SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

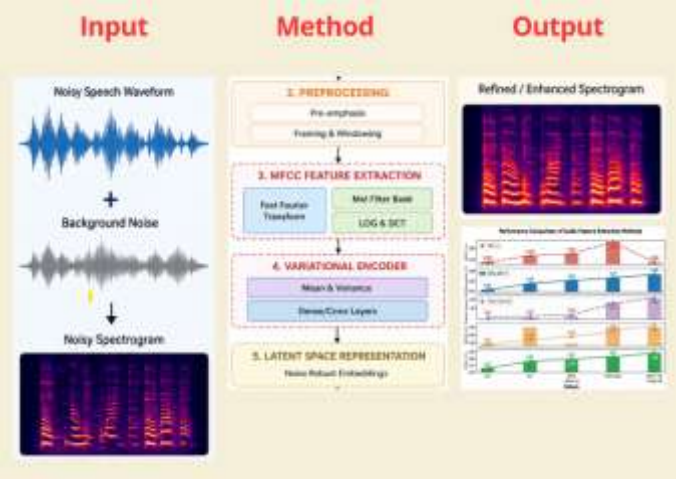
Disentangled Latent Representations For Robust Speech Recognition: A Probabilistic Mel-Frequency Cepstral Coefficient–Variational Autoencoder Approach

Mandar Diwakar[1], Brijendra Gupta[2]

[1]Ph.D. scholar, Smt. Kashibai Navale College of Engineering, Pune, Savitribai Phule Pune University, Pune, India mpdiwakar30@gmail.com
[2]Siddhant College of Engineering, Pune, Savitribai Phule Pune University, Pune, India gupbrij@rediffmail.com

Abstract: The performance of speech recognition systems greatly depends on how well we extract features from sound samples. Standard tools like MFCC (Mel-Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), and LPC (Linear Predictive Coding) work well in quiet environments, but they often have trouble in the real world, where background noise distorts the sound signal. Even newer models, such as MFCC-GANs (Mel-Frequency Cepstral Coefficient - Generative Adversarial Network), sometimes fail to capture the complex and shifting nature of noisy settings. To address this issue, we suggest a hybrid feature extraction system that combines the reliability of MFCC with the flexibility of a VAE (Variational Autoencoder). The main idea is to use the VAE to map noisy speech data into a structured latent space. This allows the model to ignore random noise and focus on the key patterns of the human voice. Our results indicate that this hybrid approach is much more robust than traditional methods. In tests conducted under noisy conditions, the proposed model achieved the lowest Mean Squared Error (MSE) and surpassed MFCC, LPC, PLP, and GAN-based models. Additionally, a high silhouette score shows that our VAE-processed features are cleaner and easier for the system to differentiate. These findings suggest that using probabilistic models like VAEs can greatly enhance the reliability of speech recognition in everyday noisy environments.

Index Terms: Feature Extraction, MFCC, Variational Autoencoder, Speech Recognition, Machine Learning.



Graphical Abstract

I. INTRODUCTION

Feature extraction is a primary part of speech recognition system. The significant feature extracted from raw audio waveforms could build a good foundation for a robust, accurate and enhanced speech recognition system [1]. Speech recognition -based applications are widely used in essential human centric services. The effectiveness and accuracy of

any speech recognition system depend upon the quality of features extracted from raw audio waveforms [2][3]. The speech enhancement model uses the representation of speech, such as MFCC, spectrogram, and embeddings. If the speech representation were poor, the speech enhancement model cannot distinguish between noise and speech components. There are some conventional methods used for feature extraction, which work well and produce good results in a clean environment, but their performance degrades significantly in noisy and distorted environments. [4][5]. MFCC is widely used for feature extraction due to its ability to capture the features that are perceptual to human hearing. However, MFCC is highly sensitive to background noise [6]. This study proposes a new hybrid architecture for extracting the features from raw audio with effective accuracy in noisy and distorted environment. For this purpose, we introduced a framework that integrates the conventional MFCC and VAE. We use VAE to improve the performance of feature extraction for denoising the features and showing generalization under noisy conditions. Here is the flow of the research paper. Section II presents the related work on the conventional system. Section III describes the proposed MFCC-VAE framework, while Section IV gives the details of the experiment performed. The research and analysis explained in section V and finally section VI covers the part of the overall conclusion of the research.

Unlike MFCC-GAN methods, which depend on adversarial training between generator and discriminator networks, the proposed MFCC-VAE learns a smooth probabilistic latent representation through Kullback–Leibler (KL) regularization. This helps the model organize speech features in a more structured latent space while reducing the effect of background noise. In comparison, GAN-based approaches may suffer from training instability, sensitivity to parameter settings, and mode collapse, which can affect performance under different noisy conditions. The VAE framework offers a smoother latent space and simpler optimization process, leading to more stable training and better generalization for unseen noisy speech samples

II. RELATED WORK

In this paper, we studied several research papers that were published on various conventional systems presently available for feature extraction of audio signals.

MFCC

MFCC is one of the most widely used feature extraction techniques. MFCC models derived from the short-time Fourier transform (STFT). There are some important characteristics of MFCC features like high correlation with human psychoacoustics, compactness and computational efficiency. MFCC can extract features that are closely tied to human auditory perception, which makes MFCC a standard feature extraction method [7]. Strengths of MFCC are effectiveness, low dimensionality, and interpretability. These qualities make MFCC one of the reliable features extraction tools in a clean environment [8].

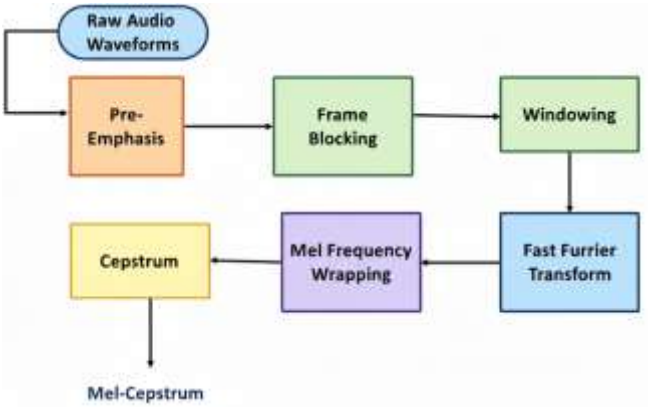


Fig 1. General block diagram of MFCC feature extraction process.

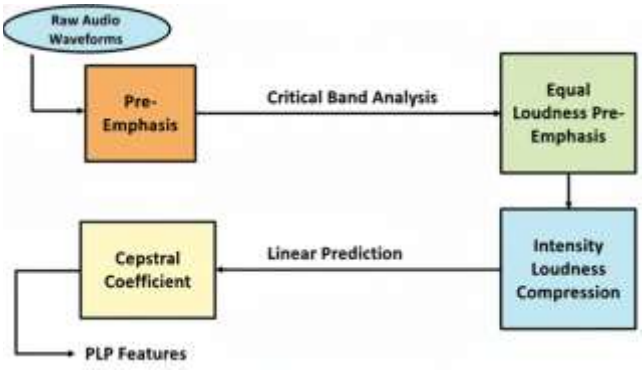


Fig 2. Block diagram of the PLP feature extraction process.

Figure 1 illustrates the processing steps involved in extracting MFCC features from the raw audio waveform. This process begins with a time-domain representation of the audio waveform. Pre-emphasis consists of passing the signal across a first-order high-pass filter. This process amplifies high frequencies, improving the signal-to-noise ratio. Frame blocking segment the signal into short overlapping frames (e.g., 20-40 frames) as it is necessary for the Fourier transform. Each short, isolated frame is multiplied by the window function to reduce spectral leakage from the signal. Windowed frames were converted from the time domain to the frequency domain by using the fast fourier transform (FFT). This conversion resulted in a magnitude spectrum of the signal. This magnitude spectrum shows a distinct energy distribution across different frequencies. The magnitude spectrum is then passed through a mel filter bank, which contains a set of triangular bandpass filters. This models non-linear perception of pitch by the human auditory system. The inverse discrete cosine transform is then applied to the output of band band-pass filter. This yields more compact and effective MFCC feature set [9].

Perceptual Linear Prediction (PLP)

PLP is considerably enhances the robustness of the features by inducting principles of human auditory perception. These robust features serve as a significant
base for a speech recognition system for various precise signal processing terminology in diverse acoustic environments. PLP is designed to suppress speaker specific variations while preserving phonetically relevant information [10].
Figure 2 shows the design of PLP features extraction system. Phases start with the time domain audio signal. Pre-emphasis involves passing the signal through a first-order high-pass filter to flatten the signal spectrally. Critical band analysis is the first key perceptual phase. The spectrum is divided into critical bands, and the total energy in every band is calculated. The equal loudness pre-emphasis phase simulates a nonlinear relationship between sound frequency and perceived loudness in the human auditory system. The linear prediction phase is for the calculation of the autocorrelation coefficient from the spectrum. The LPC coefficients are transformed, and the resulting features are the PLP features [11][12].
Linear Predictive Coding (LPC)

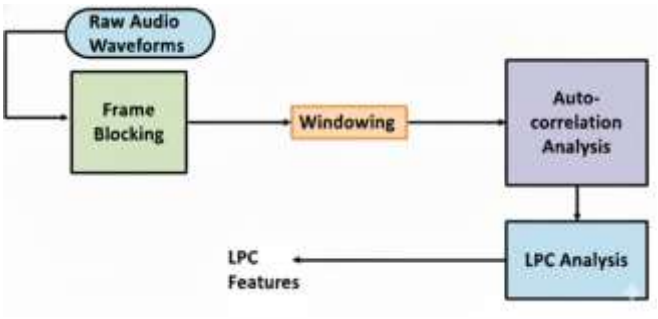


Fig 3. Block diagram LPC feature extraction process.

LPC is highly efficient methodology for modeling the physical process of speech signal. The primary advantage of the LPC lies in its extreme data composition. It typically requires only 8 to 10 coefficients per frame to offer a compact representation of features [13].
Figure 3 illustrates different phase of the LPC feature extraction system. The feature extraction process begins with a time-domain representation of the audio signal. The frame blocking phase involves the segmentation of the continuous audio signal into short frames. Every short-isolated frame gets multiplied by the window function at the windowing phase as it reduces the spectral leakages. The autocorrelation coefficient of the windowed frame is calculated. This Autocorrelation function measures the similarity between the signal and time-delayed versions of itself. LPC coefficients can be computed using the Levinson-Durbin algorithm. [14] [15].

GAN-MFCC-based feature extraction
The adversarial training is used to enhance the feature by using this method. Features are extracted with conventional MFCC and then enhanced by denoising them with the help of a GAN architecture. The feature passes through adversarial neural network training, enhanced features were very close to the original one [16]. Figure 4 illustrates the different components of the GAN and the MFCC-based feature extraction system. This hybrid architecture is designed to enhance the features extracted from raw audio waveforms by using MFCC.

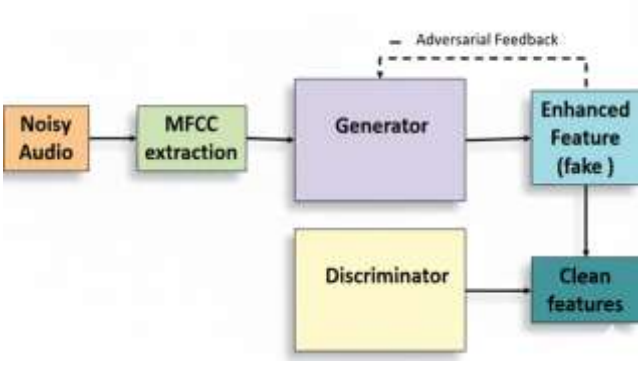


Fig 4. Block diagram of the MFCC-GAN feature extraction process.

The noisy MFCC features get enhanced by learning a mapping from the noisy and clean representations. The architecture combines two neural network generators and a discriminator. The generator learns a nonlinear mapping from noisy MFCC features to clean feature representations. The discriminator is used to distinguish real from fake. The system takes as input MFCC features extracted from noisy audio waveforms. These features serve as a benchmark representation of the speech signal, capturing both spectral and temporal information. The generator network used to produce an enhanced feature vector that's very close to a clean one. The generator learns a nonlinear transformation that maps noisy feature inputs to a clean representation [17].
The discriminator is used as a binary classifier that discriminates between the real clean and generated feature vectors. It's implemented as a fully connected neural network. The generator is progressively trained to denoise MFCC features, so that the discriminator ensures perceptual realism [18].

summary and conclusion of the literature survey

TABLE 1 SUMMARY OF LITERATURE REVIEW

Feature Extraction Method	Core Strengths	Technical Limitations	Spectral Representation	Primary Application	Future Research Scope
MFCC	Low computational overhead Well-validated in clean speech environments	Sensitive to additive noise Loss of fine-grained spectral details	Mel-scale wrapped Power Spectrum	Standard ASR & Speaker Identification	Multi-resolution filter banks Hybrid sets
PLP	Enhanced robustness at low SNR Effective channel effect suppression	Dependent on sampling and filter bank configuration	Equitoxic (Bark) scale modeling	Robust Speech Recognition & Telephony	Adaptive perceptual modeling
LPC	Efficient vocal-tract characterization Optimized for coding and data compression	Lacks auditory perceptual modeling Sub-optimal for modern high-noise ASR	Source-Filter Linear Model	Speech Coding & Synthesis (TTS)	Hybrid LPC-Spectral integration
MFCC-GAN	Mimics complex human auditory perception High sensitivity to environmental nuances	High computational cost (GPU intensive) Training instability in high-noise variants	Latent Generative Distribution	Audio Enhancement & De-reverberation	Multi-task GAN architectures

After conducting a comprehensive survey of conventional feature extraction systems, such as MFCC, PLP, and LPC. As per Table I, it is observed that these methods perform efficiently and accurately, with a higher performance rate, in a clean environment. These conventional methods exhibit low performance in noisy, high-background environments and a distorted setting; their efficiency consistently degrades and fails to meet expectations. Every conventional feature extraction system has its own benefits and limitations. MFCCs are sensitive to noisy datasets and variations in the distinguishing channels. PLP performs better and shows noticeable improvement, but it still has issues and struggles to perform under low SNR conditions. LPC exhibits good efficiency in capturing formant structure, but remains highly noisy and sensitive, with a limited ability to learn nonlinear dependencies. So, some of these limitations motivate us to research a feature extraction system and propose a hybrid system that not only overcomes these limitations but also ensures robustness and efficiency in a noisy environment.

Recently, scholarly inquiry in the domain of speech processing has delved into secure and resilient signal representation methodologies that transcend traditional feature extraction techniques. Rani et al. introduced an innovative speech steganography framework predicated on the Shift Invariant Continuous Wavelet Transform (SICWT), in conjunction with mechanisms for speech activity and message detection. Their proposed architecture illustrated that representations within the wavelet domain can proficiently maintain critical speech attributes while enhancing both robustness and security during the information embedding process. Although the principal aim of their investigation was the safeguarding of speech data rather than the facilitation of speech recognition, the research underscores the significant role of resilient signal representations under challenging conditions. Inspired by such advancements, the current study is dedicated to augmenting the robustness of speech features via probabilistic latent representations, utilizing a Variational Autoencoder that is integrated with Mel-Frequency Cepstral Coefficients (MFCC) features [27].

I. PROPOSED FEATURE EXTRACTION SYSTEM.

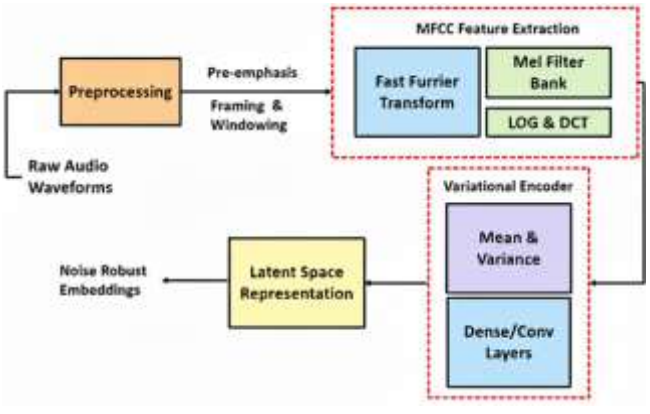


Fig 5. Proposed MFCC-VAE architecture of feature extraction

In this study, we propose a hybrid approach by combining MFCC and a Variational autoencoder. Figure 5 illustrates the detailed block diagrams of the MFCC+VAE architecture. In this architecture, MFCC provides auditorily inspired representation, while the variational autoencoder learns robust latent embedding. The raw audio input waveform is used as input. These inputs may contain background noise or some distortion. These input waveforms act as the starting phase of feature extraction [19] [20]. In the preprocessing phase, the original audio sample is converted into multiple formats. In pre-emphasis filters, higher-level frequencies are boosted to balance spectral tilt. The windowing and framing divide speech into different short frames and also ensure that the signal remains stationary. The next important stage is MFCC extraction, where the fast fourier transform

converts a time-domain signal into a frequency spectrum. These frequency spectra are applied to the Mel filter bank, where they are mapped to the Mel scale to mimic the human auditory system. Log compression firstly emphasizes perceptually relevant information and then stabilizes the variance. The role of DCT is to produce compact MFCC coefficients, i.e., a feature vector. These feature vectors are then used as input to the VAE encoder. In the VAE encoder, feature vectors pass through dense or convolutional layers and produce the output in the form of mean (**μ**) and variance ($\boldsymbol{\sigma}^2$). These outputs are then used as parameters for the latent distribution. Then, the reparameterization trick can be employed with differentiable sampling, as commonly used in Variational Autoencoder models [21].

$$Z = \mu + \sigma \Theta \ (\varepsilon \approx N(0,1)) \qquad (1)$$

In equation (1) Z represents the latent features, which combine MFCC features with the generative capability of VAE. These latent embeddings exhibit strong robustness to noise and acoustic distortion.

III. Experimental Setup.

TABLE II EXPERIMENTAL SETUP FOR THE PROPOSED FEATURE EXTRACTION SYSTEM.

Category	Parameter	Specification / Value
Acoustic Features	Input Representation	MFCC (Static + Δ + ΔΔ)
	Feature Dimensionality	60 (20 per stream)
	Windowing	25 ms Frame, 10 ms Hop (Hamming)
Model Architecture	Latent Space (z)	64-dimensional (Isotropic Gaussian)
	Objective Function	L=MSErecon+β·DKL
	KL Weight (β)	1×10−4
Training Protocol	Optimizer	Adam (β1=0.9, β2=0.999)
	Initial Learning Rate	1×10−3
	Batch Size	32
	Training Epochs	30
Data Partitioning	Dataset Source	https://doi.org/10.7488/ds/2117.
	Distribution	80% Train, 10% Val, 10% Test

Table II summarizes experimental setup adopted in this study. The dataset consists of 11752 clean files and the same number of corresponding noisy files sampled at 16 KHz. For consistency preservation, the mirror folder structure was maintained. The extracted features (.npy) were saved in separate clean and noisy folders. We applied MFCCs with their first-order (Δ) and second-order (ΔΔ) derivatives, resulting in 60 features per frame for feature extraction. Extraction parameters were fixed as follows: n_mfcc=20, n_fft=400, hop_length =160, and win_length = 400. At the utterance level, cepstral mean and variance normalization were applied to improve the robustness. The metrics used for validation were Mean Absolute Error (MAE), Mean Squared Error (MSE), Median error, and Standard deviation (Std). Adam optimizer was used for training the proposed model. The learning rate was set to 0.001 and training was performed for 30 epochs with batch size of 32. The latent dimension was fixed at 64 to obtain a compact yet expressive representation. To balance the contribution of KL divergence against reconstruction loss, a 'KL weight of 1e-4' was applied. Kullback–Leibler (KL) divergence is playing a vital role in VAE for a feature extraction system. It is useful in measuring whether one probability distribution differs from another.

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} \, d(x) \qquad (2)$$

Where:

- P(x)= true or approximate posterior distribution
- Q(x)= prior distribution (often a standard normal distribution $\mathcal{N}(0, I)$)

The Kullback–Leibler divergence measures the difference between the approximate posterior and prior distributions, and is widely used for latent-space regularization in VAE frameworks [21], [22].

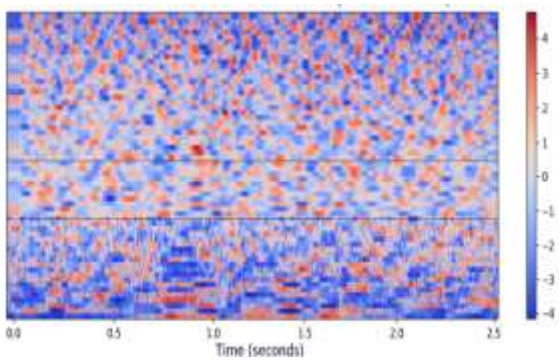


Fig. 6. MFCC feature extraction result (the color scale represents normalized cepstral coefficient magnitude, dimensionless, with values ranging from −4 to +4).

If P and Q are identical then

$$D_{KL}(P||Q) = 0$$

Hence, The Kullback–Leibler divergence regularizes the latent space by enforcing a smooth and continuous distribution, thereby improving generalization.

IV. Result and Discussion.

This section of the research paper presents the experimental results as well as a comparative analysis of MFCC, PLP, LPC, and the proposed MFCC+VAE model within the same experimental setup.

TABLE III. SUMMARY OF MFCC FEATURE EXTRACTION

Dataset Category	Clean Speech	Noisy Speech
Total Samples	11,572	11,582
Total Duration (hrs.)	14.2	14.3
Avg. SNR (dB)	> 35	5 – 15
Sampling Rate (kHz)	16	16
Feature Dim (per frame)	39 (13+Δ+ΔΔ)	39 (13+Δ+ΔΔ)
Extraction Success (%)	100%	100%

Table III illustrate the summary of feature extraction. The dataset comprises over 11,500 clean and noisy speech samples, each with approximately 14 hours of speech data recorded at 16 kHz. MFCC features with 39 dimensions (13 static, delta, and delta-delta coefficients) were successfully extracted from all samples. The noisy speech dataset contains challenging conditions with SNR values between 5–15 dB, while the clean speech dataset maintains an SNR above 35 dB, providing a comprehensive benchmark for evaluating speech enhancement and feature extraction methods.

X-axis of figure 6 represents the time frames across the duration of the speech sample, while the Y-axis represents the MFCC coefficient along with the delta and delta-delta coefficient. It includes 13 static MFCCs, 13 Delta, and 13 Delta-Delta for a total of 39 channels. On the right-hand side, a color bar represents the magnitude of the extracted coefficients, with a scale ranging from -4 to 4. Higher positive energy in the specific coefficient can be observed in red regions. The blue region represents lower energy or negative values, indicating suppression in specific frequency bands. The combination of features (MFCC + Δ + ΔΔ) provides a robust representation against a noisy and distorted environment. This is very beneficial for the downstream models trained for speech enhancement.

Comparative Analysis

To evaluate the efficiency of the proposed hybrid feature extraction system with MFFCC + VAE, we try to compare its performance with conventional systems like MFCC, PLP, and LPC using the same experimental setup, the same dataset, and the same environmental conditions. For the evaluation of extracted features, we use different matrices as follows.

1. Feature distortion methods.

The Mean Squared Error (MSE) measures the distance between the extracted features and the ideal clean reference.

$$\text{MSE} = \frac{1}{N}\sum_{i-1}^{N}(yi - \hat{y}i)^2 \qquad (3)$$

Where:

N= total number of data points (e.g., frames or feature coefficients)

y_i= original (target) value — e.g., clean speech feature

$\hat{y}_i$= predicted or reconstructed value — e.g., enhanced speech feature

$(y_i - \hat{y}_i)^2$= squared error for each data point

Equation (3) shows the formula for calculating the MSE value. In the speech enhancement model least MSE values mean model effectively reduces the noise from audio waveforms while preserving the original structures of speech. Mean Squared Error (MSE) is a standard regression-based reconstruction metric widely used for measuring distortion between predicted and target features [4], [18]. Table IV shows comparisons feature extraction system with MSE. The

Mean Squared Error (MSE) decreases progressively from 0.185 for LPC to 0.098 for the proposed MFCC-VAE method, indicating a substantial reduction in feature reconstruction distortion. MFCC-VAE achieves the lowest MSE with a 47.0% improvement over the baseline, demonstrating its superior ability to preserve important speech characteristics while suppressing noise. Figure 7 shows how different acoustic feature extraction systems work using distortion spectrograms and error maps compared to the original reference. The top part of Figure 7 is the clean reference, the standard used to check how well features are being reconstructed.
In the middle, you see what features look like after being processed by MFCC, MFCC-GAN, and the new MFCC-VAE method. This section also includes their Mean Squared Error (MSE) values. Lower MSE means the features match the clean reference better and have less distortion. The bottom part is about the absolute error maps, showing differences

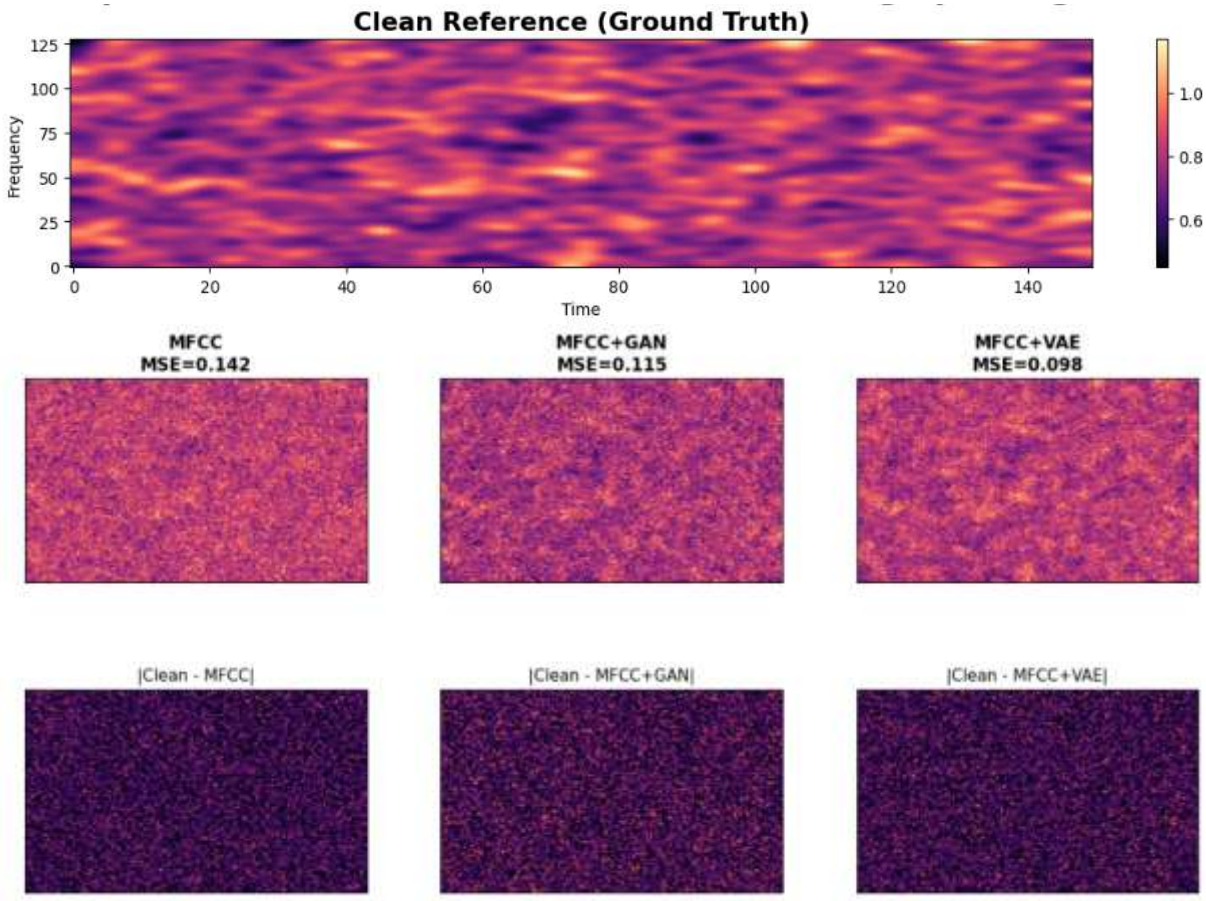


Figure 7. Comparative analysis of acoustic feature extraction architectures evaluated via distortion spectrograms and discriminative capacity

between the clean reference and the rebuilt features. Brighter areas in these maps mean bigger errors, while darker spots show better preservation of original speech. MFCC looks the worst, with the highest MSE of 0.142. MFCC-GAN does better, lowering the MSE to 0.115 thanks to its adversarial training approach. But the proposed MFCC-VAE scores the lowest MSE at 0.098, making its error map the darkest and its results closest to the clean reference. This confirms that the MFCC-VAE system is great at learning hidden speech data that cuts down noise while keeping necessary info. Compared to regular MFCC, it cuts distortion by 30.99%, and compared to MFCC-GAN, that number is 14.78%. So, the new method is clearly better for getting reliable speech features even when it's noisy.

2. Clustering Quality

The Silhouette Score (ranging from -1 to 1) evaluates the quality of the feature space. A higher score indicates that speech features belonging to the same class are tightly grouped and well-separated from other classes.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (4)$$

Where:

a(i)= average distance of point ito all other points in the same cluster (intra-cluster distance)

b(i)= average distance of point ito points in the nearest neighboring cluster (inter-cluster distance)

Silhouette Score = mean of all s(i)

Equation (4) shows how to calculate Silhouette Score

The Silhouette Score is a commonly used cluster validation metric for evaluating intra-class compactness and inter-class separation [23].

The Silhouette Score ranges between -1 and +1:

$$-1 \leq s(i) \leq 1$$

- **+1** → Maximum value: Perfect clustering
- **0** → Neutral: Clusters overlap or boundaries are fuzzy
- **–1** → Worst: Misclassified samples or wrong cluster structure

So, if Silhouette Score is close to +1, each point lies very near to the center of its own cluster. Each point is far from the other clusters. Clusters are compact and well separated. The maximum values of the silhouette score have the highest clustering quality.

As per the table IV The Silhouette Score increases from 0.52 for LPC to 0.69 for MFCC-VAE, indicating improved cluster compactness and separation among speech classes. The proposed MFCC-VAE achieves the highest silhouette gain of 32.7%, confirming that its learned latent representations are more discriminative and better organized than those produced by conventional feature extraction methods.

TABLE IV.

COMPARISON OF FEATURE EXTRACTION SYSTEMS WITH MSE AND SILHOUETTE SCORE VALUES

Feature Extraction System	Feature Distortion (MSE) ↓	MSE Improvement (%)	Silhouette Score ↑	Silhouette Gain (%)
LPC	0.185	Baseline	0.52	Baseline
PLP	0.158	14.60%	0.58	11.50%
MFCC	0.142	23.20%	0.61	17.30%
MFCC + GAN	0.115	37.80%	0.64	23.10%
MFCC + VAE (Proposed)	0.098	47.00%	0.69	32.70%

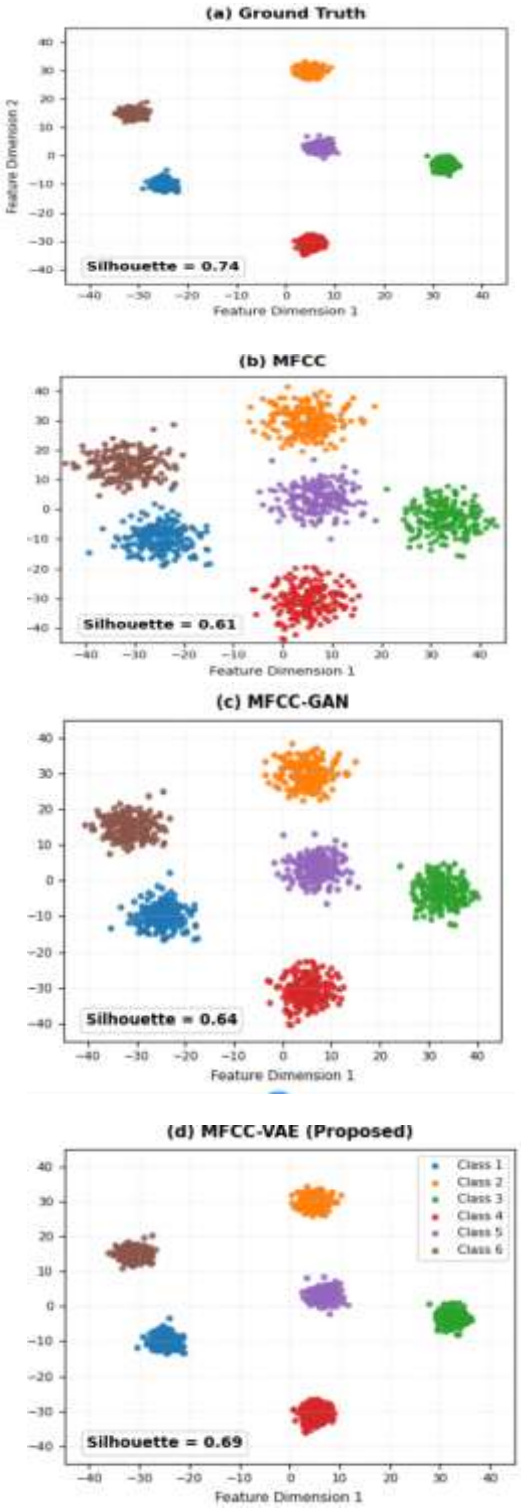


Fig 8 Visualization of Extracted Feature Spaces via t-SNE Clustering

The purpose of this figure 8 is to conduct a comparative analysis of the clustering efficacy of various methodologies employed for feature extraction from the auditory recordings of three distinct vocal sources (namely MFCC, MFCC-GAN, and the novel approach MFCC-VAE), as demonstrated through t-SNE visualizations. In order to assess the three feature extraction methodologies, we examine the clarity with which they organize into distinct clusters. The MFCC features display a moderate clustering capability, as indicated by the presence of overlapping clusters and significant within-class variability, which yields a silhouette score of 0.61 for this method. Conversely, the MFCC-GAN features demonstrate enhanced within-class clustering and diminished between-class clustering, resulting in an elevated

silhouette score of 0.64, attributable to an improved representation of the feature clusters. The features generated by MFCC-VAE exhibit the highest quality, characterized by tightly-knit, well-separated
clusters devoid of overlap, thus achieving the highest silhouette score of 0.69. This enhancement can be ascribed to the VAE's proficiency in learning distributions that encapsulate the essential attributes of speech while simultaneously mitigating the presence of extraneous noise and irrelevant factors inherent in the signal. Hence, MFCC-VAE delivers features that are more differentiable, stable, and unaffected by noise when pitted against MFCC and MFCC-GAN, culminating in enhanced classification results in the speech recognition

3. Fisher score evaluation

The Fisher score serves as primary metrics for evaluating the discriminative power of extracted feature set. The Fisher score measures how effectively feature separate different classes, e.g., men vs women, clean vs noisy, different phonemes. Higher fisher score values indicate stronger class separability. A score between 0.1 to 0.3 means features with moderate discriminability (useful features but not very strong). If the score is at a low value (0.3>), it indicates weak discriminability.

$$Fj = \frac{\sum_{c=1}^{c} n_c(\mu_{cj} - \mu_j)}{\sum_{c=1}^{c} n_c \, \sigma_{cj}^2} \qquad (5)$$

Where:

C= number of classes (e.g., 2 for male/female)
n_c= number of samples in class c
μ_{cj}= mean of feature jin class c
μ_j= global mean of feature jacross all classes
σ_{cj}^2= variance of feature jin class c

Equation (5) describes how to calculate the Fisher Score of a given feature (j).
Fisher Score is a supervised feature selection criterion used to measure discriminative power between classes [24].

TABLE V COMPARISON OF AVERAGE FISHER SCORES FOR DIFFERENT FEATURE EXTRACTION METHODS

Feature Extraction Architecture	Avg. Fisher Score (SF) ↑	Discriminative Gain (%)	Rank
LPC	0.37	Baseline	5
MFCC	0.43	16.20%	4
PLP	0.49	32.40%	3
MFCC + GAN	0.69	86.50%	2
MFCC + VAE (Proposed)	0.82	121.60%	1

Table V describes the detailed Comparison of Average Fisher Scores for Different Feature Extraction Methods. The proposed MFCC+VAE shows the maximum average Fisher score value. As shown, the proposed MFCC+VAE achieves the lowest MSE value (0.098), which indicates features having minimal distortion and preserving the originality in speech characteristics. Traditional methods like LPC (0.37) and MFCC (0.43) exhibit relatively low Fisher Scores. This suggests that in noisy environments, these features suffer from high intra-class variance. The representations of the same phoneme become scattered with other classes that leads to classification errors. MFCC + GAN framework provides a substantial improvement (0.69), it is outperformed by the Proposed MFCC + VAE. The VAE framework achieves a peak Fisher Score of 0.82, representing a 121.60% gain over the LPC baseline.
Figure 9 delineates the comparative distributions of acoustic features derived from the MFCC, MFCC-GAN, and the novel MFCC-VAE framework, in conjunction with the pristine reference signal. The traditional MFCC representation displays a relatively dispersed feature distribution, signifying a constrained discriminative efficacy in the presence of noise. The MFCC-GAN methodology enhances the organization and spatial coherence of features by acquiring improved representations through adversarial learning techniques. Conversely, the proposed MFCC-VAE yields a feature distribution that closely parallels the clean reference, thereby illustrating its capacity to acquire robust and noise-resistant latent representations. The enhanced spatial organization and consistency of the extracted features are further corroborated by a Fisher Score of 0.82, thereby validating the superior discriminative capability of the proposed methodology.

4. Evaluation through the PESQ value

PESQ (Perceptual Evaluation of Speech Quality) is an assessment metric for evaluating speech quality.

$$\mathrm{PESQ}(y,\hat{y}) = \alpha.\mathrm{D}_{\mathrm{sym}}(y,\hat{y}) + \beta.\mathrm{D}_{\mathrm{asym}}(y,\hat{y}) + \gamma \quad (6)$$

Where:

- $y(t)$— clean (reference) speech signal
- $\hat{y}(t)$— degraded or enhanced speech signal
- $\mathrm{D}_{\mathrm{sym}}$— symmetric disturbance (perceptual difference common to both ears)
- $\mathrm{D}_{\mathrm{asym}}$— asymmetric disturbance (perceptual difference due to added or missing components)
- α, β, γ— empirically determined coefficients according to ITU-T P.862
- For wideband speech: $(\alpha = -0.1, \beta = -0.0309, \gamma = 4.5)$

The PESQ score typically lies in the range:

$$1.0 \leq \mathrm{PESQ} \leq 4.5$$

Equation (6) shows how the PESQ value is calculated. The PESQ value near 1 is considered as poor speech quality, and its value close to 4.5 is considered as nearly identical to the original audio waveform. PESQ is standardized by ITU-T Recommendation P.862 for perceptual speech quality assessment [25].

5. Evaluation through the STOI value

STOI is another assessment metric used for the evaluation of the intelligibility of enhanced speech. STOI works by time-aligning clean and degraded audio signals, also framing them in short-time overlapping segments. The STOI course lies between 0.0 (unintelligible) and 1.0 (perfect intelligibility).

$$\mathrm{STOI}(x,\hat{x}) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Corr}(x_n, \hat{x}_n) \quad (7)$$

Where:

- x= clean reference speech
- $\hat{x}$= degraded or enhanced speech
- $\mathbf{x}_n, \hat{\mathbf{x}}_n$= short-time spectral envelope vectors in the n-th segment
- Corr = linear correlation coefficient
- N= total number of segments

Equation (7) describes how the STOI value is calculated. If the STOI value is closer to 1, the speech enhanced is intelligible perfectly. Similarly, if the STOI score is near 0, the speech sample is unintelligible. STOI is a widely accepted objective intelligibility metric for enhanced speech signals [26].

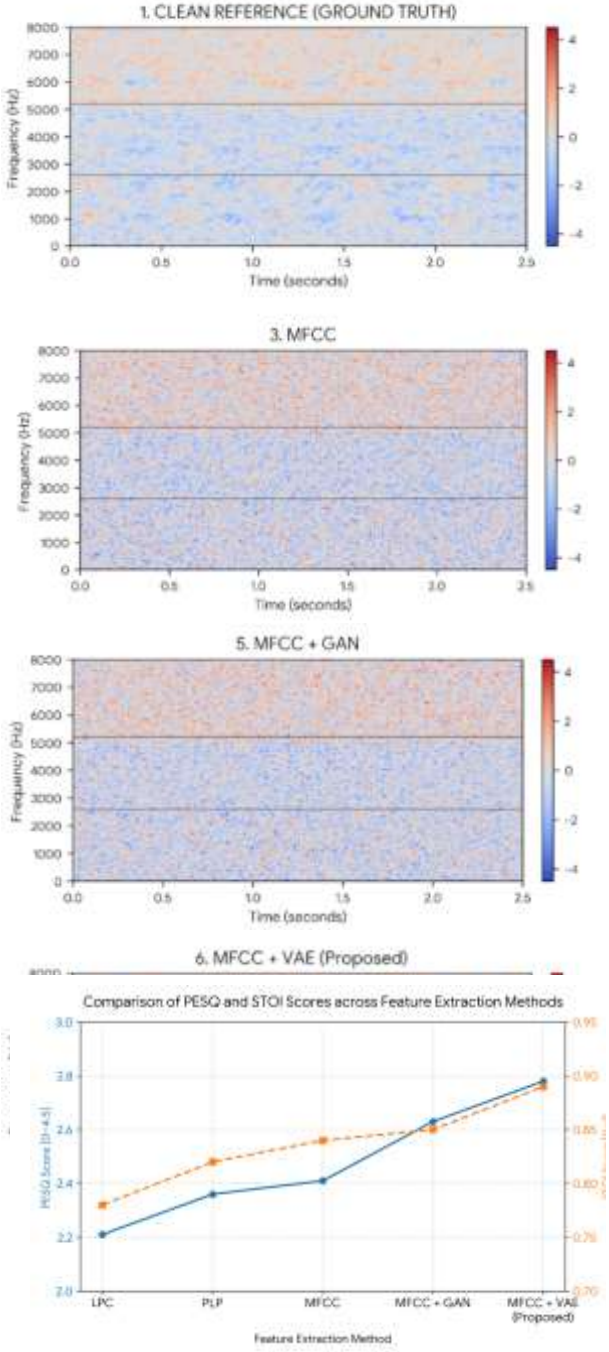


Fig. 10 Comparative PESQ and STOI scores

TABLE VI COMPARISON OF FEATURE EXTRACTION SYSTEMS WITH PESQ AND STOI VALUES

Feature Extraction Method	**PESQ Score [0–4.5]**	**ΔPESQ (vs. MFCC)**	**STOI Score [0–1]**	**ΔSTOI (vs. MFCC)**
LPC	2.21	-0.2	0.78	-0.06
PLP	2.36	-0.05	0.82	-0.02
MFCC	2.41	Baseline	0.84	Baseline
MFCC + GAN	2.63	0.22	0.85	0.01
MFCC + VAE (Proposed)	2.78	0.37	0.89	0.05

Table VI evaluates the proposed framework using two industry-standard perceptual metrics: Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI). Traditional methods like LPC (2.21) and PLP (2.36) underperform compared to the MFCC baseline (2.41), confirming that standard MFCC remains superior for general perceptual features in low-noise settings. The Proposed MFCC + VAE achieves a score of 2.78, yielding a ΔPESQ of 0.37 over the MFCC baseline. In speech enhancement literature, an improvement of > 0.1 in PESQ is considered statistically significant. The 0.37 gain suggests that the VAE's ability to model the speech manifold effectively "reconstructs" missing spectral components that are vital for human hearing. The MFCC + VAE achieves the highest STOI score of 0.89, a 0.05 improvement over the baseline and achieve intelligibility benchmark. The perceptual efficacy of the proposed MFCC + VAE framework is summarized in Table VI. The system achieved a PESQ score of 2.78 and a STOI score of 0.89, outperforming the standard MFCC baseline by 15.3% and 5.9% respectively. Notably, the VAE-based approach surpassed the GAN-based enhancement in both quality and intelligibility.
Figure 10 shows the proposed MFCC + VAE method achieves the highest speech quality and intelligibility, significantly outperforming traditional methods and the standard MFCC baseline. While generative models consistently improve performance, the VAE-enhanced approach delivers the most substantial gains with a PESQ of 2.78 and a STOI of 0.89.

TABLE VII EFFECT OF VAE AND LATENT SPACE SIZE ON FEATURE QUALITY

Architecture	**KL Loss (β)**	**Latent Size (d)**	**MSE**	**Silhouette**	**Fisher Score**	**PESQ**	**STOI**
MFCC (Baseline)	—	—	0.142	0.61	0.43	2.41	0.84
MFCC-AE	No	64	0.118	0.64	0.61	2.63	0.86
MFCC-VAE	Yes	16	0.11	0.65	0.71	2.67	0.87
MFCC-VAE	Yes	32	0.104	0.67	0.77	2.72	0.88
MFCC-VAE (Optimum)	**Yes**	**64**	**0.098**	**0.69**	**0.82**	**2.78**	**0.89**
MFCC-VAE	Yes	128	0.101	0.67	0.78	2.74	0.88

6. Ablation study

An ablation study was conducted to analyze the impact of each design component in proposed MFCC-VAE framework. In this study individual elements like KL divergence term and latent space dimensions were systematically varied and observation were recorded at the end of experiment. The table VII summarizes the effect of each modification, demonstrating the necessity of VAE architecture and chosen latent space dimensions for the performance that is very near to optimal. MFCC is raw feature set it act as “floor” for performance. MFCC-AE is standard autoencoder that compresses data but doesn’t really use probabilistic bottleneck (with no KL loss). MFCC-VAE uses KL loss to even up the latent space. This making it more structured and continuous. The proposed model achieved the lowest MSE value (0.098). This observation indicates that the VAE can compress the audio features and reconstruct them with least amount of information loss. The Silhouette Score (0.69) and
Fisher Score (0.82) are maximum for the proposed model. This indicates that VAE is better at clustering identical speech sounds together and separating odd ones in latent space. Which is very effective in tasks like speech recognition and speaker identification. The proposed model meets the highest PESQ (2.78) and STOI (0.89) scores, suggesting that the features extracted by this model retain the most "human-
audible" clarity.
When the latent size is underfitting (16 and 32) i.e. too small, the model cannot capture enough details so as observations shows higher MSE values and lower quality scores. When the latent size is 64 it provides optimum results, the perfect balance between compression and details. When the latent size is overfitting (128), performance drops (MSE rises to 0.101 and Silhouette drops to 0.67). This may introduce noise or lose the benefit of KL divergence

7. Final result and analysis Summary

8. The Table VIII summarizes following observations

TABLE VIII RESULT AND ANALYSIS SUMMARY

Method	MSE	Silhouette	Fisher	PESQ	STOI
LPC	0.19	0.52	0.37	2.21	0.78
PLP	0.16	0.58	0.49	2.36	0.82
MFCC	0.14	0.61	0.43	2.41	0.84
MFCC-GAN	0.12	0.64	0.69	2.63	0.85
MFCC-VAE	**0.1**	**0.69**	**0.82**	**2.78**	**0.89**

Reconstruction fidelity (MSE): The proposed model meets significant reduction in Mean Square Error (0.098). It clearly outperforms the baseline MFCC by 30% and MFCC-GAN by 14%. This representing significant improvement over conventional methods.

Enhanced Feature Identification: The high Silhouette (0.69) and Fisher (0.82) score indicates that VAE create more organized and distinct latent space i.e. better clustering of similar features

Optimal Human Perception: The leading scores in PESQ (2.78) and STOI (0.89) confirmed that proposed method maintains the better speech quality and intelligibility

The primary reason for significant performance spike observed in result is due to integration of VAE into feature extraction pipeline. The VAE learns a structured, probabilistic representation of speech instead of just extracting static coefficient.

VI. CONCLUSIONS.

After executing the MFCC + VAE model for feature extraction successfully, we perform a comprehensive comparison of the results with conventional feature extraction systems like MFCC, PLP, LPC and MFCC-GAN. Comparative results indicate that our proposed model (MFCC+ VAE) performs exceptionally well and outperforms the rest. In terms of evaluation metrics, the proposed model shows suppression with strong noise. While PESQ and STOI values are based on reconstruction, they also show higher-level readings. The lowest MSE values and higher Silhouette score indicates noise-invariant and discriminative latent features. Alone, MFCC, PLP, LPC and MFCC-GAN show improvement and robustness in a clean environment, but the performance and robustness get compromised in noisy and distorted conditions due to several reasons, such as high computational cost, higher sensitivity to noise, and low clustering quality in a noisy environment. The integration of the variational autoencoder significantly enhances feature robustness and discriminability. Although the proposed model introduces moderate computational overhead, it remains suitable for next generation intelligent edge devices.

Data Availability Statement

Noisy speech database for training speech enhancements in this study is publicly available at https://datashare.ed.ac.uk/handle/10283/2791

Additional processed features, noisy-clean speech pairs generated in the current study, are available from the corresponding author upon reasonable request.

References

[1] Labied, M., & Belangour, A. (2021). Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison. International Journal of Advanced Computer Science and Applications, 12(8). https://doi.org/10.14569/IJACSA.2021.0120821

[2] Sadique, U., Anwar, S., & Ahmad, M. (2023). Machine Learning based human recognition via robust Features from audio signals. 52–57. https://doi.org/10.1109/ICAI58407.2023.10136683

[3] Luis Lugo and Valentin Vielzeuf. 2024. Towards efficient self-supervised representation learning in speech processing. In Findings of the Association for Computational Linguistics: EACL 2024, pages 340–346, St. Julian's, Malta. Association for Computational Linguistics.

[4] Lee, Y., & Jung, K. (2024). Boosting Speech Enhancement with Clean Self-Supervised Features Via Conditional Variational Autoencoders. https://doi.org/10.1109/icassp48485.2024.10447220.

[5] Chauhan, N.; Isshiki, T.; Li, D. Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. Acoustics 2024, 6, 439-469. https://doi.org/10.3390/acoustics6020024.

[6] Hidayat, R., Bejo, A., Sumaryono, S., & Winursito, A. (2018). Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System. International Conference . on Information Technology and Electrical Engineering, 280–284. https://doi.org/10.1109/ICITEED.2018.8534807.

[7] Gourisaria, M.K., Agrawal, R., Sahni, M. et al. Comparative analysis of audio classification with MFCC and STFT features using machine
learning techniques. Discov Internet Things 4, 1 (2024). https://doi.org/10.1007/s43926-023-00049-y.
[8] Varma, V., & Saravanan K, A. (2023). Advancements in Speaker Recognition: Exploring Mel Frequency Cepstral Coefficients (MFCC) for Enhanced Performance in Speaker Recognition. International Journal For
Science Technology And Engineering, 11(8), 88–98. https://doi.org/10.22214/ijraset.2023.55124
[9] Chauhan, N.; Isshiki, T.; Li, D. Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. Acoustics 2024, 6, 439-469. https://doi.org/10.3390/acoustics6020024.
[10] Bernard, M., Poli, M., Karadayi, J. et al. Shennong: A Python toolbox
for audio speech features extraction. Behav Res 55, 4489–4501 (2023). https://doi.org/10.3758/s13428-022-02029-6.
[11] Sneha Basak, Himanshi Agrawal, Shreya Jena, Shilpa Gite, Mrinal Bachute, Biswajeet Pradhan, Mazen Assiri, Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems, CMES - Computer Modeling in Engineering and Sciences, Volume 135, Issue 2, 2022, Pages 1053-1089, ISSN 1526-1492, https://doi.org/10.32604/cmes.2022.021755.
[12] Saldanha, J. C., & Suvarna, M. (2020). Perceptual Linear Prediction Feature as an Indicator of Dysphonia (pp. 51–64). Springer, Singapore. https://doi.org/10.1007/978-981-15-4676-1_5
[13] Këpuska, V. and Elharati, H. (2015) Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP, and Hidden Markov Model Classifier in Noisy Conditions. Journal of Computer and Communications, 3, 1-9. doi: 10.4236/jcc.. 2015.36001.
[14] Saldanha, J. C., & Suvarna, M. (2020). Perceptual Linear Prediction Feature as an Indicator of Dysphonia (pp. 51–64). Springer, Singapore. https://doi.org/10.1007/978-981-15-4676-1_5
[15] Vary, P., & Martin, R. (2023). Linear Prediction. 181–210. https://doi.org/10.1002/9781119060970.ch6
[16] Leem SG, Fulford D, Onnela JP, Gard D, Busso C. Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech. IEEE/ACM Trans Audio Speech Lang Process. 2024;32:917-929. doi: 10.1109/taslp.2023.3340603.
[17] Mohammad Reza Hasanabadi,(2023), Electrical Engineering and Systems Science > Audio and Speech Processing,"MFCC-GAN Codec: A New AI-based Audio Coding", https://doi.org/10.48550/arXiv.2310.14300
[18] Mandar Diwakar and Brijendra Gupta. "Advancing Speech Enhancement with Generative Adversarial Network-Autoencoder: A Robust Adversarial Autoencoder Approach". International Journal of Advanced Computer Science and Applications (ijacsa) 16.9 (2025). http://dx.doi.org/10.14569/IJACSA.2025.0160932
[19] Do, H. D., Tran, S. T., & Chau, D. T. (2020). A Variational Autoencoder Approach for Speech Signal Separation (pp. 558–567). Springer, Cham. https://doi.org/10.1007/978-3-030-63007-2_43
[20] Villafuerte-Lucio, D. Á. (2023). MFCC feature extraction for COPD detection.
https://doi.org/10.35429/jti.2023.27.10.1.7
[21] Yang, Z.-H. (2022). Training Latent Variable Models with Auto-encoding Variational Bayes: A Tutorial. arXiv.Org, abs/2208.07818. https://doi.org/10.48550/arXiv.2208.07818
[22] Punnoose, A. K. (2023). Speech Enhancement Using Variational Autoencoders. 1–4.
https://doi.org/10.1109/IConSCEPT57958.2023.10170608
[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.
[24] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York, NY, USA: Wiley, 2001.
[25] International Telecommunication Union, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs, ITU-T Recommendation P.862, Feb. 2001.
[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[27] Rani, S., et al., "A Novel Speech Steganography Mechanism to Securing Data Through Shift Invariant Continuous Wavelet Transform with Speech Activity and Message Detection," Engineering Science, DOI: 10.30919/es950.