Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 1, February, 2026

Accent Classification Across Continents: A Deep Learning Approach

Md. Fahad Hossain [1*], Anzir Rahman Khan [2], Md. Sadekur Rahman [2], Ohidujjaman [3]

[1] *Department of Software Engineering, Daffodil International University, Birulia, Dhaka 1216, Bangladesh.*

[2] *Department of Computer Science and Engineering, Daffodil International University, Birulia, Dhaka 1216, Bangladesh.*

[3] *Department of Computer Science and Engineering, United International University, Badda, Dhaka, 1212, Bangladesh.*

Abstract

This study focuses on a deep learning based accent classification across continents and greatly enhances speech recognition systems by identifying the accents of Asia, Europe, North America, Africa, and Oceania. The Convolutional Neural Network (CNN) was trained on the Mozilla Common Voice dataset, which comprises the features extracted - Mel-Frequency Cepstral Coefficients, Delta, Delta-Delta, Chroma Frequency, and spectral features- and trained to classify accents. Multiple convolutional and dense layers for accent classification were combined with dropout and batch normalization layers to avoid overfitting during training. Out of the total validation data, 82% accuracy has been achieved. The Asian and European accents were classified with greater accuracy since their datasets were larger, whereas African and Oceanian accents were more misclassified due to limited representation and the greater diversity of languages. In contrast to the past research, which focused only on country-based accent classification, this work introduced a feature based deep learning approach of continent-based accent classification along the way. The recognition of this accent variation, in turn, helps integrate and improve various aspects of speech recognition systems and makes their application more inclusive for voice assistants and language learning tools with diverse linguistic patterns. The future work will concentrate on extending the dataset to the seven continents while enhancing classification accuracy via better feature engineering and model tuning.

Keywords:

CNN;
Accent Classification;
Speaker Recognition;
Continent;
MFCC.

Article History:

Received:	11	November	2024
Revised:	09	January	2026
Accepted:	22	January	2026
Published:	01	February	2026

1- Introduction

Accents are speech characteristics that reflect socio-linguistic and cultural backgrounds. They affect pronunciation, tonal variation, and lexical choice, and form a distinctive feature of spoken language [1]. Accent studies have great relevance in some areas, such as speech recognition, sociolinguistics, and artificial intelligence. While accents help identify individuals, they also impact the usability of speech-based technologies, particularly Automatic Speech Recognition (ASR) systems [2, 3]. Although ASR has improved, these systems are still plagued by issues in recognizing and processing speech from people of varying accents, consequently leading to performance degradation [4, 5]. Voice-assisted technology has transformed human interaction with digital systems. Voice assistants like Google Assistant, Siri, and Alexa have now found their home in commerce, healthcare, and education, granting hands-free control and a real-time response. Sheng & Edmund [6] show these systems struggle with accent variation, tending more to Standard English pronunciations and misinterpreting non-standard accents. This bias forces many users to modify their natural speech patterns to be understood, thereby reducing accessibility and inclusiveness.

* **CONTACT**: fahadhossain.swe@diu.edu.bd

DOI: http://dx.doi.org/10.28991/ESJ-2026-010-01-030

Arslan et al. [7] explained that accent is one of the most significant characteristics affecting speaker-independent recognition algorithms, alongside gender [8, 9] that influence speaker-independent recognition algorithms. In addition, there are age and sex, socioeconomic background, and bilingualism that further complicate accent classification. For instance, multi-lingual speakers tend to switch their accents depending on the language they are switching to. Consequently, these are variance phenomena that old-style classification methods will not satisfactorily capture [10, 11]. Accents have the power to influence not only how people communicate but also to serve as identifiers by symbolizing their home country or origin. The motivation behind this study is based on the necessity to develop a robust ASR system that can deal with different accents on a certain continental scale. While most of the existent research targets language-specific accent or a country's accent as a whole, this study fills in the gap by introducing the classification of accents at their continental affiliation using a deep learning approach. Thus, allowing advance to be applied in ASR systems for more inclusivity in managing the different linguistic trends with covering new grounds in a well-organized manner.

ASR systems depend on huge corpora used for training, which usually have biases for the most widely spoken or standard-accented voices. In these acoustics, certain accents remain underrepresented, such as most African and Oceanic ones, which have been proven to have a higher misclassification score for speakers from such regions Most ASR systems perform well with American, British, and Australian accents but fail to recognize the speech of speakers with African, South Asian, or other regional accents. Hence, the lack of linguistic diversity in ASR training datasets seems to be a major constraint limiting the applicability of the technology in use for users worldwide.

This study elaborates upon a novel deep learning architecture for accent classification at a continental level while seeking to ameliorate challenges posed in earlier studies. The presented model uses a CNN architecture trained on samples of speech from the Mozilla Common Voice dataset that provided a bird's-eye view of accent peculiarities present in the English language as spoken by people across Asia, Europe, North America, Africa, and Oceania.

The CNN model was thus expected to capture the phonetic and spectral variations characteristic of different continents for accurate classification. Feature extraction is highly significant in accent classification because of the acoustic features such as MFCC, Delta, Delta-Delta, Chroma Frequency, Spectral Centroid, etc. which are used in the process of accent classification performance enhancement. They capture phonetic nuances, as well as pronunciation differences of speakers across continents [12]. With dropout and batch normalization for minimizing overfitting, convolutional and dense layers constitute the model architecture. CNN-based architectures are compared to conventional machine learning models to find that they are better suited to tackling the complexities posed by variation in accents [13].

This study makes several key contributions to accent classification:

- This research discusses the designing and building in a solid computational system for the identification and classification of accents in various continents, with special emphasis on accents of different kinds of English. The model furthermore enhances speaker verification and identity confirmation by highlighting different aspects of such an individual's language background as well as the authenticity of an individual related to it.
- Speech-Enabled Technologies Improvement: The technique suggested would upgrade existing speech applications, such as virtual assistants, automated transcription systems, and machine translation systems. The addition of accent accommodation expands and improves the functionalities of a speech-acoustic in terms of understanding among and across different multilingual environments.
- The study conducts a thorough investigation into the phonetic and pronunciation differences that aid in second-language acquisition and pedagogy. Findings therefore improve the language modeling to be more open to other accent variations while easing communication and understanding with regard to language teaching and cross-cultural communication.

2- Related Works

Accent classification is an active area of research. Heteroscedastic Linear Discriminant Analysis (HLDA) and Maximum Mutual Information (MMI) were first used by Choueiter et al. [14] to identify the correct groups of English accents. They were able to achieve 32% accuracy. Several studies have since explored different methods to enhance accent recognition across various languages and dialects.

Zheng et al. [15] and Long et al. [16] worked on the categorization of the Chinese speech accents. The accuracy attained by Long et al. has been 80.8 using RASTA-PLP algorithm along with a Naïve Bayes classifier. Joseph & Upadhya [17] assessed native Indian accents like those of Bengali, Gujarati, Malayalam, and Marathi using Dynamic Time Warping algorithm and obtained 63.4% accuracy. The work of Kibria et al. [18] focused on regional Bengali accents (Sylhet & Dhaka). Mannepalli et al.'s [19] research used k-NN which proved successful in classifying different Telugu dialects (Telangana, Rayalaseema, and Coastal Andhra) accurately. By using k-NN, Ma et al. [20] reported an accuracy of 79.78% in recognizing Malaysian-accented English. Danao et al. [21] Found Multi-Layer Perceptron (MLP) to be the best classifier for Telugu speech; exhibiting an accuracy of 93.33%. Hossain et al. [22] differentiated six

regional accents of the UK using traditional machine learning from 20 MFCC features and achieved an impressive accuracy of 99%. Pedersen et al. [23] performed classification of two English accents using SVM based on MFCC features extracted from segments of short speeches. Deshpande et al. [24] differentiated between American English and Indian-accented English by formant frequencies.

More recent research has shifted toward deep learning-based methods. Berjon et al. [25] attempted to analyze accent classification issues in speech recognition, contrasting classical machine learning with CNNs. They proposed a spectrogram-based French accent classification scheme. Lesnichaia et al. [26] created a CNN for an English accent classification algorithm into Germanic, Romance, and Slavic accents. This good outcome illustrates the effectiveness of the mel-scale amplitude spectrogram in discriminating accent classes. A study by Kashif et al. [27] created the Multi-Kernel Extreme Learning Machine (MKELM) framework for classifying foreign-accented English. It combines MFCCs and prosodic features with pairwise binary classifiers. Zhang et al. [28] suggested a way to use an extra ASR task to pull out phonetic features that are important for identifying languages. Their method uses both fixed and trainable acoustic models to combine embeddings. This makes language-related acoustic features more stable by using a hybrid framework. Gomez et al. [29] used the SpeechBrain and Common Voice datasets for English, Italian, German, and Spanish accents along with the ECAPA-TDNN and Wav2Vec 2.0/XLSR architectures to classify accents in multiple languages. Gong et al. [30] proposed a layer-wise adaptation method for ASR to handle diverse accents dynamically. Multi-DenseNet, PSA-DenseNet, and MPSE-DenseNet models were created by Song et al. [31] to help classify English accents. These models combine multi-task learning and PSA attention mechanisms. Accent variability also impacts speech emotion recognition (SER). Dharshini & Rao [32] looked into how accent recognition could improve SER performance by using statistical functions to pull out features at the utterance level. They conducted tests on the CREMA-D dataset and discovered that SSC features only functioned effectively in noisy environments after training on noisy data. On the other hand, MFDWC features were strong in both clean and noisy settings.

Several studies have focused on building datasets for accent classification. Demirşahin et al. [33] introduced open-source, multi-speaker speech corpora for English accents, including Southern England, Midlands, Northern England, Welsh English, Scottish English, and Irish English. Singh et al. [34] trained a two-layer CNN on a dataset covering Arabic, English, French, Mandarin, and Spanish accents, achieving promising results.

Despite significant advancements, there remains a gap in continent-based accent classification. Existing studies primarily focus on country-level or language-specific accents. This research aims to address this gap by implementing a comprehensive pipeline that includes data preparation techniques such as noise reduction, pre-emphasis, and post-emphasis. Our study also emphasizes extracting voice features to enhance classification accuracy. Table 1 details a comparison between the proposed work and previous research.

Table 1. Comparison between proposed work and other related works

Authors	Language	Region	Accent Category	Reported Performance
Jayne et al. [3]	English	UK	Irish, Midland, Northern, Scottish, Southern, Welsh	-
Kashif et al. [27]	English	-	Native English - Non native English	84.72%
Sheng & Edmund [6]	English	-	Native English - Non native (Korean, Chinese)	69%
Lesnichaia et al. [26]	English	-	Germanic, Romance and Slavic	98.70%
Hossain et al. [22]	English	UK (Ireland), Midland, Northern England, Scotland, Southern England, Wales	Irish, Midland, Northern, Scottish, Southern, Welsh	98.48%
Upadhyay & Lui [35]	English	China, India, France, Germany, Turkey, and Spain.	Native English - Non native English	71.9%
MA et al [20]	English	Malaysia	Malaysian English	79.78%
Danao et al [21]	Tagalog	Philippines.(Batangas, Cavite, Laguna, Quezon and Rizal)	Talisay, Maragondon, Paete, Lucban, and Taytay	>90%
Joseph & Upadhya [17]	Indian	India	Bengali, Gujarati, Malayalam and Marathi	63.4%
Present research	English	Asia, Europe, North America, Africa, and Oceania	Asian, European, North American, African, Oceanian.	82%

3- Research Methodology

3-1- Process Workflow

A process workflow is a set of actions or processes that are sequentially carried out in order to complete a task or acquire the desired outcome. Figure 1 shows the process workflow for this research.

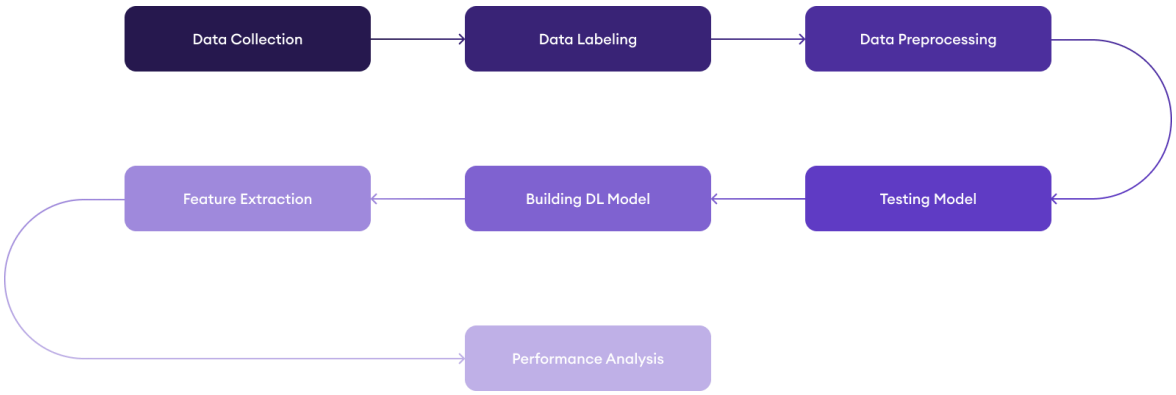

Figure 1. Working Process

3-2- Data Collection

In order to conduct the research, the dataset was collected from Mozilla Common Voice [36] dataset. Volunteers provided voice samples in a range of languages, which were collected to generate the Common Voice dataset. But exclusively recorded speech in English was used as voice metadata data to generate the continent-based dataset. This research separated the speech data into the appropriate nations. In the next step, the nation data was moved to the relevant continent. Figure 2 shows the entire data collecting process, while Table 2 lists continent wise collection of this dataset.

Figure 2. Data Collection Process

Table 2. Continent wise dataset

Continent	Dataset
Europe	7500
North America	7500
Asia	7194
Oceania	3687
Africa	518

3-3- Feature Extraction

Among all the features, Cepstral domain features such as MFCCs and LPCCs have shown the best performance in terms of continuous speech display. After experimenting with a number of features and testing different algorithms with different subsets of these features, it was found that the CNN algorithm combined with MFCCs, Deltas, Delta-Deltas, Chroma Frequency, Zero Crossing Rate, Spectral flux, Spectral centroid, Spectral bandwidth, RMS, Pitch, and Line Spectral Frequencies performed the best for accent classification based on continent.

- **Mel Frequency Cepstral Coefficients (MFCCs):**
  - **Preemphasis:** The first step is to apply a pre-emphasis filter to the audio signal to amplify the higher frequencies. This step helps in equalizing the signal and improving its spectral characteristics. The pre-emphasis filter is typically a first-order high-pass filter, and the equation for pre-emphasis is given by Equation 1:

$$h[k] = g[k] - \alpha * g[k-1] \quad (1)$$

where, h[k] is the pre-emphasis signal, g[k] is the actual signal, and α is the preemphasis coefficient.

  - **Framing:** In this step, the pre-emphasized signal is split up into equal-length short frames. This step helps in capturing the temporal variations in the signal. Common frame durations are between 20-40 milliseconds, and the frame size is usually a power of two to facilitate efficient computation [37]. As 13 features were taken only, thus, 13 MFCC characteristics constitute the pre-emphasized signal. These 13 sections had to be in the same place in order for this to be more practical.

- **Calculating Power Spectrum:** The windowed frames are then passed through the FFT algorithm to obtain each frame's frequency spectrum. In this stage, the signal is converted from the time domain to the frequency domain. Each frame has been subjected to the DFT in order to compute that [38]. The following Equation 2 was used for this step:

$$Pi = \frac{1}{N}\left|\sum_{n=1}^{N} Si(n)\, h(n) e^{-i2\pi kn/N}\right|^2,\ 1 \le k \le \mathrm{K} \tag{2}$$

Here, Si: Input signal at time index, $h(n)$: Window function (e.g., Hamming, Hanning), N: Number of points in the Discrete Fourier Transform (DFT), k: Frequency bin index, K: Maximum frequency bin, $e^{-i2\pi kn/N}$: Basis function of the Fourier Transform

- **Applying Mel Filtebank:** The MFCC's implementation requires the use of Mel Filterbank. This approach makes use of a collection of 40 equal-area filters [23]. Each filter's mathematical representation is shown below:

$$H_i(k) = \begin{cases} 0 & \text{for} \quad k < f_{c_{i-1}} \\ \frac{2\left(k - f_{c_{i-1}}\right)}{\left(f_c - f_{c_{i-1}}\right)\left(f_{c_{i+1}} - f_{c_{i-1}}\right)} & \text{for} \quad f_{c_{i-1}} \le k \le f_{c_i} \\ \frac{2\left(f_{c_{i+1}} - k\right)}{\left(f_{c_{i+1}} - f_{c_i}\right)\left(f_{c_{i+1}} - f_{c_{i-1}}\right)} & \text{for} \quad f_{c_i} \le k \le f_{c_{i+1}} \\ 0 & \text{for} \quad k > f_{c_{i+1}} \end{cases}, \tag{3}$$

Here, $H_i(k)$ the magnitude response of the i[th] Mel filter, k is the frequency bin index in the FFT,$f_{c_{i-1}}$, f_{c_i},$f_{c_{i-1}}$ is the boundary frequencies of the triangular Mel filter; where $f_{b_{i-1}}$ is left boundary (starting point) of the filter, f_{b_i}, is Center frequency of the filter and $f_{b_{i+1}}$ Right boundary (ending point) of the filter.

- **Calculating the energy log:** The logarithm of the magnitude of the filtered spectrum is computed to obtain the log-scale representation. This process compresses the dynamic range of the spectrum.
- **Calculating the log energies' discrete cosine transform (DCT):** Applying the discrete cosine transform DCT to the log-scaled spectrum finally yields the MFCC coefficients. The DCT takes away the correlations between the Mel filterbank energies, which gives us a set of coefficients that describe the signal's spectral features.

This study has only used the first 13 features out of the 40 energy features that were calculated. Because the signal is primarily represented by the first 13 features, which are linear features. The first 13 MFCCs were used, as they capture most of the essential spectral information while excluding higher coefficients that may introduce noise. These coefficients represent the short-term power spectrum of sound and help the model identify unique accent patterns. MFCC feature values for each label is shown in Figure 3.

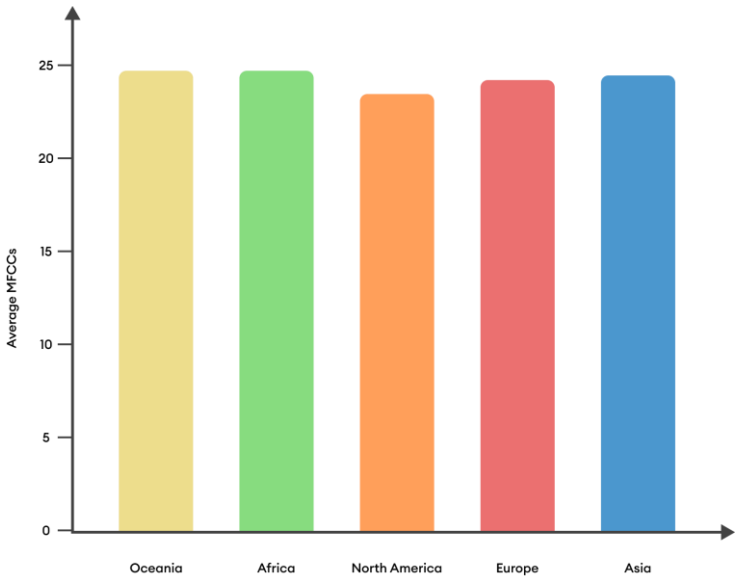

Figure 3. Average MFCCs for each label

- **Delta Features:** The main disadvantage of MFCC, despite the fact that it has been the most successful method for expressing speech, is that it only employs static data. Combining both static and dynamic features can cut error rates in half while improving the model, as it was later demonstrated that dynamic features

also convey some crucial information [39]. This study used a single set of dynamic characteristics called delta features to do this. From MFCC features, delta features are calculated using the first order derivation. 13 MFCC characteristics were used to generate 13 delta features. These dynamic characteristics supplemented our model's features and helped us make improvements. Figure 4 shows the mean delta feature for each continent.

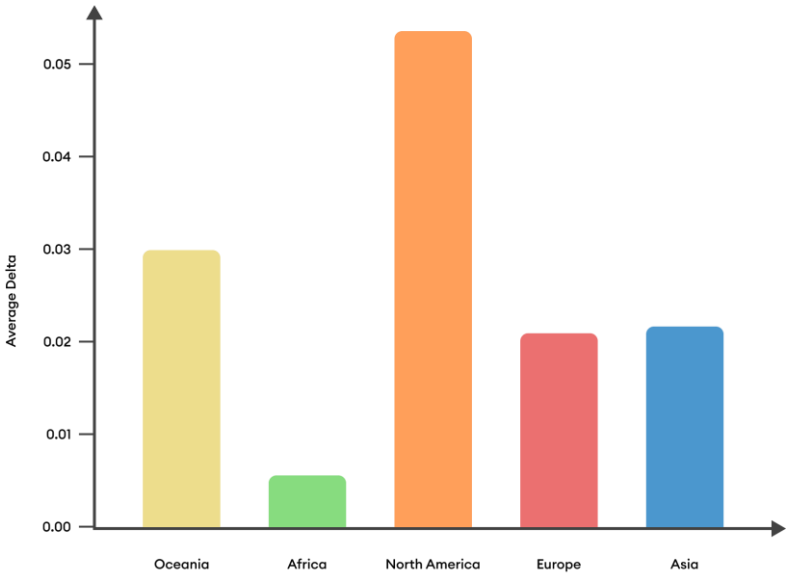

Figure 4. Average Delta for each label

- **Delta-Delta Feature:** The Delta-Delta algorithm, an extension of the Delta algorithm, computes the second-order temporal derivatives of a series of feature vectors. It offers further details on the feature sequence's acceleration or rate of change. The resulting delta-delta coefficients reveal details about the feature sequence's rate of change's acceleration or change over time. Figure 5 depicts the mean delta-delta characteristic for each continent.

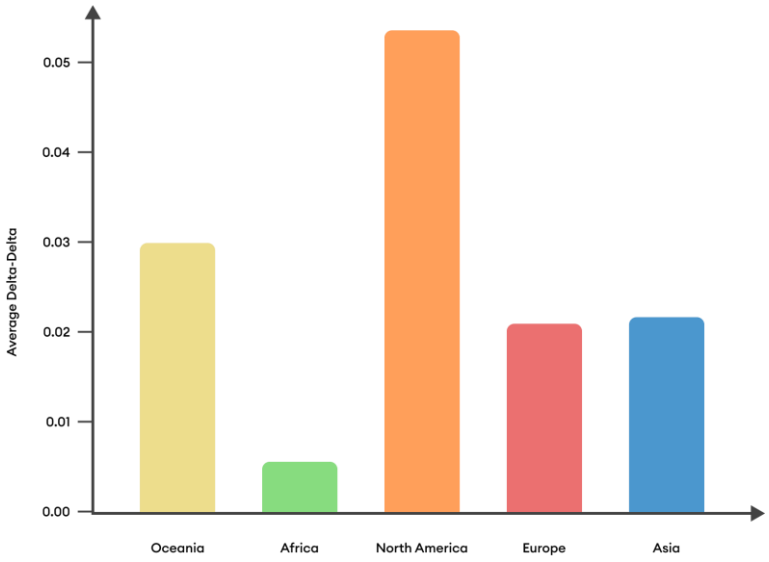

Figure 5. Average Delta-Delta for each label

Delta-delta features record the acceleration, while delta features represent the MFCCs' rate of change. These dynamic features provide temporal context, helping the model detect variations in speech flow and transitions unique to different accents.

- **Chroma Frequency:** This function takes care of analyzing the harmonic structure and timbre attributes of the incoming audio stream. For the purpose of reducing spectral leakage, the audio is split into frames, and every frame has a window function applied to it. The Short-Time Fourier Transform (STFT) is used to obtain the magnitude spectrum, which contains the spectral content of the given frame. The filter bank maps the frequencies onto chroma bands. After that, Chroma energies are normalized, such that their sum is one, to facilitate comparison across different audio streams. Optional post-processing techniques might be applied for

stability enhancement and reduction of temporal variability, such as smoothing or temporal averaging. Chroma features inspire chromatic mapping of audio frequency for speech harmonic structure, emphasizing tonal differences and aiding in recognizing regional speech characteristics. Figure 6 displays the average chroma frequency for each continent.

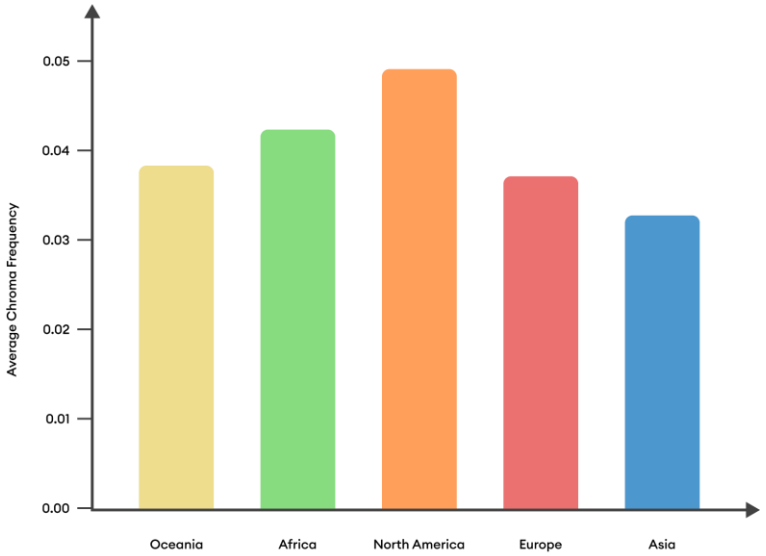

Figure 6. Average Chroma Frequency for each label

- **Zero Crossing Rate:** A measurement of the speed at which an audio stream passes the zero amplitude level is the Zero Crossing Rate (ZCR) characteristic. The stream is divided into frames for calculating the ZCR, and each frame's sample sign is determined. The ZCR value for each frame is calculated by counting the instances in which the sign switches from positive to negative or negative to positive and dividing that number by the overall number of samples within the frame. Each frame in the audio signal goes through this process once. The audio waveform's quick transitions or changes are described by the ZCR characteristic. This feature helps capture rhythmic and temporal patterns in speech, which vary across accents. Figure 7 displays the average Zero Crossing Rate for each continent.

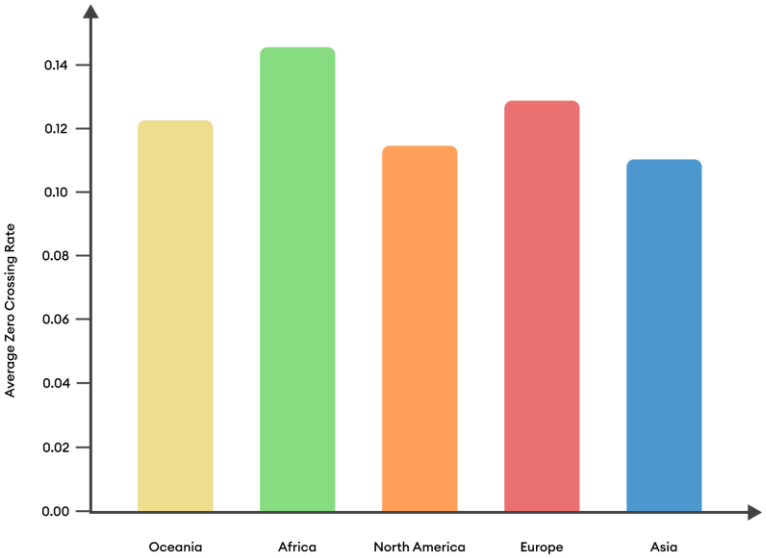

Figure 7. Average Zero Crossing Rate for each label

- **Spectral flux:** This feature is designed to estimate how the spectral content of an audio signal varies over time. It provides information on the dynamics and variability of spectral energy distribution. The spectral flux feature is computed by comparing the magnitude spectrum of successive frames. After segmentation of the audio stream into frames, a windowing function is used on each frame, and the Fourier transform is performed to obtain the magnitude spectrum. Spectral flux is then determined by adding up the squared variations in the magnitude spectra of subsequent frames. This is done for every frame and thus provides the spectral flux values, which indicate how fast the spectral content changes over time. Therefore, the spectral flux measures changes in the

power spectrum and aids in identifying dynamic alterations in spoken language. Figure 8 shows the average Zero Crossing Rate for each continent.

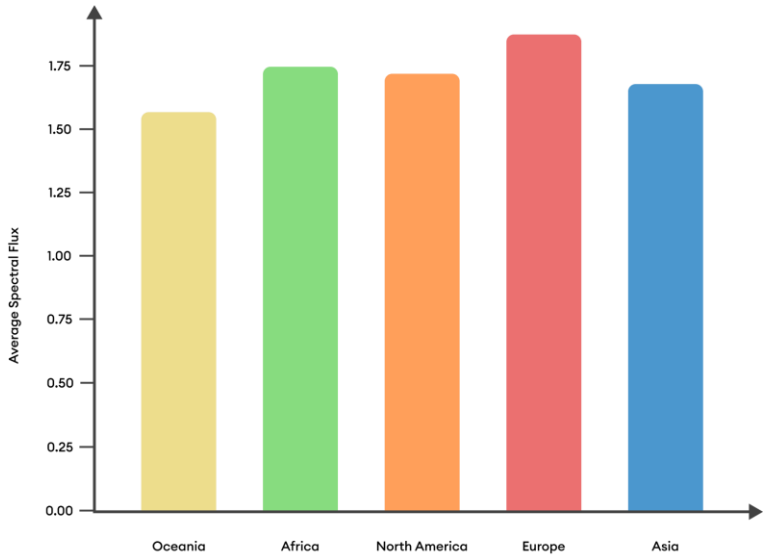

Figure 8. Average Spectral flux for each label

- **Spectral Centroid:** Spectral Centroid is a feature used to extract the average frequency or the gravity center of the spectrum of an audio stream. It gives information about the most favored frequency in a particular frame or segment of the signal. For the computation of the Spectral Centroid, the audio signal is first framed, followed by applying a window function to each frame and computing the Fourier transform, yielding the magnitude spectrum. The spectrum's weighted frequencies are averaged, with the weighting being the magnitude of the respective frequency bins, to yield the spectral centroid for each frame. This produces a series of Spectral Centroid values that portray the frequency center for each frame that attempts to capture frequency characteristics. Figure 9 shows the average Zero Crossing Rate for each continent.

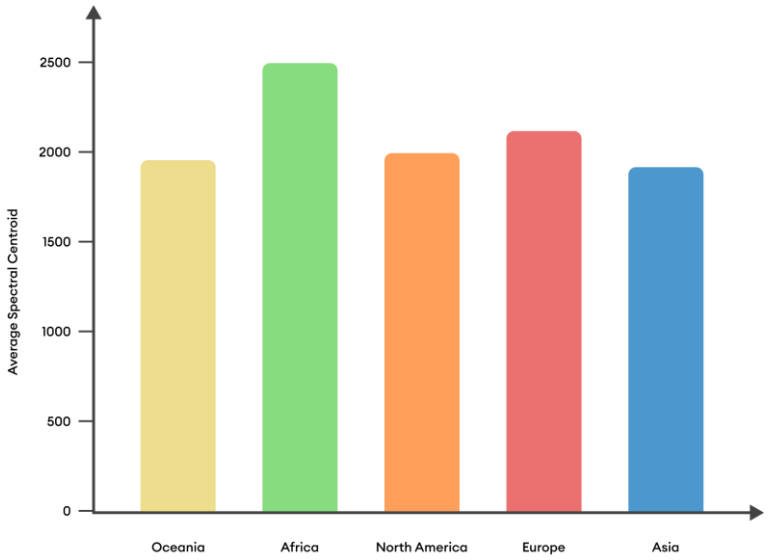

Figure 9. Average Spectral Centroid for each label

- **Spectral Bandwidth:** The Spectral Bandwidth function can be defined as the means by which to calculate the width or spread of different frequencies in the spectrum of an audio signal. It gives information about the dispersal of spectral energy around the spectral centroid. The magnitude spectrum is obtained by Fourier-transforming a windowed function to each frame of the audio stream to compute the Spectral Bandwidth. The spectral bandwidth is calculated per frame by adding the weighted squared deviations of each frequency bin from the spectral centroid, with the weights being the magnitudes of the frequency bins in question. The repetitions give rise to a string of spectral bandwidth values, which represent the dispersion or width of frequencies surrounding the

spectral centroid within each frame. It says the width of the frequency spectra, which gives information about the sharpness of the sound. Figure 10 shows the average spectral bandwidth for each continent.

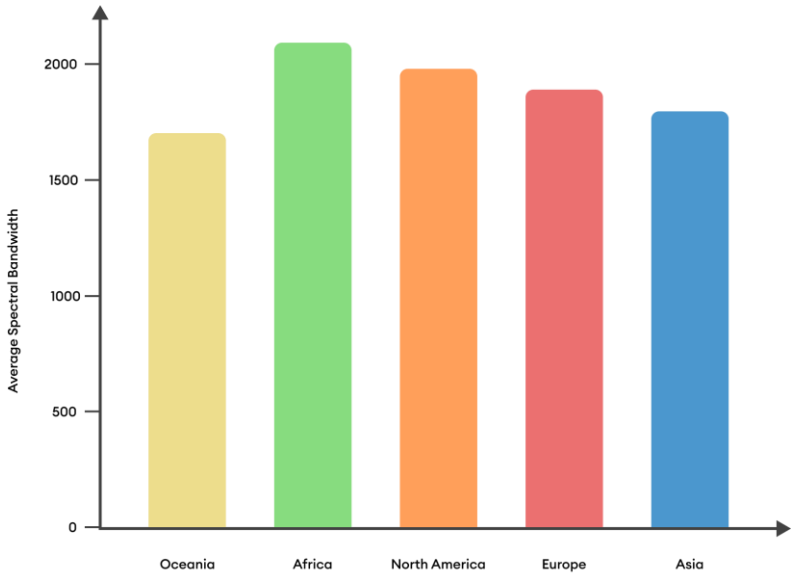

Figure 10. Average Spectral Bandwidth for each label

- **RMS:** The amplitude and energy of an audio signal can be computed using the RMS feature; this indicates the average signal power level. In calculating the root mean square feature, the signal is framed into slices, and each slice is applied to a window function. The windowed samples, or the samples contained in the sliced frame, have their values squared based on that window function. The RMS value for each frame is said to be computed by averaging the squared samples and taking the square root of that average. The above steps are repeated until a sequence of RMS values is obtained that would reflect the energy of the signal over time. The Root Mean Square energy gives information on expressive strength-the measure of the intensity of speech signals that vary due to accent differences locations. Figure 11 displays the mean RMS value for each continent.

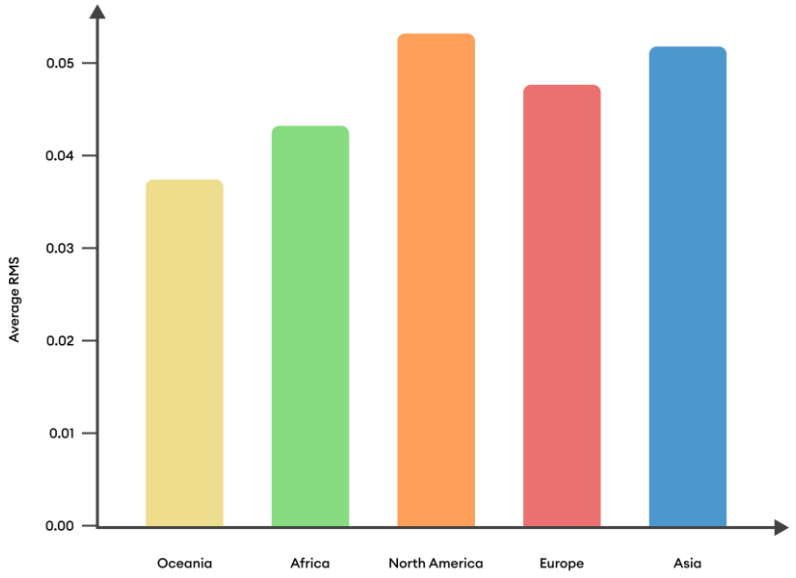

Figure 11. Average RMS for each label

- **Pitch:** The pitch feature calculates the perceived pitch or fundamental frequency of an audio signal. It provides information about the pitch or melody of the sound. To calculate the pitch feature, audio streams are segmented into frames, a windowing function is used on each frame, and the pitch is then estimated using pitch estimation methods such as autocorrelation or cepstral analysis. The methods usually analyze the periodicity and/or harmonic structure of the signal to assess the fundamental frequency. The fundamental frequency of speech contribute to the realization of intonation and melody, which are crucial in the distinguishing of accents. Figure 12 displays the mean pitch value for each continent.

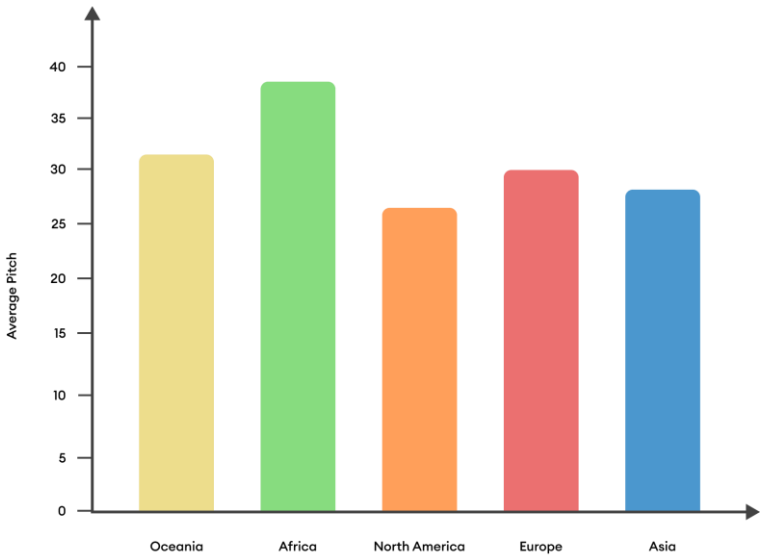

Figure 12. Average Pitch for each label

- **Line Spectral Frequencies:** Line Spectral Frequencies (LSF), more commonly referred to as Line Spectral Pairs (LSP), is a property that characterizes the spectral envelope of an audio source of different audio sources. Line Spectral Frequencies offer information about the formants or the resonant frequencies contained in the signal. The method begins by segmenting the audio signal into frames, followed by application of a window function to each frame, and finally by estimation of the coefficients of the vocal tract filter through Linear Predictive Coding analysis, thus yielding the LSF. The LSFs so computed are exploited for Speech Synthesis, Speaker Identification, and Speech Recognition applications as they stand for the formant frequencies of each frame. The LSF features capture the unique properties of the vocal tract while also elucidating the spectrum characteristics of the audio stream. LSFs describe the spectral envelope of the speech signal, pinpointing formant frequencies that are distinct in different accents. The mean Line Spectral Frequencies value for each continent is displayed in Figure 13.

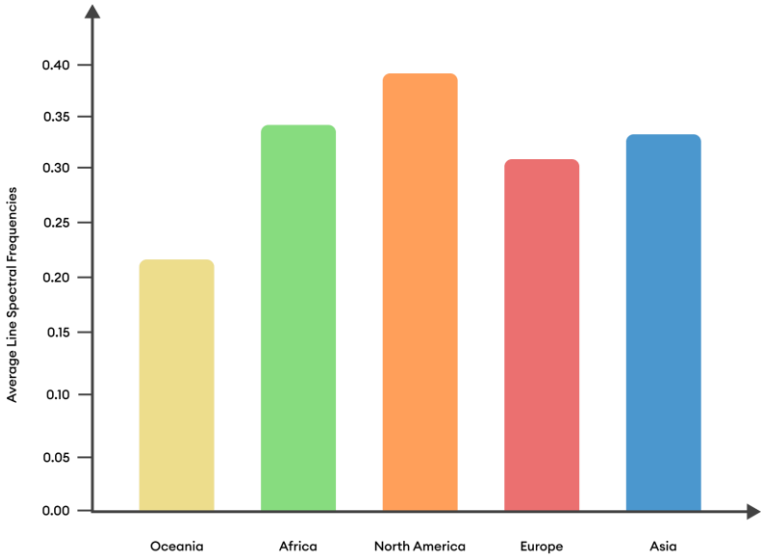

Figure 13. Average Line Spectral Frequencies for each label

3-4- Relevance of Extracted Features

In Summary, each of these features contributes uniquely to the model's ability to classify accents:

- **MFCCs, Delta, and Delta-Delta Features:** Used for static and dynamic spectral information, providing a robust representation of speech signals.
- **Chroma and Pitch Features:** Highlight tonal and melodic aspects of speech that vary across accents.
- **Spectral Features and ZCR:** Reflect rhythm, clarity, and other acoustic patterns tied to regional speech

By combining these features, the CNN model gains a comprehensive understanding of both the spectral and temporal characteristics of speech, enabling effective accent classification.

3-5-Accent Classification Model

Proposed Continent-wise Accent Classification used fully connected Convolutional layer, max pooling layer, fatten layer Dense layer for classifying Continent-wise accent. Additionally, this model used a few regularization techniques, such as dropout [40] and batch normalization [41].

Continent-wise Accent Classification Model

1 : ADAM (learning rate)
2 : For 964 iterations in all batch do:
3 : Convolution 1 (Filter, Kernel Size, Stride, Padding, Activation)
4 : MaxPooling 1 (Pool Size, Stride)
5 : Dropout (Rate)
6 : Convolution 2 (Filter, Kernel Size, Stride, Padding, Activation)
7 : MaxPooling 2 (Pool Size, Stride)
8 : Dropout (Rate)
9 : Convolution 3 (Filter, Kernel Size, Stride, Padding, Activation)
10 : MaxPooling 3 (Pool Size, Stride)
11 : Dropout (Rate)
12 : Convolution 4 (Filter, Kernel Size, Stride, Padding, Activation)
13 : MaxPooling 4 (Pool Size, Stride)
14 : Dropout (Rate)
15 : Dense (Units, Activation)
16 : Dropout (Rate)
17 : Dense (Units, Activation)
18 : Dropout (Rate)
19 : Dense (Units, Activation)
20 : Dropout (Rate)
21 : Dense (Units, Activation)
22 : Dropout (Rate)
23 : End for

The first layer of convolution has 64 convo1D filters each of which has a (3 × 3) size kernel, a (1 × 1) size stride, and an ReLu (4) activation function. Followed by a MaxPooling layer, and it has (2 × 2) size pool size and (2 × 2) size stride. After that, a dropout layer with a 20% dropout rate.

$$\text{ReLU}(Y) = \text{MAX}(0, Y) \quad (4)$$

The second layer of convolution has 128 convo1D filters, each of which has a (3 × 3) size kernel, a (1 × 1) size stride, and an ReLu (4) activation function. Followed by a MaxPooling layer, and it has (2 × 2) size pool size and (2 × 2) size stride. In the following step, another dropout layer with a 20% dropout rate was added.

The third layer of convolution has 256 convo1D filters, each of which has a (3 × 3) size kernel, a (1 × 1) size stride, and an ReLu (4) activation function. Followed by a MaxPooling layer, and it has (2 × 2) size pool size and (2 × 2) size stride. In the following step, another dropout layer with a 20% dropout rate was added again.

The fourth layer of convolution has 128 convo1D filters, each of which has a (3 × 3) size kernel, a (1 × 1) size stride, and an ReLu (4) activation function. Followed by a MaxPooling layer, and it has (2 × 2) size pool size and (2 × 2) size stride. Again a dropout layer with a 20% dropout rate was added.

Next, apply 150 unit's dense layer with ReLu activation, and 25% dropout after flattening the layer. With ReLu activation and a 25% dropout, the output of this layer connects to the 100 units Dense Layer. The Dense layer was then coupled to the output of this layer through 20 units of ReLu activation and 25% dropout. Utilize 5 units with SoftMax (5) activation at the final output layer.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_j}} \quad for \quad j = 1, \dots, k \quad (5)$$

Here, σ(z)j is Softmax function output for the j-th class, e is Euler's number (approx. 2.718), base of natural logarithm, z_j is Input value (logit) for the j-th class, Σ is Summation symbol (sum of all exponentiated values), k is total number of classes, j is index representing a specific class, and z_k is input value (logit) for the k-th class.

In Summary,

- **Input Layer:** Accepts input of size (56, 64), corresponding to processed speech features.
- **Convolutional Layers:**
  - **Layer 1:** 64 filters with (3 × 3) kernel, (1 × 1) stride, and ReLU activation. Followed by MaxPooling with (2 × 2) pool size and (2 × 2) stride, and 20% dropout.
  - **Layer 2:** 128 filters, (3 x 3) kernel, (1 × 1) stride, ReLU activation, MaxPooling, and 20% dropout.
  - **Layer 3:** 256 filters, (3 x 3) kernel, (1 × 1) stride, ReLU activation, MaxPooling, and 20% dropout.
  - **Layer 4:** 128 filters, (3 x 3) kernel, (1 × 1) stride, ReLU activation, MaxPooling, and 20% dropout.
- **Fully Connected Layers**:
  - Flattened output connected to a 150 unit dense layer, ReLU activation, and a 25% dropout rate.
  - Sequential dense layers with 100, 50, and 20 neurons, each using ReLU activation and 25% dropout.
  - Final layer with 5 neurons and SoftMax activation for classification.

Architecture of Continent-wise Accent Classification Model is displayed in Figure 14.

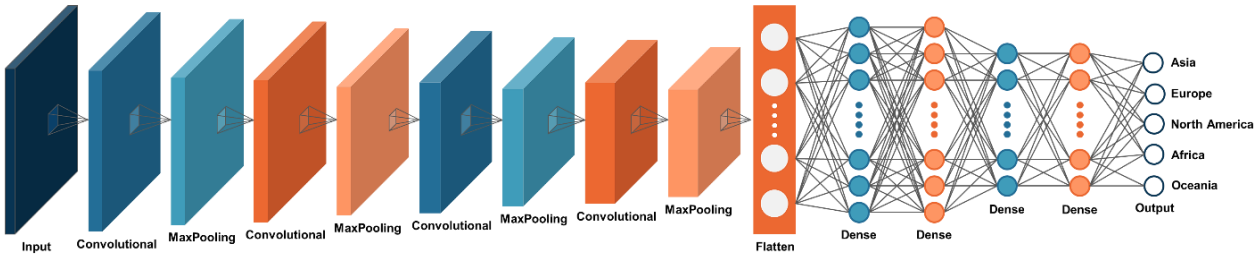

Figure 14. Architecture of Continent-wise Accent Classification Model

The architecture is designed to capture intricate speech patterns and variations in accents across continents:

- **Convolutional Layers:** Extract local features, such as spectral characteristics.
- **Dropout:** Prevent overfitting by randomly deactivating nodes during training.
- **Dense Layers**: Combine learned features for classification.

The choice of ReLU activation ensures efficient training by addressing the vanishing gradient problem. MaxPooling layers downsample feature maps, reducing computational load while preserving essential features. This architecture was finalized after experimenting with different configurations and achieved optimal performance on validation data.

The model can be accurately identified with the support of the model architecture overview. A simplified overview of the full model can be seen in Table 3.

Table 3. Continent-wise Accent Classification Model Architecture summary

Layer	Type	Output Shape	Parameter	Connected to
(Input Layer)	Conv1D	(None, 56, 64)	256	-
1()	MaxPooling1D	(None, 28, 64)	0	Input Layer
2()	Dropout	(None, 28, 64)	0	1
3()	Convo1D	(None, 26, 64)	12352	2
4()	MaxPooing1D	(None, 13, 64)	0	3
5()	Dropout	(None, 13, 64)	0	4
6()	Convo1D	(None, 11, 128)	24704	5
7()	MaxPooing1D	(None, 5, 128)	0	6
8()	Dropout	(None, 5, 128)	0	7
9()	Convo1D	(None, 2, 256)	98560	8

10()	MaxPooing1D	(None, 1, 256)	0	9
11()	Dropout	(None, 1,256)	0	10
12()	Flatten	(None, 256)	0	11
13()	Dense	(None, 128)	32896	12
14()	Dropout	(None, 128)	0	13
15()	Dense	(None, 256)	33024	14
16()	Dropout	(None, 256)	0	15
17()	Dense	(None, 128)	32896	16
18()	Dropout	(None, 128)	0	17
19()	Dense	(None, 64)	8256	18
20()	Dropout	(None, 64)	0	19
21()	Dense	(None, 5)	325	20
Total params:			**243,269**	
Trainable params:			**243,269**	
Non-trainable params:			**0**	

3-6- Optimizer and Learning Rate

One of the key components for training a neural network model is optimizer. It is responsible for adjusting the weights and biases of the model during the training process to minimize the loss function. Proposed Continent-wise Classification model used the very popular Optimizer Adam Optimizer [42]. The stochastic gradient descent algorithm has been modified by this optimizer. Network weight is being updated by an optimizer that can adjust hyper-parameters. Adam optimizer (6) was employed in the proposed continent-wise accent classification, and its learning rate was set at 0.001.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt[1]{\hat{v}_t}+\varepsilon}\hat{m}_t \quad (6)$$

Here, t is Time step (iteration index), η is step size, $\hat{m}_t$ is mean of gradients, $\hat{v}_t$ is variance of gradients, and $\mathcal{E}$ is Small constant for numerical stability.

A function known as categorical cross entropy (7) was used to calculate model error [43].

$$Li = -\sum J\, Ti,j \mid \log(Pi,j) \quad (7)$$

Here, Σ_J is Summation over all classes J, $T_{i,j}$ is true label (1 if correct class, 0 otherwise), $P_{i,j}$ is Predicted probability for class j, and $\log(P_{i,j})$ is Logarithm of the predicted probability.

Learning rate is a hyperparameter that determines the step size or rate at which a machine learning model adjusts its parameters during training. It regulates how much the model's weights and biases are modified in response to the determined gradients. Larger updates are possible with a greater learning rate. It may lead to faster convergence but runs the danger of going beyond what would be considered the ideal answer. Conversely, a smaller learning rate leads to smaller updates, which may slow down convergence but could potentially improve the model's accuracy [43]. A greater learning rate of 0.001 was initially stated, however it was changed in response to the validation accuracy.

3-7- Training the Model

The proposed CNN model was trained using Mozilla Common Voice dataset with a batch size of 56, achieving an accuracy of 82% after 50 epochs.

4- Results and Discussion

The final CNN model demonstrated strong performance on train, test, and validation sets. With a total of 26396 voices from five different continents, the model was trained using 80% of the voice data then tested with 10%, and validated with 10%. The accuracy and loss of the proposed continent-wise accent classification model are shown in Figure 15.

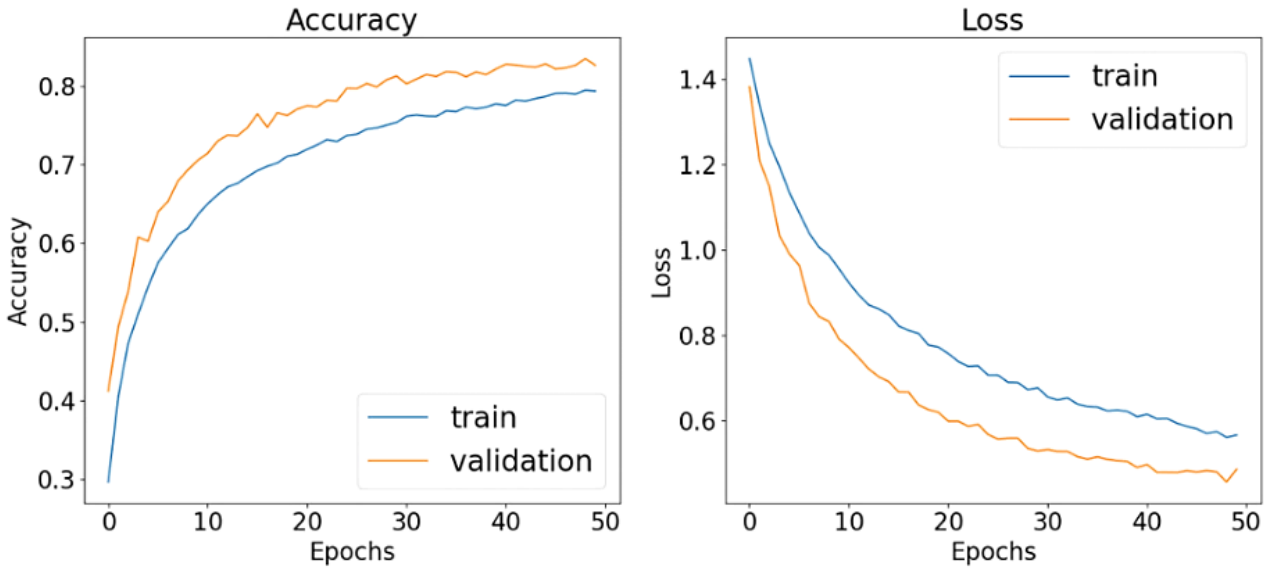

Figure 15. Training and Validation Accuracy and Loss of Continent-wise accent classification Model

After 50 iterations, the model's accuracy was 84.79% on the training set and 82.34% on the validation datasets.

Table 4. Summary of the performance metrics for each continent

Continent	Accuracy	Precision	Recall	F1-Score
Asia	85%	84%	86%	85%
Europe	83%	82%	83%	82%
North America	80%	79%	80%	79%
Africa	76%	75%	77%	76%
Oceania	74%	73%	75%	74%

Summary of the accuracy of all the classes is shown in Table 4. That indicates an overall accuracy of 82%, the model recorded significant variations in performance among the five continents. Asia (85%) and Europe (83%) secured the highest classification accuracy as these two continents had bigger and more balanced datasets that allowed them to generalize better. On the contrary, Africa (76%) and Oceania (74%) had the lowest classifications because of the limited data representation and higher linguistic diversity within these two continents that contribute to more errors in classification. North America (80%) has a fair performance but has common languages with the European accent leading to some mistaken identifications within the two. The dependence of the model on spectral as well as on phonetic features enabled cross-continental wise effective classification, but intra continental linguistic diversity and dataset imbalances highlighted differences in performance. to further improve the model's ability to separate the accents within the continent while improving overall accuracy, data representation, refinements in extracting features, and hierarchical classification are some ways forward.

According to our model, continental-level classification takes priority over the regional classification of dialects or sub-accents within a continent. Thus, while it is able to categorize accents in broad continental groups with an accuracy of about 85%, it does not make an explicit distinction between finer accent variations like the ones between British English versus Australian English or North American versus Caribbean English. The CNN-based architecture captures phonetic and spectral patterns common to accents within a continent, but due to limitations of the dataset and universal linguistic reasons, specific local accents may get put into one category. CFs like MFCCs with Delta and Chroma Frequency might help in identifying pronunciation niceties, but better enhancements or just fine-tuning with region labels and hierarchical classification could allow for more distinguishing power among sub-accents. In the future, multi-level classification or other methods like self-supervised learning could help in the model's further refinement toward appreciating intra-continental accent variations while retaining accuracy in the broad distinction. Further analysis of the model is presented in the confusion matrix shown in Figure 16.

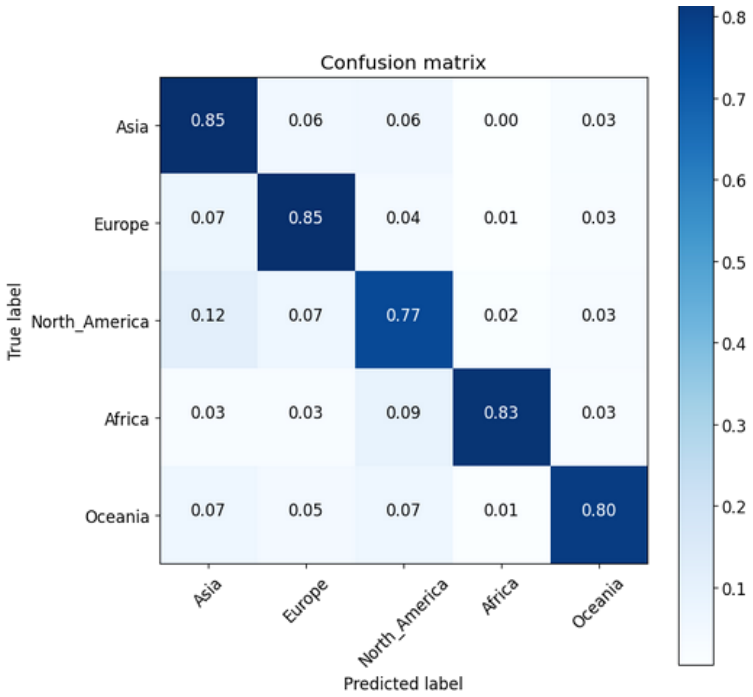

Figure 16. Confusion matrix of Continent-wise accent classification model

The evaluation of the confusion matrix illuminates the model's positive and negative distinguishing powers across five continents in the accent classification space. Among all of them, the highest accuracies were delivered for Asia (85%) and Europe (85%); these were supported by larger and more balanced datasets, though Asia faced a 6% misclassification against Europe and North America, possibly due to alignment in phonetic proximity. On the other hand, North America with 77% accuracy had a little lesser accuracy with 12% classifying falsely as having Asian accents, possibly due to highly contrasting linguistic influences. Africa (83%) showed a 9% chance of misclassification against North America and that was on account of their phonetic similarities, while Oceania (80%) had the lowest possible score, with 7% being misclassified as Asian and 5% as European, owing in great part to its smaller database and the commonality in accents. The overlap between North American and European accents and some overlaps with some African and North American accents present some areas for improvement, including expanding the datasets, adopting hierarchical classification, and refining extraction of phonetic features. If a model could sensibly differentiate accents, it would be able to do so with greater precision if linguistic diversity, imbalances in datasets, and regional variations were considered.

5- Conclusion

A deep learning-based accent classification across continents was introduced in this paper, which made use of convolutional neural networks (CNN) along with spectral feature extraction methods to enhance the accuracy of speech recognition systems. The model classified accents from five continents: Asia, Europe, North America, Africa, and Oceania, with an impressive validation accuracy of 82%, as verified through the use of the Mozilla Common Voice dataset. The results reveal that higher accuracy has been obtained for accents from Asia and Europe due to the existence of larger, well-represented datasets, whilst African and Oceanian accents were more frequently misclassified due to their linguistic diversity combined with lesser dataset availability. The study emphasizes the critical nature of encompassing diverse speech patterns into the ASR systems for maintaining inclusivity and enhancing a worldwide niche. Findings have also helped contribute to a variety of speech-based applications that will tailor themselves to various accents, thus minimizing recognition bias—for instance, voice assistants, speech-to-text systems, and language translation technologies.

This research would shape future avenues beyond ASR performance advancements in understanding speech variations at a continental level in language and cultural contexts. Unlike most works that concentrated on country-specific native vs. non-native English accents, this research takes speech from much broader classification perspectives to improve its technology adaptability to global users. It would include future applications to add all accents around the seven continents to this dataset, continue refinement of feature extraction, and incorporate some of the latest deep learning techniques, such as transformer-based architectures, to improve classification performance. Real-time accent adaptation mechanisms integrated into ASR systems could be of further improvement in their effectiveness for multilingual and cross-accent scenarios. Therefore, this study points out the existing gaps in accent classification research and leads towards global diversity by creating more democratic and inclusive speech processing technologies.

6- Declarations

6-1- Author Contributions

Conceptualization, M.F.H. and M.S.R.; methodology, M.F.H..; software, M.F.H. and A.R.K.; validation, M.F.H., A.R.K. and M.S.R.; formal analysis, M.S.R.; investigation, O.; resources, M.F.H.; data curation, M.F.H.; writing—original draft preparation, M.F.H.; writing—review and editing, A.R.K.; visualization, A.R.K.; supervision, O. and M.S.R.; project administration, M.F.H. and M.S.R.; funding acquisition, O. and M.F.H. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding

This research was funded by the Institute for Advanced Research Publication Grant of United International University, Ref. No.: IAR-2025-Pub-102.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language learning, 49, 285-310. doi:10.1111/j.1467-1770.1995.tb00963.x

[2] Esquivel, P., Gill, K., Goldberg, M., Sundaram, S. A., Morris, L., & Ding, D. (2024). Voice Assistant Utilization among the Disability Community for Independent Living: A Rapid Review of Recent Evidence. Human Behavior and Emerging Technologies, 2024(1), 6494944. doi:10.1155/2024/6494944.

[3] Jayne, C., Chang, V., Bailey, J., & Xu, Q. A. (2022). Automatic Accent and Gender Recognition of Regional UK Speakers. Communications in Computer and Information Science, 1600 CCIS, 67–80. doi:10.1007/978-3-031-08223-8_6.

[4] Rizwan, M., & Anderson, D. V. (2018). A weighted accent classification using multiple words. Neurocomputing, 277, 120–128. doi:10.1016/j.neucom.2017.01.116.

[5] Huang, C., Chen, T., Li, S., Chang, E., & Zhou, J. (2001). Analysis of speaker variability. EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology, 1377–1380. doi:10.21437/eurospeech.2001-356.

[6] Mak, L., Sheng, A., & Wei Xiong, M. E. (2018). Deep Learning Approach to Accent Classification. CS229, 1–6.

[7] Arslan, L. M., & Hansen, J. H. L. (1996). Language accent classification in American English. Speech Communication, 18(4), 353–367. doi:10.1016/0167-6393(96)00024-6.

[8] Badhon, S. M. S. I., Rahaman, M. H., & Rupon, F. R. (2020). A Machine Learning Approach to Automating Bengali Voice Based Gender Classification. Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019, 55–61. doi:10.1109/SMART46866.2019.9117385.

[9] Bahari, M. H., & Van Hamme, H. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. BioMS 2011 - 2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications, Proceedings, 27–32. doi:10.1109/BIOMS.2011.6052385.

[10] Hansen, J. H. L., Williams, K., & Bořil, H. (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. The Journal of the Acoustical Society of America, 138(2), 1052–1067. doi:10.1121/1.4927554.

[11] Mporas, I., & Ganchev, T. (2009). Estimation of unknown speaker's height from speech. International Journal of Speech Technology, 12(4), 149–160. doi:10.1007/s10772-010-9064-2.

[12] Sreelatha, M. B., Pranathi, M., Reddy, K. M., & Shirisha, J. (2024). A Survey on Telugu Accent Classification and Conversion. International Journal of Advances in Engineering and Management (IJAEM), 6(03), 624. doi:10.35629/5252-0603624629.

[13] Singh, M. K. (2024). A text independent speaker identification system using ANN, RNN, and CNN classification technique. Multimedia Tools and Applications, 83(16), 48105–48117. doi:10.1007/s11042-023-17573-2.

[14] Choueiter, G., Zweig, G., & Nguyen, P. (2008). An empirical study of automatic accent classification. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 4265–4268. doi:10.1109/ICASSP.2008.4518597.

[15] Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., & Yoon, S. Y. (2005). Accent detection and speech recognition for Shanghai-accented Mandarin. 9th European Conference on Speech Communication and Technology, 217–220. doi:10.21437/interspeech.2005-112.

[16] Zhang, L., Zhao, Y., Zhang, P., Yan, K., & Zhang, W. (2015). Chinese accent detection research based on RASTA - PLP algorithm. Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things, ICIT 2015, 31–34. doi:10.1109/ICAIOT.2015.7111531.

[17] Joseph, J., & Upadhya, S. S. (2018). Indian accent detection using dynamic time warping. IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017, 2814–2817. doi:10.1109/ICPCSI.2017.8392233.

[18] Kibria, S., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2020). Acoustic Analysis of the Speakers' Variability for Regional Accent-Affected Pronunciation in Bangladeshi Bangla: A Study on Sylheti Accent. IEEE Access, 8, 35200–35221. doi:10.1109/ACCESS.2020.2974799.

[19] Mannepalli, K., Narahari Sastry, P., & Rajesh, V. (2015). Accent detection of Telugu speech using supra-segmental features. International Journal of Soft Computing, 10(5), 287–292. doi:10.3923/ijscomp.2015.287.292.

[20] Ma, Y., Mp, P., Yaacob, S., Ab, S., & Mokhtar, N. F. (2013). Statistical formant descriptors with linear predictive coefficients for accent classification. Proceedings of the 2013 IEEE 8th Conference on Industrial Electronics and Applications, ICIEA 2013, 906–911. doi:10.1109/ICIEA.2013.6566496.

[21] Danao, G., Torres, J., Tubio, J. V., & Vea, L. (2017). Tagalog regional accent classification in the Philippines. HNICEM 2017 - 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, 2018-January, 1–6. doi:10.1109/HNICEM.2017.8269545.

[22] Hossain, M. F., Hasan, M. M., Ali, H., Sarker, M. R. K. R., & Hassan, M. T. (2020). A machine learning approach to recognize speakers region of the United Kingdom from continuous speech based on accent classification. Proceedings of 2020 11th International Conference on Electrical and Computer Engineering, ICECE 2020, 210–213. doi:10.1109/ICECE51571.2020.9393038.

[23] Pedersen, C., & Diederich, J. (2007). Accent classification using support vector machines. Proceedings - 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007; 1st IEEE/ACIS International Workshop on e-Activity, IWEA 2007, 444–449. doi:10.1109/ICIS.2007.47.

[24] Deshpande, S., Chikkerur, S., & Govindaraju, V. (2005). Accent classification in speech. Proceedings - Fourth IEEE Workshop on Automatic Identification Advanced Technologies, AUTO ID 2005, 139–143. doi:10.1109/AUTOID.2005.10.

[25] Berjon, P., Nag, A., & Dev, S. (2021). Analysis of French phonetic idiosyncrasies for accent recognition. Soft Computing Letters, 3, 100018. doi:10.1016/j.socl.2021.100018.

[26] Lesnichaia, M., Mikhailava, V., Bogach, N., Lezhenin, I., Blake, J., & Pyshkin, E. (2022). Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 3669–3673. doi:10.21437/Interspeech.2022-462.

[27] Kashif, K., Alwan, A., Wu, Y., De Nardis, L., & Di Benedetto, M. G. (2024). MKELM based multi-classification model for foreign accent identification. Heliyon, 10(16), e36460. doi:10.1016/j.heliyon.2024.e36460.

[28] Zhang, Z., Wang, Y., & Yang, J. (2021). Accent Recognition with Hybrid Phonetic Features. Sensors (Basel, Switzerland), 21(18), 6258. doi:10.3390/s21186258.

[29] Zuluaga-Gomez, J., Ahmed, S., Visockas, D., & Subakan, C. (2023). CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 5291–5295. doi:10.21437/Interspeech.2023-2419.

[30] Qian, Y., Gong, X., & Huang, H. (2022). Layer-Wise Fast Adaptation for End-to-End Multi-Accent Speech Recognition. IEEE/ACM Transactions on Audio Speech and Language Processing, 30, 2842–2853. doi:10.1109/TASLP.2022.3198546.

[31] Song, T., Nguyen, L. T. H., & Ta, T. V. (2025). MPSA-DenseNet: A novel deep learning model for English accent classification. Computer Speech and Language, 89, 101676. doi:10.1016/j.csl.2024.101676.

[32] G, P. D., & Rao, K. S. (2023). Accent classification from an emotional speech in clean and noisy environments. Multimedia Tools and Applications, 82(3), 3485–3508. doi:10.1007/s11042-022-13236-w.

[33] Demirsahin, I., Kjartansson, O., Gutkin, A., & Rivera, C. (2020). Opensource multispeaker corpora of the english accents in the british isles. LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 6532–6541.

[34] Singh, Y., Pillay, A., & Jembere, E. (2020). Features of speech audio for accent recognition. 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, IcABCD 2020 - Proceedings, 1–6. doi:10.1109/icABCD49160.2020.9183893.

[35] Upadhyay, R., & Lui, S. (2018). Foreign English Accent Classification Using Deep Belief Networks. Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018, 2018-January, 290–293. doi:10.1109/ICSC.2018.00053.

[36] Mozilla. (2021). Mozilla Common Voice. Common Voice. Available online: https://commonvoice.mozilla.org/id/about (accessed on December 2025).

[37] Haton, J. P. (2003). Automatic speech recognition: A Review. ICEIS 2003 - Proceedings of the 5th International Conference on Enterprise Information Systems, 1, IS5–IS10. doi:10.5120/9722-4190.

[38] Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. October, 1(3), 191–194. doi:10.1.1.75.8303.

[39] Furui, S. (1981). Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(3), 342–350. doi:10.1109/TASSP.1981.1163605.

[40] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

[41] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In 32nd International Conference on Machine Learning, ICML 2015, 448–456.

[42] Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, United States.

[43] Schaul, T., Zhang, S., & LeCun, Y. (2013). No more pesky learning rates. 30th International Conference on Machine Learning, ICML 2013, PART 2, 1380–1388.