Noise-Robust Speech Signal Processing: Enhancing Speech-to-Text Model Performance in Noisy Environments through Advanced Feature Engineering

1st Veera Venkata Megha Shyam Ankem
VeeraVenkataMeghaShyamAnkem@my.unt.edu
University of North Texas
Denton, Texas, USA

2nd Suguna Sai Navaneeth Rentala
SugunaSaiNavaneethRentala@my.unt.edu
University of North Texas
Denton, Texas, USA

3rd Nikhitha Rayabarapu
NikhithaRayabarapu@my.unt.edu
University of North Texas
Denton, Texas, USA

4th Dinesh Teja Lingala
DineshTejaLingala@my.unt.edu
University of North Texas
Denton, Texas, USA

***Abstract*—Speech-to-text models have always had a difficult time in noisy environments where the presence of irrelevant sounds, and overlapping speech significantly reduces transcription accuracy. When dealing with real-world scenarios like healthcare, telecommunication, and assistive technologies, this challenge hinders the process of communication. Hence, managing noise types, ensuring data quality, and addressing class imbalances are crucial to improving model scalability and performance in formal and informal real-world settings.**

***Index Terms*—speech-to-text, Noise, Polynomial Features, Feature Extraction, Feature Engineering, Feature Selection, Spectral Features, Mel Frequency Coefficients, Support Vector Machine, Hyperparameter tuning**

I. Introduction

Speech-to-text systems have become essential tools in a number of applications in recent years, such as accessibility technology, virtual assistants, and transcription services. Accurately converting spoken language into text is the foundation of these systems, and this process is highly reliant on the input speech signal's quality. However, noise, such as background talk, ambient sounds, or electronic interference, frequently contaminates speech signals in real-world situations, which can seriously impair these systems' effectiveness. This difficulty emphasizes how crucial it is to provide noise-resistant speech signal processing methods that guarantee dependable performance in a range of acoustic conditions.

Noise presents two challenges: it can generate artifacts that confuse machine learning algorithms and conceal the acoustic properties that are essential for phoneme and word recognition. Advanced techniques that combine domain expertise and data-driven innovations are required because traditional denoising methods frequently fail to generalize across different noise types and intensities. Feature engineering, in particular, plays a pivotal role in bridging this gap by transforming raw speech signals into representations that are more resilient to noise while preserving the information required for accurate transcription.

The goal of this project is to apply sophisticated feature engineering techniques to improve speech-to-text models' performance in noisy situations. Our goal is to provide a framework that enhances recognition accuracy in a variety of acoustic circumstances by utilizing a combination of noise suppression algorithms, reliable feature extraction, and cutting-edge signal processing techniques. By making communication more dependable in noisy real-world situations, the suggested method not only improves the robustness of speech-to-text systems but also advances the field of human-computer interaction.

This project is part of the CSCE 5222.002: Feature Engineering Course, University of North Texas, Texas, USA.

II. Related Works

Rabiner et al. [14] pioneered early feature extraction techniques with linear predictive coding (LPC), which laid the foundation for modern methods. MFCC features, introduced by Davis and Mermelstein [2], have become a standard in speech processing. Recent studies have incorporated complementary features such as spectral centroid and RMS energy to improve noise-robustness [5], [15].

Dimensionality reduction techniques like PCA have been widely used to minimize noise effects in feature space. Tipping and Bishop [8] demonstrated how PCA aids in compressing high-dimensional speech data while retaining discriminative power. Applications of PCA in noisy speech datasets show its effectiveness in enhancing classifier robustness.

Vapnik's seminal work [10] introduced SVM as a robust alternative to neural networks for small to medium datasets. Subsequent studies demonstrated that SVM with linear, RBF, and polynomial kernels could achieve high accuracy in noise-affected speech classification tasks. Grid search and cross-validation, as used by Wang et al. [16], further optimize model performance.

Recent advances in deep learning have shown promise in handling noise, as demonstrated by Amodei et al. [17]. However, simpler models like SVM remain relevant due to their computational efficiency and explainability, especially when combined with advanced feature engineering techniques [18].

III. About the Dataset

We used the LibriSpeech dataset [19], a well-established benchmark for speech recognition tasks. The LibriSpeech ASR Corpus is a corpus of read speech designed to aid the training and testing of Automatic Speech Recognition (ASR) systems. This subset was made from combining dev-clean, test-clean sets provided and we named it as Clean-data. Similarly the corresponding noise sets were combines to get final noisy-data along with respective transcriptions. Overall, 4-5 hours of audio data has been taken.

A. Dataset Highlights:

Corpus Origin: Based on LibriVox's public domain audio books.

Purpose: Enable the training and testing of ASR systems.

B. Data Overview:

Audio Data: Segmented and aligned audio from audiobooks.

Text Data: Original transcriptions corresponding to the audio.

C. Directory Organization:

A root directory with metadata and a subfolder specifically for the subset are recreated for each of the train and test sets. Every speaker's audio is kept in a different subfolder, and each audio chapter is kept in a different subdirectory. Transcripts for every utterance are contained in *.trans.txt files, whilst the audio itself is contained in FLAC files.

IV. Methodology

The foundation of this project is feature engineering, which converts unprocessed signals into representations that are more resilient to noise, thereby bridging the gap between raw audio data and efficient model learning. The techniques used were intended to improve the model's generalization in noisy settings while maintaining speech's fundamental properties. The three primary steps of the method—**feature extraction, feature transformation**, and **feature selection**—are each designed to tackle distinct problems presented by noisy data.

A. Feature Extraction:

Feature Extraction stage focuses on extracting relevant characteristics from raw audio signals that encapsulate the spectral, temporal, and energy-related aspects of speech. These features are designed to help the model distinguish between useful speech information and irrelevant noise.

1) Mel-frequency cepstral coefficients (MFCCs):: MFCCs use the Mel scale, which closely resembles human auditory perception, to record the frequency content of audio signals. Because it highlights the most perceptually important frequency bands, this feature works very well in speech recognition tasks.

2) Spectral Features:: We have extracted the spectral features of the audio data that includes:

Spectral Centroid: Indicates the "center of mass" of the spectrum, helping identify brightness in the signal.

Spectral Rolloff: Measures the frequency below which a specified percentage (e.g., 85%) of total energy lies.

Spectral Bandwidth: Quantifies the width of the spectrum, distinguishing between noise and speech.

Spectral Flatness: Describes the tonal or noise-like quality of the signal

3) Temporal Features:: Some of the temporal features were also extracted.

Zero-Crossing Rate (ZCR): Measures how often the signal changes its sign, useful for detecting unvoiced segments.

RMS Energy: Computes the root mean square energy to capture the signal's intensity.

B. Feature Transformation:

Feature transformation, which aims to maximize the extracted features' interpretability, consistency, and noise resistance. This stage entails transforming the retrieved or raw features into a format that addresses dataset variability and guarantees better interoperability with machine learning techniques. The relevance, application, and anticipated effects of each transformation technique used in the project are covered in detail below.

1) Audio Normalization:: The amplitude of audio signals is adjusted to a specified range using audio normalization. The loudness of recordings in dataset frequently varies because of variations in recording equipment, speaker distance, or ambient conditions. This variation may affect the model's capacity to generalize and result in uneven feature extraction. Hence, normalization was applied to ensure all recordings had the same maximum amplitude. This process was automated using Librosa's amplitude normalization function, which rescales each sample in the audio dataset.

C. Feature Selection:

In the Feature selection stage, we determine the features that are most significant and eliminate others that are irrelevant. By concentrating on the characteristics that most influence the model's predictive power, this stage not only lowers computational complexity but also improves performance. By

separating features that are resistant to noise while preserving speech-related information, feature selection becomes even more crucial in the situation of noisy voice data.

1) Recursive Feature Elimination (RFE):: RFE iteratively removes the least important features based on their contribution to model performance, focusing on retaining the most impactful subset. We prioritize the features that significantly affect classification accuracy and eliminates the noise-sensitive and redundant features. We have conducted RFE using a Support Vector Machine (SVM) model as the estimator and retained the top 10 features after multiple iterations. As a result, it improved model generalization by focusing on noise-resilient features and simplified model complexity while maintaining robust performance.

2) Principal Component Analysis (PCA):: PCA is a dimensionality reduction technique that projects features onto orthogonal components, ranked by the variance they capture. It reduces computational complexity by retaining only the most informative components. It also removes correlations between features to improve model efficiency.

TABLE I
SUMMARY OF EXTRACTED FEATURES

Feature Name	Description	Dimension
MFCCs	13 coefficients representing spectral properties on the Mel scale	13
Spectral Centroid	Mean "center of mass" of the spectrum	1
Spectral Rolloff	Mean frequency containing 85% of signal energy	1
Spectral Bandwidth	Average frequency spread of the signal	1
Spectral Flatness	Ratio of geometric to arithmetic mean of power	1
Zero-Crossing Rate	Rate of zero-amplitude crossings in the signal	1
RMS Energy	Average root mean square energy of the signal	1

We have reduced the dimensionality of the feature space by 40% on average and also enhanced model training speed without sacrificing accuracy.

The final feature set, after applying transformations and selection, was reduced from 19 original features (i.e., MFCCs, spectral, and temporal features) to a compact set of 10 features, depending on the applied technique (i.e., PCA or RFE).

V. EXPERIMENTAL SETUP AND MODEL EVALUATION

A. Data Preprocessing:

To normalize feature values, this step is essential for optimizing the performance of models like SVM, which are sensitive to the scale of input features.

From the audio data, we have extracted the features including MFCCs, spectral features, zero-crossing rate, and RMS energy and transfromed them to the standardized scale using StandardScaler from scikit-learn library in Python.

Fig. 1 compare the overall amount of Clean and Noisy data and we can observe that there is more noisy data than clean data and the data is pretty much balanced.

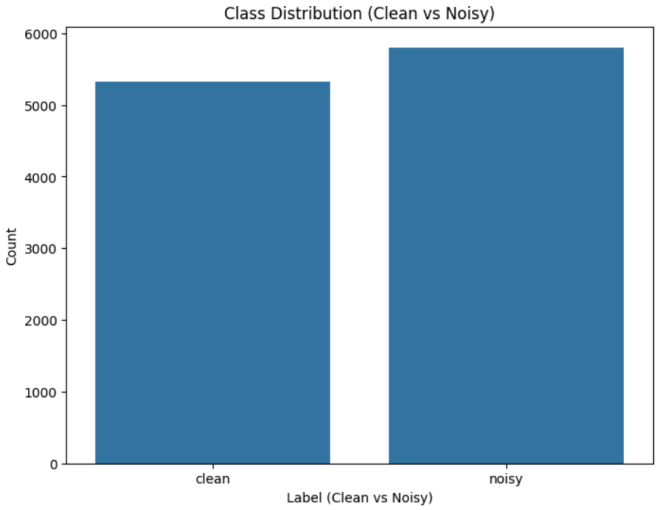

Fig. 1. Distribution of Clean and Noisy data

Fig. 2 shows the distribution of Mel-Frequency Cepstral Coefficients (MFCC) across the dataset. The values range between -500 to 100. The median values and the overall shape of the distributions differ across the MFCC features, indicating that they capture different aspects of the audio signal. Similarly Fig. 3 shows the distribution of Spectral Features while their values range between 0 to 7000.

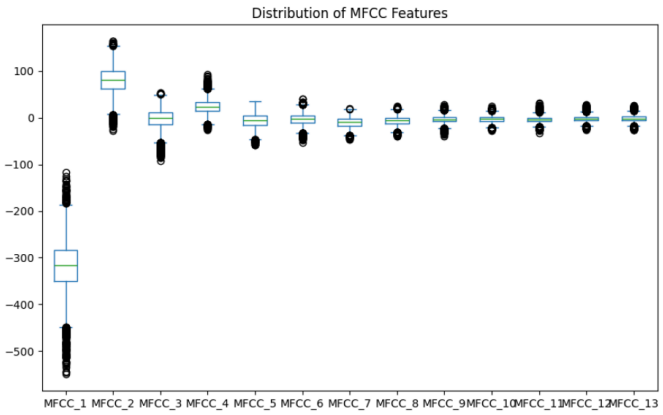

Fig. 2. Distribution of MFCC Features

Fig. 4. shows the correlation heatmap of all the audio features extracted. The MFCC features have a high negative correlation indicating individual contribution to the classification. However, Spectral features are strongly correlated with each other (0.85–0.96), suggesting redundancy. This was later avoided by performing feature selction (PCA or RFE) on the data.

B. Baseline Model:

In order to evaluate the effect of feature engineering strategies on speech-to-text model performance in noisy environments, we are usinf a Support Vector Machine (SVM) model as the baseline classifier in this project. SVM works well with

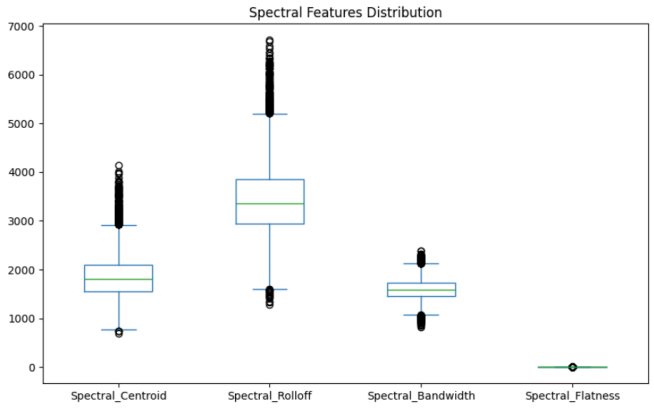

Fig. 3. Distribution of Spectral Features

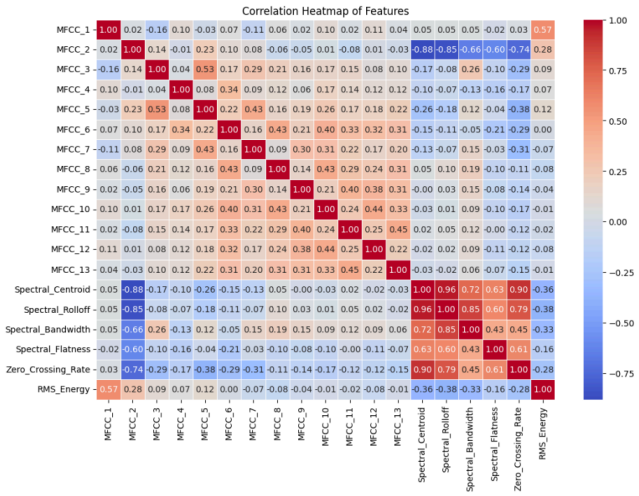

Fig. 4. Correlation between all the features

a range of kernels and is resistant to overfitting, especially in small- or medium-sized datasets. Assessing how effectively designed features, such MFCCs and spectral features, as well as transformations like PCA, RFE, and polynomial features, might improve the model's capacity to accurately identify speech signals in the presence of noise was the main goal of utilizing SVM.

C. Evaluation Metrics:

The model was assessed using a number of crucial parameters that are crucial for comprehending speech recognition task performance, particularly in noisy environments:

Accuracy: Accuracy is the percentage of correctly classified examples to all occurrences, which indicates how accurate the model's predictions are overall. When it comes to speech-to-text models, accuracy gives a broad indication of how frequently the model correctly classifies a particular audio sample as clean or noisy.

Precision: Out of all projected positives, precision quantifies the percentage of accurately predicted positive cases (such as noisy speech). When a model predicts a noisy signal with high precision, it is considered dependable.

Recall: The model's recall assesses how well it can recognize all real positives (noisy speech). Although the model may contain more false positives, high recall indicates that it is sensitive to identifying noisy speech. These measures provide a more thorough understanding of the model's performance on unbalanced classes and its capacity to prevent errors in data that is prone to noise.

ROC: The ROC curve plots the Sensitivity (TPR) against the 1-Specificity(FPR) across various thresholds, assessing a model's performance.

Confusion Matrix: This is used to analyze errors made by the model, showing how often it incorrectly identified noisy audio as clean or vice versa.

After splitting the data, we have taken about 1065 clean data samples and 1161 noisy data samples for training the model.

VI. Results and Discussion

A. Baseline Model - SVM:

We train the baseline SVM model with a linear kernel and balanced class weights to handle imbalanced data. The model is evaluated using 5-fold cross-validation, splitting the training data into five subsets to ensure robust performance assessment.

```
Classification Report:
              precision    recall  f1-score   support

       clean       0.66      0.69      0.67      1065
       noisy       0.70      0.67      0.69      1161

    accuracy                           0.68      2226
   macro avg       0.68      0.68      0.68      2226
weighted avg       0.68      0.68      0.68      2226
```

Fig. 5. Classification Report for Baseline SVM Model

The baseline SVM model has given an accuracy of 68% as observed from Fig. 5. About 1515 samples were correctly classified and the rest of the 711 samples were misclassified by the model (from Fig. 6). After performing Cross-validation, the accuracy scores are as follows: [0.66901408 0.67645332 0.69406929 0.68467 41 0.65590135]

Fig. 7. shows the trade-off between TPR and FPR. The AUC value of 0.72 indicates that the model has performed better than random chance, however more feature engineering is required. Hence, we perform hyperparameter tuning in the next step.

B. Hyperparameter Tuning:

The SVM model's hyperparameters were adjusted using GridSearchCV. We have re-applied 5-fold cross-validation (cv=5) using accuracy as the scoring measure and created a parameter grid with values for C, kernel, and gamma. Following hyperparameter tweaking, accuracy reached about 97%. The model was optimized for processing loud voice signals through this procedure.

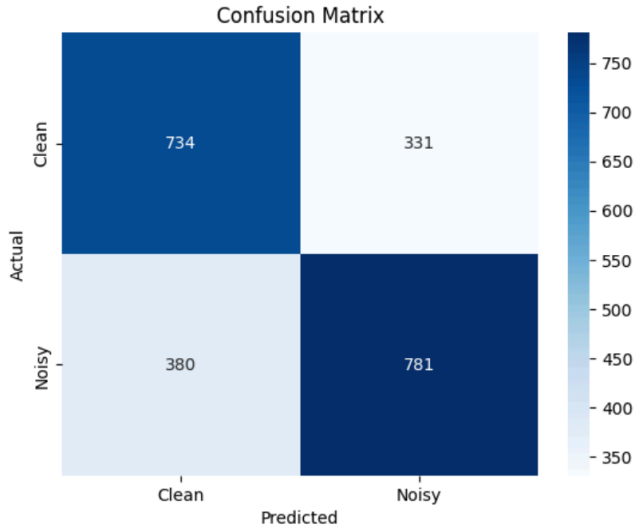

Fig. 6. Confusion Matrix for Baseline SVM Model

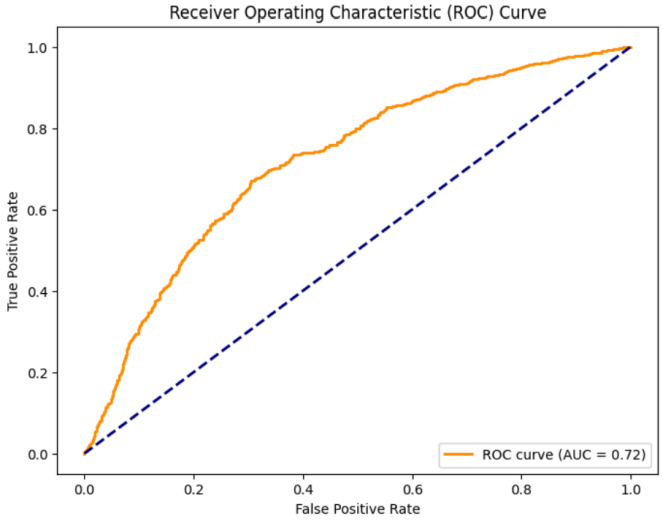

Fig. 7. ROC for Baseline SVM Model

After performing Cross-validation, the accuracy scores are as follows: [0.67402616 0.81892973 0.77125585 0.67402 616 0.81810765 0.76867238 0.67578755 0.93365553 0.87341635 0.67578755 0.93236377 0.8723596 0.67637461 0.96688659 0.91181373 0.67637461 0.96676915 0.91087428]

We can see some significant improvement in the model's performance given its accuracy of 97% as observed from Fig. 8. Most of the data (about 2100 samples) has been correctly classified by the model (from Fig. 6). The high AUC value of 0.99 shows that our model performed really well.

C. Advanced Feature Engineering (RFE and PCA)

1) Using PCA:: To minimize dimensionality while maintaining the maximum variance in the data, Principal Component Analysis, or PCA is applied. Prior to applying the same transformation to the test data, we convert the training data into

Classification Report:

	precision	recall	f1-score	support
clean	0.95	0.98	0.97	1065
noisy	0.98	0.95	0.97	1161
accuracy			0.97	2226
macro avg	0.97	0.97	0.97	2226
weighted avg	0.97	0.97	0.97	2226

Fig. 8. Classification Report after Hyperparameter tuning

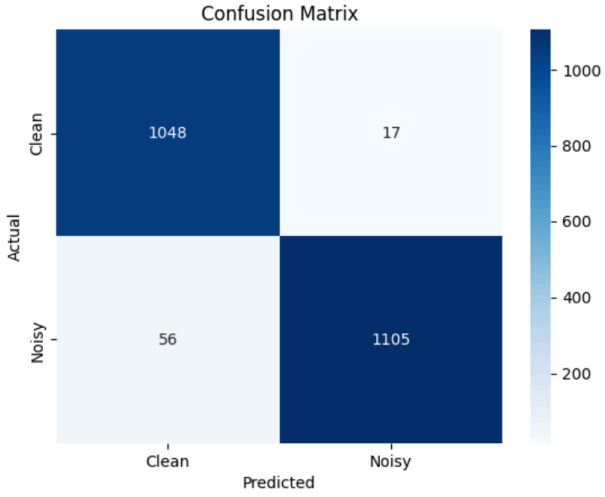

Fig. 9. Confusion Matrix after Hyperparameter tuning

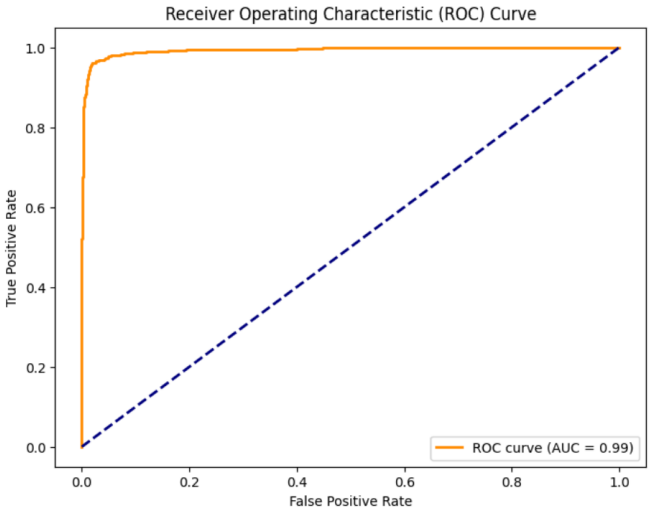

Fig. 10. ROC after Hyperparameter tuning

10 principal components. Cross-Validation Accuracy Scores for PCA: [0.59683099 0.59953024 0.61421022 0.59541985 0.59130945]

Classification Report with PCA:

	precision	recall	f1-score	support
clean	0.57	0.63	0.60	1065
noisy	0.63	0.57	0.60	1161
accuracy			0.60	2226
macro avg	0.60	0.60	0.60	2226
weighted avg	0.60	0.60	0.60	2226

Fig. 11. Classification Report with PCA

From Fig. 11., We can observe that the SVM Model with PCA has given and accuracy of 60%. The model have majorly misclassified Noisy data as Clean data (about 502 samples). Furthermore, the low AUC value and given how closer the ROC Curve is to the random classifier (blue dashed line) suggests that the PCA-transformed features did not significantly improve the model's ability to distinguish between clean and noisy classes.

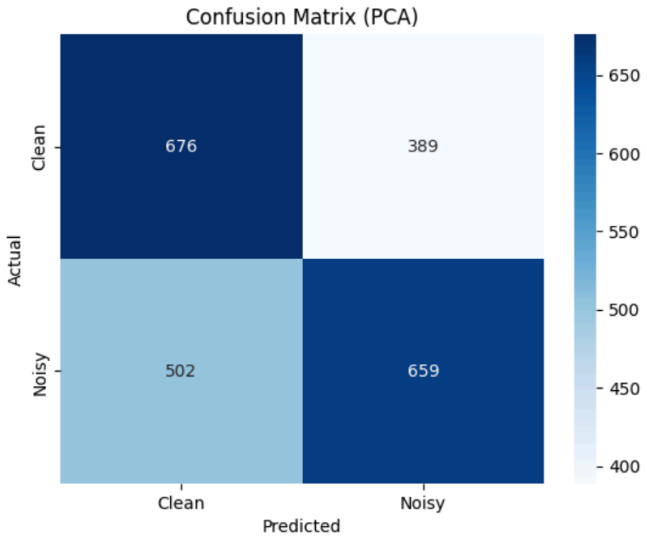

Fig. 12. Confusion Matrix with PCA

2) Using RFE:: We take the top 10 most relevant features to use RFE with an SVM (linear kernel). RFE finds the subset that contributes most to prediction accuracy by repeatedly eliminating the least significant characteristics and retraining the model. The cross validation accuracies for RFE are: [0.65551643 0.67410452 0.66294774 0.66529654 0.63358779]

From Fig. 14., We can observe that the SVM Model with RFE has given and accuracy of 65%. Though RFE has performed slightly better that PCA, we can observe that there is not much difference in RFE results compared to PCA.

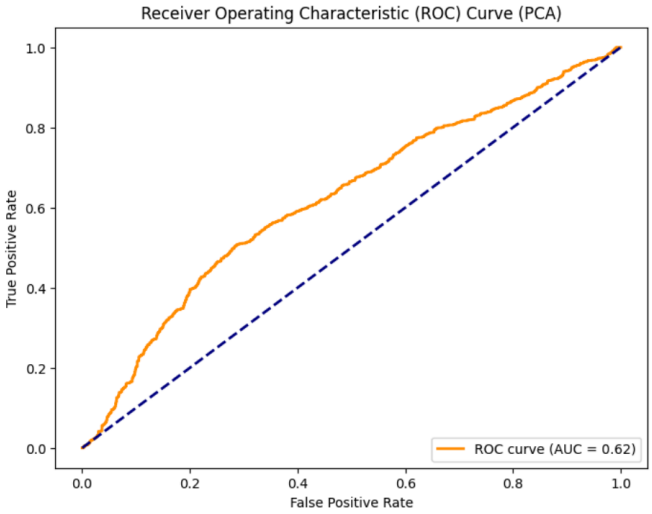

Fig. 13. ROC with PCA

Classification Report with RFE:

	precision	recall	f1-score	support
clean	0.63	0.67	0.65	1065
noisy	0.68	0.64	0.66	1161
accuracy			0.65	2226
macro avg	0.66	0.66	0.65	2226
weighted avg	0.66	0.65	0.66	2226

Fig. 14. Classification Report with RFE

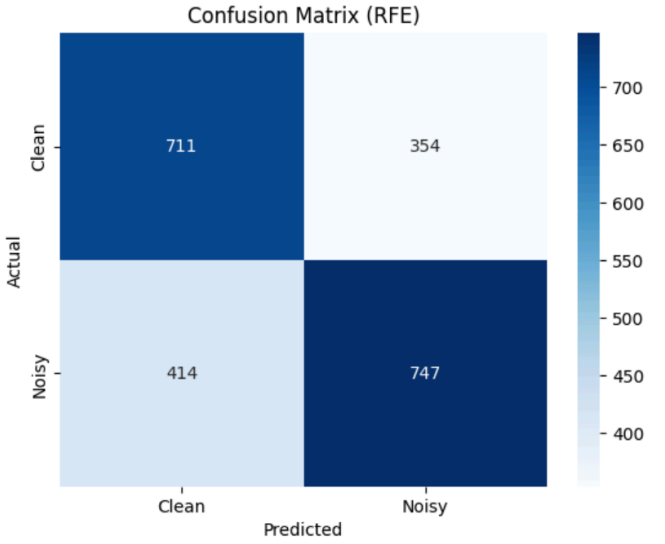

Fig. 15. Confusion Matrix with RFE

The RFE model correctly classified 1458 samples and misclassified the remaining 768 samples as observed from Fig. 15. The AUC Value is noted to be 0.71.

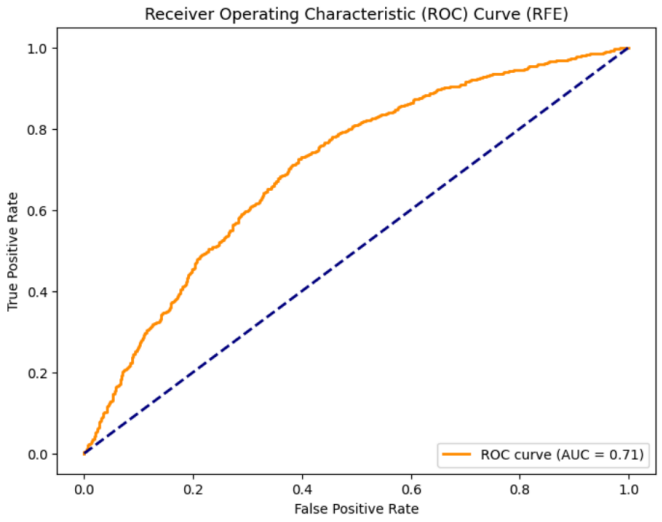

Fig. 16. ROC with RFE

D. Polynomial Features:

In order to capture non-linear correlations, this algorithm adds polynomial interactions with degree = 2 to improve features. This technique used to modify the training data and provide more feature combinations. The converted features are used to train a linear kernel SVM model. Cross-Validation Accuracy Scores for Polynomial Features are as follows: [0.83861502 0.8608 3382 0.83910746 0.85731063 0.85613623]

SVM Model with Polynomial Features Classification Report:

	precision	recall	f1-score	support
clean	0.81	0.87	0.84	1065
noisy	0.87	0.82	0.84	1161
accuracy			0.84	2226
macro avg	0.84	0.84	0.84	2226
weighted avg	0.84	0.84	0.84	2226

Fig. 17. Classification Report with Polynomial Features

As noted from Fig. 17, While the noisy class obtains greater accuracy (0.87), indicating fewer false positives, the clean class has higher recall (0.87), demonstrating better recognition of actual positives.

For the Clean data class, 925 samples were correctly classified, while 140 were misclassified as Noisy. For the Noisy data class, 949 samples were correctly identified, but 212 were misclassified as Clean.

From Fig. 19, it is noted that AUC = 0.92 which is a decent value indicating better model performance compared to baseline SVM, and SVM with PCA & RFE.

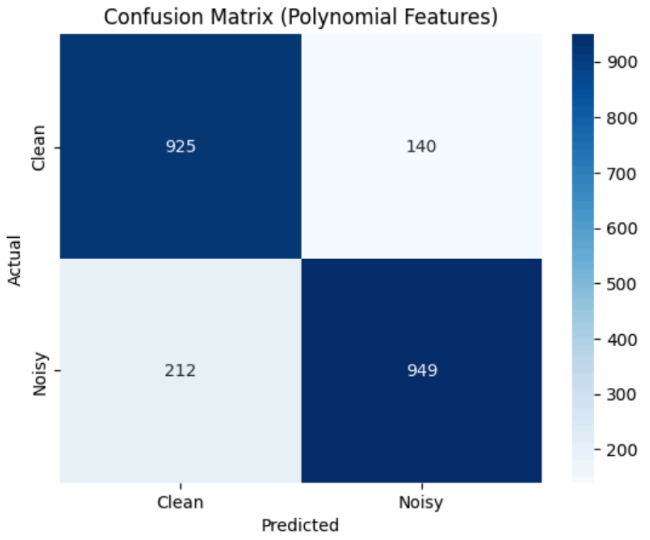

Fig. 18. Confusion Matrix with Polynomial Features

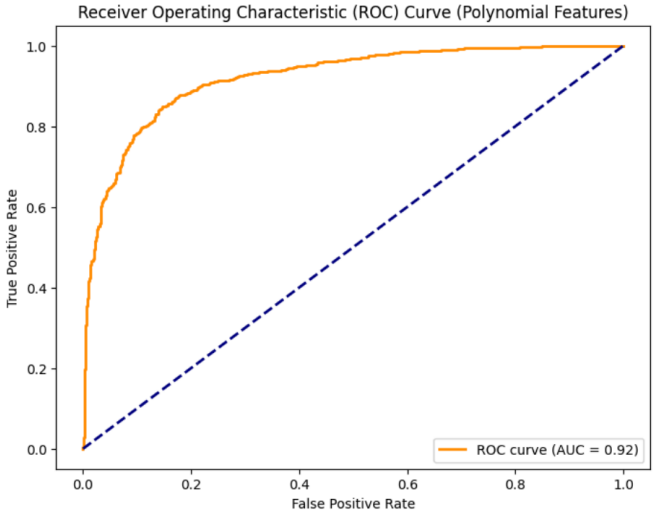

Fig. 19. ROC with Polynomial Features

E. Model Comparision:

After comparing the results of all the models' performance, we can observe that SVM after hyperparameter tuning has give the highest Accuracy (96.72%), Precision (98.48%), Recall (95.18%), F1-Score (96.80%), and ROC AUC (99.31%). The Polynomial Features model, while not as strong as Tuned SVM, shows relatively high scores, with an Accuracy of 84.19% and a ROC AUC of 92.09%, highlighting its effectiveness in capturing non-linear patterns. Overall, model tuning and feature engineering significantly impact performance.

VII. Conclusion and Future Work

In this project, we explored various methods for improving noise robustness in speech signal processing, with a focus on feature extraction, dimensionality reduction, and machine

TABLE II
PERFORMANCE METRICS OF DIFFERENT MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Baseline SVM	68.06%	70.23%	67.27%	68.72%	72.17%
Tuned SVM	96.72%	98.48%	95.18%	96.80%	99.31%
PCA	59.97%	62.88%	56.76%	59.67%	62.20%
RFE	65.50%	67.85%	64.34%	66.05%	70.56%
Polynomial Features	84.19%	87.14%	81.74%	84.36%	92.09%

learning techniques. Key techniques like Mel-Frequency Cepstral Coefficients (MFCC), spectral features, and dimensionality reduction methods such as Principal Component Analysis (PCA) were analyzed to enhance the robustness of speech-to-text systems under noisy conditions.

From the results, it was evident that a combination of feature engineering and optimization strategies led to improved performance in noisy environments. In terms of accuracy, precision, recall, and F1-score, the tuned-SVM outperformed the other models that were assessed with an accuracy of 97%. SVM with Polynomial Features turned out to be second-best given its accuracy of 84% .

A. Future Work:

- While traditional ML models like SVMs are effective,deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) could be used for handling more complex, noisy speech data.
- Multimodal data, such as textual, visual, and acoustic information, could be incorporated into future research to improve the robustness of voice recognition systems.
- Creating real-time noise-adaptive systems for more adaptable and reliable speech recognition in a range of real-world scenarios. This would dynamically modify their parameters to maximize performance depending on the noisy environment.

REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[3] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*. New York, NY, USA: Springer, 2004, pp. 191–245.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[5] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR)*, 2000, pp. 1–3.

[6] A. T. Cemgil and S. Dikici, "Bayesian filtering of spectrogram features for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1782–1786, Dec. 2016.

[7] K. K. Paliwal and L. A. Taylor, "Noise estimation algorithms for robust speech enhancement," *Speech Commun.*, vol. 12, no. 1, pp. 55–67, Mar. 1993.

[8] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. R. Statist. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Mar. 2002.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

[11] M. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[12] B. Schoelkopf *et al.*, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, May 1998.

[13] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[14] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[15] Z. Wang, Y. Han, and Y. Wang, "Improved spectral subtraction method for speech enhancement," *IEEE Trans. Consum. Electron.*, vol. 50, no. 1, pp. 86–89, Feb. 2004.

[16] J. Wang, M. Li, and C. Fang, "A grid-search algorithm for parameter tuning in SVMs," *J. Soft Comput.*, vol. 18, no. 7, pp. 1165–1173, Oct. 2014.

[17] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[18] Y. Bengio *et al.*, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[19] Dataset Link: https://www.openslr.org/12