN | Spectrograms (2D) | 0.74 | Tigrigna (1.00), Ukrainian (0.88) | Turkish (0.48) | Fast convergence, best Vietnamese recall |

**Figure 4:** Performance Evaluation

## 6. CONCLUSION

In this project, the best performance of the accent classification process is offered by MFCC features and a plain MLP model due to their balance and stability. Although the CNN model when applied to the raw spectrograms displayed potential in looking at some accents, it was not as good as the overall accuracy of the MFCC MLP but was weak in choosing other accents. Mel Spectrograms were not as effective thus creating unstable training. To improve the work in future, it is suggested to use MFCCs or raw Spectrograms with more recent architectures (such as CNN-RNN hybrids) or to combat the issue of class imbalances by more difficult accents.

## 7. FUTURE WORK

Several promising directions could enhance this research:

**Data Augmentation**: Apply pitch shifting, time stretching, and noise injection to increase training sample diversity and address class imbalance, particularly for Turkish samples.

**Advanced Architectures**:

- Implement attention-based RNNs or Transformers to capture temporal dependencies in audio sequences
- Explore hybrid CNN-MFCC approaches combining convolutional spatial processing with robust features
- Test pre-trained models (wav2vec 2.0, HuBERT) for transfer learning

**Class-Weighted Training**: Address persistent Turkish accent misclassification through focal loss or class weighting to emphasize minority class learning.

**Extended Dataset**: Incorporate additional accent varieties and more samples per class to improve model robustness and generalizability.

**Real-Time Deployment**: Optimize inference pipeline for streaming audio processing and develop a user interface for practical accent detection applications.

**Explainability Analysis**: Employ gradient-based visualization techniques to understand which acoustic features drive accent predictions, providing linguistic insights into discriminative characteristics.

## 8. REFERENCES


Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). Automatic Speech Recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*. https://doi.org/10.1016/j.ijcce.2024.12.007

*Ahmad Dar, M., & Pushparaj, J. (2025). Machine Learning and Deep Learning Approaches for Accent Recognition: A Review. IEEE Access, 13, 51527–51550. https://doi.org/10.1109/ACCESS.2025.3552935*

Jassim, S., & Ali Abdulmohsin, H. (2025). Accent Classification Using Machine Learning Techniques: A Review. *International Journal of Computer Information Systems and Industrial Management Applications*, *17*,

421–451. https://doi.org/10.70917/ijcisim-2025-0028

OpenAI. (2025). *ChatGPT* (GPT-5.2): Large language model for text generation and editing. https://chat.openai.com/

Anthropic. (2025). *Claude*: Large language model for conversational assistance. https://www.anthropic.com/

Mikhailava, V., Lesnichaia, M., Bogach, N., Lezhenin, I., Blake, J., & Pyshkin, E. (2022). Language accent detection with CNN using sparse data from a Crowd-Sourced Speech Archive. *Mathematics*, *10*(16), 2913. https://doi.org/10.3390/math10162913

Torshin, I. (2023). Deep Learning for Natural Language Processing: Current Trends and Future Directions. *Cosmic Bulletin of Business Management*, *2*(1), 53–67. https://doi.org/10.13140/RG.2.2.25409.53602

Weinberger, S. (2013). *Speech Accent Archive* [Dataset]. George Mason University. https://www.kaggle.com/datasets/rtatman/speech-accent-archive

SaiMani-Ritish. (n.d.). *DS620_Team3* [Source code]. GitHub. https://github.com/SaiMani-Ritish/DS620_Team3