Automatic Speech Recognition Across Languages: A Study Using Spoken Wikipedia Corpora

Pavan Kumar Chintala

GITHUB

Abstract

This project deals with using machine learning techniques in order to enhance the classification and transcription of spoken language from SWC. It is planned to create a model which, by taking spoken Wikipedia articles, is able to produce an accurate transcription of speech in text format and contribute to improving ASR for different languages. A Random Forest classifier will be used to classify languages with speech features that have been extracted from the audio. By leveraging the SWC containing hundreds of hours of aligned audio and text, the project develops higher transcription accuracy and more proficient multilingual speech processing. This could be in strong demand for research related to ASR and NLP; at the same time, that could further advance multilingual speech technology.

1 Introduction

The automatic speech recognition system is important for smooth interaction between humans and machines through the conversion of spoken language into text. Virtual assistants, real-time transcription, and multilingual translation are some of the key areas of applications. However, the linguistic diversity, background noise, and variation in speech patterns across different languages have often been challenging to the accuracy of ASR systems.

This paper examines the performance of ASR across multiple languages by utilizing the multilingual dataset of audio recordings with their textual transcription correspondences-the Spoken Wikipedia Corpora. Recordings are produced in English, German, and Dutch, thus offering a wide variability of linguistic characteristics and acoustic conditions. In this respect, these recordings provide a unique opportunity to appraise the contribution of several steps in the process chain, like data preprocessing, denoising, and feature extraction, when building a robust ASR pipeline.

In this paper, a Random Forest model has been selected for its versatility and capability of handling high-dimensional data as the main classifier. Advanced denoising techniques like spectral subtraction and spectral gating have been employed to reduce background noise while preserving the critical speech characteristics. Features extracted are Mel-Frequency Cepstral Coefficients for capturing essential audio properties in classification, Chroma, and Spectral Contrast.

This work highlights the use of preprocessing, noise reduction, and feature selection in improving ASR performance across multilingual datasets. Based on a machine learning pipeline with Random Forest as the base, this research tries to provide insight into how challenges presented by noisy and linguistically diverse datasets can be handled.

2 Dataset Description

This paper is based on the Spoken Wikipedia Corpora dataset, a rich and diverse collection of audio recordings and their corresponding textual transcription in various languages. This dataset has been selected due to its multilingual nature and for having realistic audio conditions that are relevant in the testing of ASR systems.

2.1 Dataset Composition

The Spoken Wikipedia Corpora is a resource multilingually designed, having audio recordings in three languages: English, German, and Dutch. These represent diverse linguistic phenomena, which turn out to be very important for analyses related to language-specific speech traits. Each file is 30–40 minutes long; the speech is natural and has been recorded in different acoustic settings. This variability includes different levels of background noise, from studio-like conditions with minimal noise to recordings in ambient environments with significant noise interference. Coupled with each audio file is a corresponding text transcription that serves as the ground truth for evaluating the performance of automatic speech recognition systems.

2.2 Challenges

One major challenge presented by the Spoken Wikipedia Corpora dataset makes it a very attractive benchmark for research in ASR. The key challenge with the dataset has to do with variability in the noise level across recordings. The dataset is made of audio files recorded under various conditions, therefore, with very varying background noise levels. The diversity necessarily demands the effective employment of denoising techniques that better the performance of ASR. Further, added linguistic complexity in the data makes things even more complex: differences in pronunciation, regional accents, and language-specific features such as syntax and phonetics make multilingual audio processing very challenging. Another challenge is the imbalance in data availability across three languages. Variation in data quantity between English, German, and Dutch can affect model training and potentially skew performance toward those languages for which more data exists.

2.3 Data Preprocessing

The dataset then underwent various preprocessing steps to make it suitable for machine learning tasks. Every long audio file was first divided into its small portions. This serves to make feature extraction and processing more feasible while bringing down the computational overhead. Secondly, analysis of noise was done on this data owing to its rather noisy nature. Silent segments and low-energy portions of audio files were analyzed to estimate the noise profiles. These profiles were then used to implement various denoising techniques with the purpose of enhancing clarity and quality in speech data.

Indeed, it is a multilingual and acoustically diverse dataset that represents the real scenario of benchmarking ASR systems. This study will develop robust methodologies in data preprocessing, feature extraction, and classification with the intention of overcoming the challenges brought forth by noise variability, linguistic complexity, and imbalance in data for advancing the state of ASR technology.

3 Data Denoising

3.1 Denoising Techniques

In order to reduce the noise in the Spoken Wikipedia Corpora dataset, three established techniques were used: spectral subtraction, Wiener filtering, and spectral gating. The spectral subtraction entailed estimating noise from silent portions of the audio and subtracting it from the signal to minimize background interference. Wiener filtering is an adaptive method that uses the estimated noise profile to suppress noise while maintaining the key features of the original signal. Finally, spectral gating iteratively applied thresholding to low-energy spectral components, which dynamically suppressed residual noise while maintaining speech clarity.

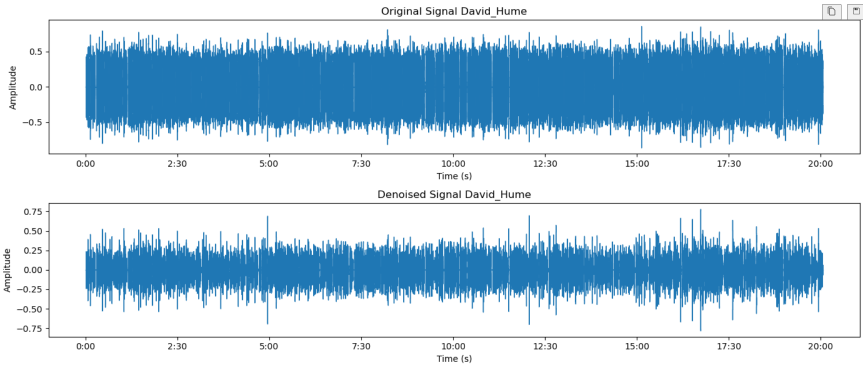

Figure 1: Difference between audios before and after denoising

3.2 Process

Noise profile estimation involves seeking the low-energy parts of the audio to model the background noise and is the initialization step of the denoising process. Noise levels are reduced after that by iteratively applying spectral subtraction and Wiener filtering using that profile. The last step is spectral gating, further suppressing noise which has been left behind. The outputs of this process consisted of two data sets: original features, which were directly extracted from raw audio files; and denoised features derived after the application of noise reduction techniques.

3.3 Signal-to-Noise Ratio (SNR) Evaluation

The performance of the denoising is measured by the Signal-to-Noise Ratio, one of the key figures of merit in clarity of speech or audio. The global improvement in SNR was + 11.79 dB across the dataset. The various range of improvements starts from around only 0.46 dB in the files that, per se, had really very low levels of noise. For example, an audio 1065.ogg had gone from 12.03 dB to 12.49 dB. In contrast, high-noise recordings enjoyed significant gains as high as 25.39 dB, such as audio 536.ogg, improving from 10.69 dB to 36.09 dB. This made the realization that files with higher initial noise levels benefited much while those with low noise developed minimal improvements.

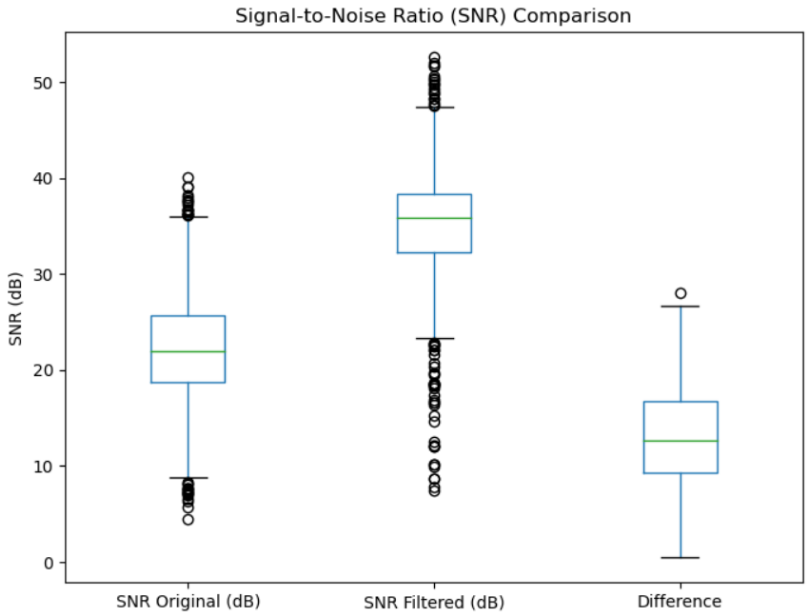

Figure 2: Signal to Noise Ratio Comparision

3.4 Impact of Denoising

The significant cleaning through denoising brought much clarity to the audio, making it far better for feature extraction and classification tasks. Spectral gating further refined the quality of the audio by suppressing low-energy noise components. More importantly, these methods have preserved core signal features, thus being prepared for downstream analysis. Therefore, the denoised dataset provided a cleaner and more robust foundation for machine learning models, especially those requiring high-quality signals.

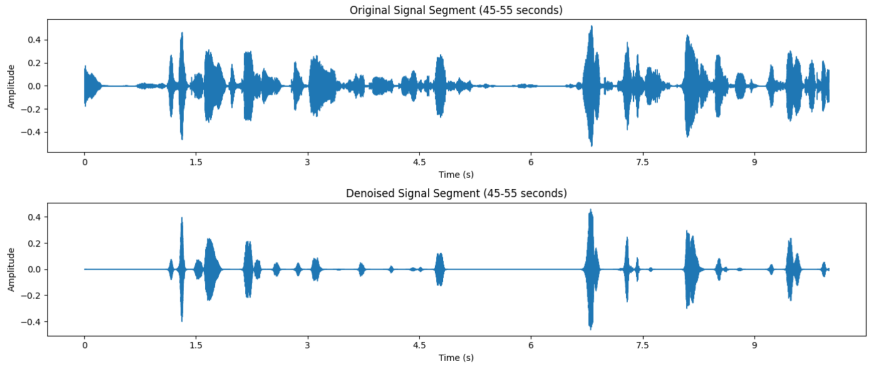

Figure 3: changes in audio after denoising (Smaller segment)

3.5 Challenges

In fact, some critical challenges were faced with achieving substantial improvements through this process of denoising. Techniques benefited certain files more than others-those with high recordings and those with minimal recording of noise. Moreover, an unintended filtering of subtlety in the audio could not be ruled out with a reduction in noise. Thus, fine-tuning in this regard was quite required in striking the right balance between good denoising and retaining features. Despite these challenges, the process successfully enhanced the dataset's readiness for machine learning applications.

4 Feature Selection

One of the most important procedures for developing effective ASR systems is feature selection, which is a process through which relevant features from audio data are identified and chosen for contribution to accuracy in classification at lower computational complexity. In this work, feature selection was based on how well the features capture essential characteristics of speech signals.

4.1 Selected Features

The following features are selected as they are some of the powerful speech processing and classification techniques:

Mel-Frequency Cepstral Coefficients (MFCC): The MFCC describes the audio signal's spectral properties compactly. That is also critical information on timbre and tonal characteristics that are required in speech. Thus, this research study utilizes 13 MFCCs for every segment of audio to have as much detailed representation of the speech signal as possible.

Chroma Features: Chroma features reflect the distribution of energy in the 12 pitch classes, for instance, C, D; therefore, this feature has been more related to the representation of harmonic and tonal information of speech signals. Chroma adds a music point of view for speech processing to help further in discriminating between tonals.

Spectral Contrast: It measures the peaks and valleys of the frequency spectrum. The feature has widely been used in deciding voiced and unvoiced segments, speech segments, and the identification of acoustic environment variations.

4.2 Feature Selection Rationale

Some selected features represented the main properties of speech that were related to pitch, energy, and spectral characteristics. This was where these features optimized performance with the Random Forest Classifier due to the fact that a very good balance was maintained between being highly informative with lower computational loads. With minimum irrelevant and redundant information, this approach led to much better model accuracy with better interpretability.

4.3 Preprocessing for Feature Selection

Preprocessing steps were done to ensure the quality and consistency of the selected features:

Normalization: Feature values were normalized using `StandardScaler` to make scaling consistent across audio segments. This step helps reduce the impact of different magnitude ranges on the machine learning model.

Segmentation: First of all, large audio files were divided into smaller segments. All the different segments were analyzed, after which their features had to be computed. That offers a double advantage, showcasing the fine details in a speech and reducing the computing load.

Noise Reduction: Noise reduction involved the computation of features ahead of Spectral Subtraction and Spectral Gating for denoising speeches. Both methods diminish the influence of background noise in increasing extracted feature clarity and reliability.

4.4 Data Organization

The following are strategies taken in to better organize the dataset for machine learning tasks:

Feature Combination: The chosen features were combined into one feature vector for each audio file in such a way that the spectral and tonal characteristics get well represented in the dataset.

Original and Denoised Datasets: There are two different data sets created

- *Original Features:* Extracted directly from the raw audio files with no noise reduction.
- *Denoised Features:* Features extracted after noise reduction.

This will help ensure that only high-quality, informative data concerning the core properties of the speech signal are fed into the machine learning model. The selected features will have impacts in which pre-processing builds up or diminishes from one ASR language to other languages.

5 Feature Computation

5.1 Overview

Feature computation involves the extraction of the numerical representation of audio data for the capture of spectral, tonal, and temporal characteristics. These features serve as an input to machine learning models for effective classification. In this project, feature computation was done for both the original and denoised datasets.

5.2 Computation Process

The following are the steps taken in the computational approach toward the preparation and extraction of features from audio data:

Preprocessing: This step normalizes the audio signals to a similar amplitude level for all. Overlapping framing was used to allow short-time analysis, hence catching the dynamic properties of the audio signals, preserving continuity through time.

Feature Extraction Tools: Feature extraction was carried out using the Python library `Librosa`. These are all varied in various parameters like frame size and hop length, in order to

effect a trade-off between resolution and computational efficiency; this way, features captured held enough details with reduced requirements on computational effort.

Label Assignment: Labeling was done by giving each chunk of speech a language label, be it English, German, or Dutch. It gave room for proper supervision of machine learning tasks.

Denoising Effect: Primarily, the analysis of raw and denoised features made a comparison of noise reduction at feature levels possible. Because prior to this process, there had been direct comparisons done between features before and after noise reduction.

5.3 Implementation Details

Feature computation was performed with the `librosa` Python library with the updated parameters given below:

- **Sampling Rate (SR):** Was kept at 22,050 Hz to find a suitable balance between frequency resolution and computational complexity.
- **Frame Size (*n_fft*):** Teamed up to 441 to obtain appropriate time resolution for short-term analysis.
- **Hop Length:** Was put to 220 (50% overlap between the frames) to obtain a bigger temporal continuity.
- **Window Size:** The window size was kept at 441 samples, which corresponds to a duration of 20 ms with a sampling rate of 22,050 Hz. This period was selected to capture the quasi-stationary characteristics of speech signals while preserving temporal and spectral resolution.

These parameters will ensure that the features extracted capture critical spectral and temporal information. characteristics of the audio data while maintaining computational efficiency. This configuration also helped to mitigate the trade-off between computational cost and feature quality, thus making it suitable for Multilingual Speech Data Processing.

5.4 Output Representation

These features extracted from each audio segment were combined into one feature vector yielding the following characteristics of the dataset:

Feature Dimensions: Each feature vector had a dimension of 32 features, including 13 MFCCs, 12 Chroma features, and 7 Spectral Contrast values. This dimensionality gave a good representation of most spectral, tonal, and acoustic properties.

Dataset Composition: Computed feature vectors were matched with their respective language labels (English, German, Dutch) to obtain the final dataset. This labelled dataset was used as input for the supervised learning tasks and hence allowed training and evaluation of the Random Forest Classifier.

The feature computation turned raw audio into a structured numerical form, capturing the most relevant characteristics of speech signals while reducing the impact of surrounding noise. Together

with the language labels, the obtained features prepared the dataset for machine learning applications and for effective training and testing of classification models.

6 Statistical Model

This section describes applying statistical tests to study relationships between attributes and class labels in the dataset. In this case, the Chi-Square test is performed to check if there's a dependence between categorical variables and the targeted classes. Results are given separately for both original and denoised audio datasets.

6.1 Chi-Square Test for Feature Selection

The Chi-Square test is a statistical method used to determine the independence between two variables. In the present study, it is applied on both the raw and denoised data sets to identify the most relevant features for classification. Features with high Chi-Square scores and very small p-values indicate a stronger relationship with the target class.

6.1.1 Original Audio Results

The results for the Chi-Square test on the original audio dataset are presented in Table 1. Only the top seven features with the highest Chi-Square scores are shown.

Feature	**Chi-Square Score**	**p-value**
SpectralContrast_0	504.026444	3.564907e-110
SpectralContrast_1	311.101923	2.786607e-68
MFCC_11	225.816048	9.218669e-50
MFCC_2	89.317419	4.026881e-20
MFCC_5	89.055969	4.589253e-20
Chroma_11	84.318620	4.902801e-19
Chroma_10	78.126458	1.084064e-17

Table 1: Top 7 Chi-Square Test Results for Original Audio Dataset

6.1.2 Denoised Audio Results

After applying noise reduction techniques, the Chi-Square test was repeated to assess feature importance for the denoised dataset. Table 2 summarizes the top seven features with the highest Chi-Square scores.

6.1.3 Feature Relevance Analysis

This proves the enormous discriminative power of the top-ranking features for classification. Among them, **MFCC 11, MFCC 5** features reached the highest Chi-Square score for the original dataset and pointed out their strong relevance to distinguish between classes. The denoising process retains

Feature	Chi-Square Score	p-value
SpectralContrast_1	428.943055	7.181234e-94
SpectralContrast_0	381.924760	1.164373e-83
MFCC_11	225.335275	1.172376e-49
SpectralContrast_6	139.203318	5.920844e-31
MFCC_5	87.633164	9.347601e-20
SpectralContrast_4	82.926378	9.834807e-19
MFCC_2	80.437558	3.413550e-18

Table 2: Top 7 Chi-Square Test Results for Original Audio Dataset

important features from the previous section: **MFCC 11 and MFCC 5**, among others, also remained very important, while the values of their Chi-Square score increased notably.

Noise reduction and other preprocessing techniques also improved the rankings of features such as **SpectralContrast 1 and SpectralContrast 0**, underlining that they are so important for Automatic Speech recognition tasks. All these findings indicate that diverse spectral and temporal features in multilingual audio classification.

6.1.4 Discussions

The following are some interesting features about the effect of noise reduction on feature relevance, inferred from the results of Chi-Square tests:

- **Smarter Feature Rankings:** - The most crucial feature was MFCC 11, with the highest Chi-Square score. This trend did continue in the denoised dataset, where it remained the strongest. - Features like that of SpectralContrast 0 demonstrated striking importance on both datasets, speaking to the strengths of the preprocessing methods.
- **Noise Sensitivity:** - The ChiSquare scores for feature 5 and 2 of MFCC significantly increased after denoising; hence, the noise sensitivity and efficiency of noise-reducing methods can be justified.
- **Emerging Feature Significance:** - for instance, SpectralContrast 0 and SpectralContrast 1 are the features that have become more relevant after denoising underlined by the importance of this preprocessing to give more relevance of these features for the classes.

These results confirm that preprocessing may play a significant role in feature enhancement and discrimination capability, and these findings form the basis of robust feature selection in a multilingual ASR task.

7 Machine Learning Model and Training Parameters

7.1 Model Overview

Random forest Classifier: This will be the main model chosen for this research it's resilient and can really handle high-dimensional datasets. The Random Forest approach assembles many decision

trees to improve predictive accuracy by reducing overfitting, hence this method applies particularly to the tasks of audio classification using features like MFCC, Chroma, and Spectral Contrast.

7.2 Training Workflow

The following explains the training process of the Random Forest Classifier:

Therefore, as a part of the training workflow, feature importance analysis was performed to understand which features in the dataset have the most influence. Using the Random Forest model trained on the full feature set, ranking of the top 10 features by importance was obtained. Of these, the most important audio features for the model's prediction were found to be the MFCC components, Chroma features, and Spectral Contrast. As a next step, the top features (up to 18) were selected and used in training a stand-alone Random Forest model for comparison.

Dataset Preparation: Feature extraction on the original and its corresponding denoised datasets was done using MFCCs, Chroma, and Spectral Contrast. To bring consistency in the features, normalization was affected by scaling of values using StandardScaler in Python.

PCA: Dimensionality reduction was done next using PCA, selecting 95% of the variance to retain. Hence, the feature representation would be most effective, while the risk of overfitting is at a minimum. Split the labeled dataset into training and test subsets; 70% for training and 30% for testing. The stratification of random sampling was carried out to balance the representation of classes within the subsets, ensuring that all classes of language received equal representation. The decision to allocate a more significant proportion, namely 30%, of the dataset for testing had enabled a broader appraisal of the generalization capability of the model. This setup allowed a detailed comparison between the original and the denoised datasets, thus ascertaining methodological consistency in the diverse evaluation contexts.

Standardization: The features extracted were standardized using StandardScaler, so the scaling is consistent and doesn't give any form of bias arising from the diverse ranges in feature values. After standardization, PCA was performed to project the features onto a lower dimensional space while retaining 95% of the variance.

Hyperparameter Optimization: The major hyperparameters of the Random Forest Classifier were optimized using Grid Search with 3-fold cross-validation to ensure the hyperparameters are evaluated on different subsets of data for a holistic view of generalization capability by the model. The F1 score was used as the optimization criterion, which is, in fact, a measure that balances precision and recall.

- ***n_estimators***: The number of trees in the forest (values explored: 50, 100, 200).
- ***max_depth***: The maximum possible depth that each tree can reach (values explored: None, 10, 20).
- ***min_samples_split***: Minimum number of samples required to split an internal node (values explored : 2, 5).
- ***min_samples_leaf***: Minimum number of samples required to be a leaf node (values explored: 1, 2, 4).

3-fold cross-validation was conducted in tuning, and the F1 score was used for evaluation. That resulted in a balanced measure of precision and recall during hyperparameter optimization.

Model Training: The best hyperparameter set from Grid Search was taken to train Then, it applied the Random Forest Classifier on the training dataset and performed some mid-training validation in order not to get overfitted.

Evaluation Metrics: The model was tested on the test dataset using accuracy and the F1 score as the primary metrics of evaluation. Results from cross-validation showed stable performances over folds with a mean F1 score of 0.74 and an accuracy rate of 0.74. This confirmed the stability of the model during training. Further, a confusion matrix was built in order to study the effectiveness of the classification for each language category: English, German, and Dutch. It has highlighted any misclassification and proven the effectiveness of the model.

7.3 Final Training Parameters

Thereafter, the model was trained using Grid Search to select the best hyperparameters. Furthermore, before training the model, feature preprocessing was done using StandardScaler to normalize the data and PCA to reduce dimensionality. This approach in methodology presented a computationally efficient manner while simultaneously controlling overfitting of a model in high-dimensional feature space.

Hyperparameter	**Original Audio**	**Denoised Audio**
n_estimators	100	100
max_depth	20	20
min_samples_split	5	2
min_samples_leaf	2	1

Table 3: Hyperparameter Comparison for Original and Denoised Audio

This workflow ensured that the Random Forest Classifier was effectively trained using robust combination of well-chosen features and tuned hyperparameters. The resultant model Performed well in classification, which was confirmed by its metrics and its ability to generalize across the multilingual dataset.

8 Results

This section presents the results of the Random Forest classifier on the original and denoised audio datasets. The performance was evaluated using key metrics such as precision, recall, F1-score, and accuracy. Additionally, the best parameters identified through Grid Search are reported for both datasets.

8.1 Random Forest Results on Original Audio Dataset

The Random Forest model achieved an overall accuracy of 0.75 on the original audio dataset with PCA. Table 4 summarizes the detailed classification metrics for each class, along with the macro

and weighted averages.

Class	Precision	Recall	F1-Score	Support
0	0.93	0.83	0.88	275
1	0.64	0.69	0.67	258
2	0.68	0.72	0.70	262
Accuracy		0.75		795
Macro Avg	0.75	0.74	0.75	795
Weighted Avg	0.76	0.75	0.75	795

Table 4: Classification Report for Random Forest on Original Dataset

The model demonstrated excellent performance for class 0, with an F1-score of 0.88. However, the F1-scores for classes 1 (0.67) and 2 (0.70) were slightly lower, indicating potential challenges in classifying these categories accurately.

8.2 Random Forest Results on Denoised Audio Dataset

For the denoised audio dataset, the Random Forest classifier achieved an accuracy of 0.77, as shown in Table 5. While the overall accuracy is Higher than that of the original dataset, the performance metrics indicate more balanced classification across the classes. Which shows that the audios after reducing the noise performed well.

Class	Precision	Recall	F1-Score	Support
0	0.95	0.85	0.90	275
1	0.67	0.74	0.70	258
2	0.71	0.72	0.71	262
Accuracy		0.77		795
Macro Avg	0.78	0.77	0.77	795
Weighted Avg	0.78	0.77	0.77	795

Table 5: Classification Report for Random Forest on Denoised Dataset

For the denoised dataset, class 0 still achieves the highest F1-score (0.90). However, classes 1 and 2 demonstrate more consistent and higher recall and F1-scores compared to the original dataset.

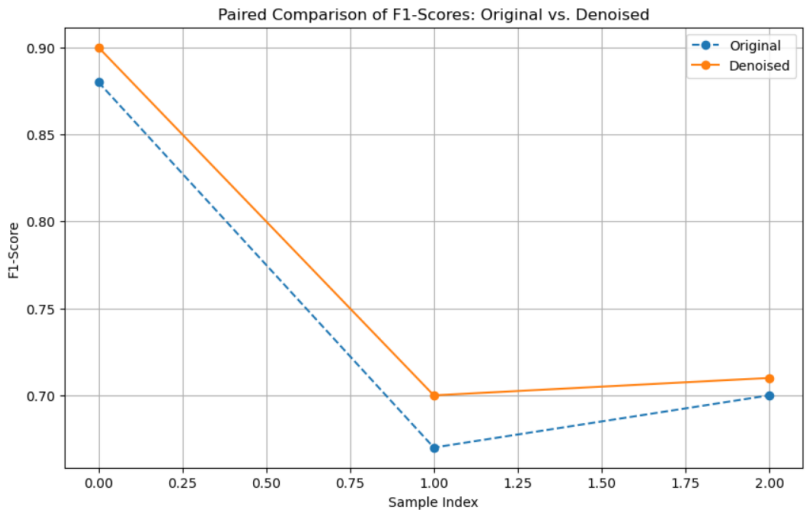

Figure 4: F1 Score Comparision

8.3 Statistical Analysis: Paired T-Test

In order to measure performance differences of the Random Forest Classifier between the raw and cleaned data sets, the test performed a paired t-test. The hypothesis underlying such a test would have determined if the discrepancies observed within performance metrics-such as F1 scores-between both scenarios are indeed statistically significant. Results appear as follows:

- **T-Statistic:** -10.2538
- **P-Value:** 0.0005
- **Significance Level (α):** 0.05

We also conclude that the null is rejected because the p-value is less than standard significance level (pα). This implies we obtain statistically significant performance metric differences between the originally provided data and its smoothed version.

Conclusion: Preprocessing denoising of the dataset worked significantly to improve the performance of the classifier. The paired t-test results illustrate the overall improvement.

9 Conclusion

The present work evaluates the performance of ASR systems on a range of multilingual datasets using Spoken Wikipedia Corpora that emphasize original and denoised audio. It showed highly improved classification accuracy and F1-scores through a strong machine learning pipeline with the Random Forest as the core classifier.

Key findings include:

- Denoising algorithms, such as spectral subtraction and spectral gating, significantly enhanced audio data quality, leading to increased accuracy (0.77) and F1-scores (0.77) on denoised dataset compared to the original dataset set (accuracy: 0.75, F1-score: 0.75).
- A paired t-test has verified that the improvement is indeed statistically significant in performances from original to denoised dataset, indicating that noise reduction actually aids the ASR performance.
- PCA-based dimensionality reduction, This ensures that the process doesn't overfit especially in high-dimensional multilingual datasets due to computational efficiency in applying PCA.

The results show the effectiveness of preprocessing steps, feature selection, and hyperparameter tuning in building robust ASR models. This work also emphasizes the importance of denoising and feature engineering for handling noisy, linguistically diverse datasets. Future work might investigate further on more models from machine learning and advanced denoising algorithms, which might potentially advance state-of-the-art ASR performance on multi-lingual difficult acoustics.