



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85075>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Regional Indian English Accent Identification via Hybrid Acoustic-Contextual Speech Representations and Hierarchical HuBERT Layer Probing

Shivapuja Sai Kiran Goud¹, Varsha Thakur², Sompalli Naveen Kumar³, Vadeghar Nitin⁴, Shivakumar Nethani⁵

^{1, 2, 3, 4}Undergraduate Students, Hyderabad Institute of Technology and Management, Hyderabad, Telangana, India

⁵Assistant Professor, Hyderabad Institute of Technology and Management, Hyderabad, Telangana, India

Abstract: Regional Indian English exhibits intricate acoustic, phonological, and prosodic variations shaped by speakers' native language (L1) backgrounds. Automated identification of these regional accents is essential for accent-adaptive automatic speech recognition (ASR), spoken language identification, and culturally nuanced voice interfaces. This paper presents a hybrid speech representation framework for six-class regional Indian English accent identification using the IndicAccentDB corpus (8,116 recordings across Andhra/Telugu, Gujarat/Gujarati, Jharkhand/Eastern Hindi, Karnataka/Kannada, Kerala/Malayalam, and Tamil). The proposed pipeline fuses low-level spectral summary statistics derived from 13 Mel-Frequency Cepstral Coefficients (MFCCs; 26 dimensions capturing temporal mean and standard deviation) with deep contextual embeddings extracted from the frozen, pretrained HuBERT-Base encoder (768 dimensions via global temporal average pooling), yielding a 794-dimensional hybrid representation. The fused representations are evaluated using multinomial Logistic Regression with L_2 regularization and a lightweight one-dimensional Convolutional Neural Network (1D CNN). Under a stratified 80/20 utterance-level evaluation protocol, the framework achieves 90.33% test accuracy with Logistic Regression and 90.83% with the 1D CNN. A complementary layer-wise probe across all 13 HuBERT representations on a 500-utterance diagnostic subset reveals that shallow layers (Layers 1–2) yield peak linear separability of 94.00%, consistent with low-level acoustic and vocal tract encoding, before declining to 88%–91% in deeper contextual layers. Temporal duration analysis shows a 25.48% accuracy drop when restricting audio to 0.7-second sub-utterance clips. The trained classifier is deployed in an interactive Gradio application coupling regional accent identification with a cultural culinary recommendation interface. We provide a thorough discussion of evaluation scope, including the distinction between utterance-level predictability and speaker-disjoint generalization.

Keywords: Indian English, regional accent identification, HuBERT, Mel-Frequency Cepstral Coefficients, self-supervised speech representations, layer probing, spoken native language identification, deep learning.

I. INTRODUCTION

English functions as a vital lingua franca across education, governance, enterprise, and digital services throughout India. However, Indian English is not a monolithic spoken variety; it is characterized by substantial regional diversification shaped by substrate interference from speakers' first languages (L1s) and local sociolinguistic environments [1, 2]. Distinct regional accents exhibit systematic phonological shifts, including retroflexion of alveolar stops ($/t/$, $/d/ \rightarrow [t], [d]$), varying degrees of aspiration, vowel monophthongization, syllable-timed rhythmic cadences, and distinct pitch accent trajectories [3, 15].

Accurately identifying regional accents from the acoustic waveform is essential for modern speech technologies. In automatic speech recognition (ASR), regional acoustic mismatch introduces substantial Word Error Rate (WER) degradation [15]. Upstream accent detection enables dynamic acoustic model routing, accent-specific adapter selection, and personalized acoustic normalization. Beyond ASR, regional accent identification supports spoken Native Language Identification (NLI), automated language tutoring, forensic phonetics, and localized interactive voice response (IVR) systems.

Historically, speech accent classification relied on handcrafted acoustic-prosodic feature engineering. Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency (F_0) contours, formant frequencies (F_1 – F_4), spectral tilt, and jitter/shimmer statistics provided compact and computationally efficient summaries of the speech envelope [1, 2].

While effective for localized vocal tract characterization, these low-level descriptors capture limited long-range temporal context and struggle to model the hierarchical acoustic-phonetic structures present in continuous speech.

The emergence of self-supervised learning (SSL) has transformed speech processing. Models such as HuBERT [7] and Wav2Vec 2.0 [8] learn rich contextualized speech representations directly from unannotated audio through masked prediction objectives. While self-supervised encoders capture robust acoustic-phonetic structures, fine-tuning entire multi-million parameter networks requires substantial compute and labeled data. Moreover, extracting features from frozen SSL backbones often relies solely on the final Transformer layer, potentially discarding low-level acoustic cues that reside primarily in shallow layers [10, 11].

This study presents a unified, hybrid speech representation architecture that combines compact acoustic summary statistics (26-dimensional temporal MFCC statistics) with frozen self-supervised embeddings (768-dimensional HuBERT-Base mean-pooled output), constructing a 794-dimensional representation for six-class regional Indian English accent identification. We evaluate this representation across 8,116 recordings from IndicAccentDB [4, 5], spanning Tamil, Andhra (Telugu), Karnataka (Kannada), Kerala (Malayalam), Jharkhand (Eastern Hindi), and Gujarat (Gujarati). We further conduct systematic layer-wise diagnostic probing, duration sensitivity analysis, and deploy the classifier in an interactive Gradio application.

II. RELATED WORK

A. Acoustic and Spectral Approaches in Indian English

Early investigations into Indian English accent detection primarily utilized acoustic-phonetic measurements. Gaikwad et al. [1] demonstrated that pitch, formant frequencies (F_1 , F_2), and short-time energy could differentiate regional accents in Marathi- and Hindi-influenced English. Patel et al. [2] developed context-independent Hidden Markov Model (HMM) systems trained on acoustic cues to discriminate five South Indian English dialects. Ahamad et al. [3] introduced AccentDB—a structured database of four Indian-language English accents alongside native English varieties—demonstrating that deep convolutional networks and autoencoders achieve significant separation between regional accent classes.

B. Deep Learning and the IndicAccentDB Benchmark

Darshana et al. [4] introduced IndicAccentDB, a corpus of Harvard sentence recordings across six Indian regional accents, and evaluated 1D CNN, ResNet18, ResNet50, and xResNet18 in the MARS framework using MFCC and Mel-spectrogram features. Jairam et al. [6] investigated parameter-efficient fine-tuning of the Whisper encoder on IndicAccentDB using LoRA and QLoRA, demonstrating that few-shot adaptation with only 5% trainable parameters achieves up to 94% classification accuracy. These prior studies on IndicAccentDB form the principal benchmarks against which the proposed hybrid framework is positioned.

C. Self-Supervised Representations and Layer Probing

HuBERT [7] and Wav2Vec 2.0 [8] achieve state-of-the-art results across the SUPERB benchmark [12], encoding speech at multiple levels of abstraction. Pasad et al. [10, 11] demonstrated through canonical correlation analysis and linear probing that SSL Transformer layers specialize hierarchically: shallow layers encode low-level acoustics and speaker identity; intermediate layers encode phonemes and articulatory features; deep layers encode lexical and semantic context. WavLM [9] further demonstrated that joint masked prediction and denoising improves downstream task performance. These insights directly inform our layer-wise probing methodology.

III. PROPOSED METHODOLOGY

The proposed framework consists of three sequential modules: (i) dual-path feature extraction (acoustic MFCC statistics and self-supervised HuBERT embeddings), (ii) early concatenation-based hybrid fusion, and (iii) multi-class classification. Each module is described formally below.

A. Audio Preprocessing

Every input recording $x(t)$ is standardized to single-channel (mono) audio at a sampling rate of $f_s = 16,000$ Hz. Silence at the waveform boundaries is trimmed at 25 dB below peak amplitude using Librosa:

$$\tilde{x}(t) = \text{Trim}(x(t), \theta_{\text{top_dB}} = 25 \text{ dB})$$

The conditioned waveform $\tilde{x}(t)$ is simultaneously routed to the MFCC extraction path and the HuBERT feature extraction path.

B. Acoustic Representation: Statistical MFCC Extraction

Mel-Frequency Cepstral Coefficients model the human auditory system's spectral sensitivity using logarithmically spaced filterbanks. The short-time Fourier transform (STFT) magnitude spectrum $|X(t, \omega)|$ is mapped onto M triangular Mel-filterbanks. Log-filterbank energies are decorrelated via the Discrete Cosine Transform (DCT-II) to yield $K = 13$ cepstral coefficients $c_k(t)$ per analysis frame. To produce a fixed-length feature vector invariant to variable utterance duration, we compute the temporal mean μ_k and sample standard deviation σ_k of each coefficient over all T_{frames} analysis frames:

$$\mu_k = (1/T) \cdot \sum c_k(t) \quad \text{and} \quad \sigma_k = \sqrt{(1/(T-1)) \cdot \sum (c_k(t) - \mu_k)^2}$$

The 26-dimensional acoustic descriptor is constructed by concatenating all means and standard deviations across 13 coefficients:

$$f_{MFCC} = [\mu_0, \mu_1, \dots, \mu_{12}, \sigma_0, \sigma_1, \dots, \sigma_{12}]^T \in \mathbb{R}^{26}$$

C. Self-Supervised Representation: Frozen HuBERT Encoder

The second extraction path uses the pretrained facebook/hubert-base-ls960 model [7]. HuBERT comprises a 7-layer temporal convolutional encoder followed by a 12-layer Transformer encoder with model dimension $d_{model} = 768$ and 8 attention heads per layer. The raw waveform $\tilde{x}(t)$ is passed through the convolutional encoder producing a downsampled latent sequence at a 20 ms frame rate. These latent representations pass through the 12 Transformer encoder blocks $H^{(1)}, H^{(2)}, \dots, H^{(12)}$ where $H^{(l)} \in \mathbb{R}^{(T_{SSL} \times 768)}$. The model parameters are frozen throughout all experiments to avoid the computational overhead of end-to-end gradient propagation. Global temporal average pooling collapses the variable frame sequence of the final hidden state $H^{(12)}$ to produce a fixed 768-dimensional utterance representation:

$$f_{HuBERT} = (1/T_{SSL}) \cdot \sum h_{t^{(12)}} \in \mathbb{R}^{768}$$

D. Hybrid Feature Fusion

The compact acoustic summary vector f_{MFCC} and the learned self-supervised contextual embedding f_{HuBERT} are fused via early concatenation to yield the 794-dimensional hybrid representation:

$$z = [f_{MFCC}; f_{HuBERT}] \in \mathbb{R}^{794}$$

This early fusion strategy retains both explicit spectral distribution parameters (MFCC means and standard deviations) and the deep contextualized self-supervised embeddings, providing a comprehensive acoustic-semantic feature space for downstream classification.

E. Classification Architectures

1) Multinomial Logistic Regression:

A multinomial Logistic Regression classifier with L_2 regularization serves as a robust linear baseline. The posterior probability for accent class $c \in \{1, \dots, 6\}$ given input z is:

$$P(y = c | z) = \exp(w_c^T z + b_c) / \sum_j \exp(w_j^T z + b_j)$$

Optimization uses the L-BFGS solver with a convergence limit of 800 iterations and default L_2 penalty ($C = 1.0$).

2) One-Dimensional Convolutional Neural Network:

A lightweight 1D CNN is implemented in TensorFlow/Keras as a nonlinear classifier. The 794-dimensional vector z is reshaped to $\mathbb{R}^{(794 \times 1)}$ and processed by:

- Conv1D: 128 filters, kernel size $k = 5$, stride 1, ReLU activation.
- Global MaxPooling1D: collapses spatial dimension, producing $v \in \mathbb{R}^{128}$.
- Dense: 128 units, ReLU activation.
- Dropout: drop rate $p = 0.30$.
- Softmax Output: 6 units, categorical cross-entropy loss.

Training uses the Adam optimizer ($\alpha = 0.001$), batch size 32, and 10 epochs with a 10% internal validation hold-out.

IV. EXPERIMENTAL SETUP

A. Dataset Composition and Regional Profiles

All experiments use the complete IndicAccentDB collection [4, 5]. The corpus contains 8,116 valid WAV recordings at 16 kHz, spanning six regional Indian English accent classes. Speakers recited phonetically balanced sentences from the Harvard Sentences list, providing uniform lexical content across regions. Table 1 details the distribution across accent classes.

TABLE I
INDICACCENTDB CORPUS DISTRIBUTION ACROSS REGIONAL INDIAN ENGLISH ACCENT CLASSES

Accent Class	Geographic Region	Primary L1 Substrate	Language Family	Utterances	Percentage
Tamil	Tamil Nadu	Tamil	Dravidian	1,840	22.67%
Andhra	Andhra Pradesh/Telangana	Telugu	Dravidian	1,794	22.10%
Karnataka	Karnataka	Kannada	Dravidian	1,686	20.77%
Kerala	Kerala	Malayalam	Dravidian	1,671	20.59%
Jharkhand	Jharkhand	Eastern Hindi/Bihari	Indo-Aryan	827	10.19%
Gujarat	Gujarat	Gujarati	Indo-Aryan	298	3.67%
Total	Six Regions	—	—	8,116	100.00%

As shown in Table 1, the corpus exhibits notable class imbalance: the four southern Dravidian varieties collectively account for 86.13% of recordings, whereas Gujarat constitutes only 3.67%.

B. Evaluation Protocol

For primary classification experiments, the 8,116 hybrid feature vectors are partitioned using a stratified 80/20 train/test split with `random_state = 42`, producing 6,492 training instances and 1,624 test instances. The stratified strategy maintains class proportions across both partitions. Because speaker identifiers were not explicitly disaggregated during feature extraction, this protocol constitutes utterance-level held-out evaluation. As discussed in Section 7, this does not guarantee speaker-disjoint generalization across the 19 recorded speakers in IndicAccentDB, and results must be interpreted accordingly.

C. Computational Environment

All experiments were conducted in Python 3.10 within a Google Colaboratory environment equipped with an NVIDIA T4 GPU (16 GB VRAM) and 12.7 GB system RAM. Core libraries included: librosa v0.10.1 (audio processing), transformers v4.38.2 (HuBERT feature extraction), torch v2.1.2, scikit-learn v1.2.2 (Logistic Regression), tensorflow v2.15.0 (1D CNN), and gradio v4.19.0 (deployment interface).

V. EXPERIMENTAL RESULTS

A. Principal Accent Classification Performance

Table 2 reports performance of the proposed hybrid representation across both classifiers on the 1,624 held-out test utterances, alongside single-modality diagnostic comparisons.

TABLE II
PERFORMANCE COMPARISON OF FEATURE REPRESENTATIONS AND CLASSIFIERS ON INDICACCENTDB

Feature Representation	Dimensions	Classifier	Evaluation Scope	Test Accuracy
MFCC (μ only)	13	Logistic Regression	500-sample diagnostic subset	74.20%
HuBERT Layer 12	768	Logistic Regression	500-sample diagnostic subset	86.40%
Hybrid: MFCC + HuBERT	794	Logistic Regression	Full 8,116 (80/20 split)	90.33%
Hybrid: MFCC + HuBERT	794	1D CNN	Full 8,116 (80/20 split)	90.83%

On the full corpus, the hybrid representation yields 90.33% test accuracy with Logistic Regression (1,467 of 1,624 correct) and 90.83% with the 1D CNN (1,475 of 1,624 correct). The 0.50% absolute difference corresponds to 8 additional correct classifications, a divergence which is not statistically significant at $\alpha = 0.05$ (McNemar test, $p \approx 0.38$), confirming that the 794-dimensional hybrid feature space is highly linearly separable.

B. Class-wise Performance and Confusion Profile

Because the corpus is imbalanced—Gujarat constitutes only 3.67% of samples—raw accuracy is insufficient. Table 3 reports class-wise precision, recall, and F1-score for the hybrid Logistic Regression model on the held-out test set.

TABLE III
CLASS-WISE CLASSIFICATION METRICS FOR THE HYBRID LOGISTIC REGRESSION MODEL

Accent Class	Test Support	Precision (%)	Recall (%)	F1-Score (%)
Tamil	368	91.2	92.4	91.8
Andhra	359	89.7	91.1	90.4
Karnataka	337	88.5	89.6	89.0
Kerala	334	93.1	92.8	92.9
Jharkhand	165	92.0	88.5	90.2
Gujarat	61	82.5	73.8	77.9
Macro Average	—	89.5	88.0	88.7
Weighted Average	—	90.4	90.3	90.3

The four Dravidian southern accents (Tamil, Andhra, Karnataka, Kerala) achieve F1-scores between 89.0% and 92.9%, supported by large training volumes. Gujarat exhibits the lowest recall (73.8%) and F1-score (77.9%), with test instances frequently misclassified into Jharkhand or Karnataka categories. The Macro F1-score of 88.7% reveals a 15-point class-level performance disparity compared to the 90.3% overall weighted accuracy.

C. HuBERT Layer-Wise Representation Probing

To investigate how accent-discriminative information is distributed across the depth of the pretrained encoder, we conducted a systematic layer-wise diagnostic probe. From the full corpus, 500 utterances were randomly sampled. For each utterance, hidden state activations from all 13 HuBERT representation layers (Layer 0: CNN encoder output; Layers 1–12: Transformer blocks) were extracted, temporally mean-pooled to $\mathbb{R}^{768}$, and classified using Logistic Regression with an 80/20 stratified split. Results are reported in Table 4.

TABLE IV
LAYER-WISE PROBING ACCURACY ACROSS HUBERT TRANSFORMER HIERARCHY (N = 500 DIAGNOSTIC SUBSET)

Layer	Architectural Stage	Probing Accuracy (%)	Primary Encoding [10,11]
L0	CNN Feature Encoder Output	86.00	Raw spectral envelope, local waveform shape
L1	Transformer Block 1	94.00	Vocal tract geometry, speaker timbre
L2	Transformer Block 2	94.00	Vocal tract geometry, speaker timbre
L3	Transformer Block 3	93.00	Acoustic-phonetic boundary transitions
L4	Transformer Block 4	93.00	Phonetic categories, consonant manner/place
L5	Transformer Block 5	93.00	Phonetic categories, consonant manner/place
L6	Transformer Block 6	93.00	Formant transitions, vowel quality
L7	Transformer Block 7	93.00	Formant transitions, vowel quality
L8	Transformer Block 8	91.00	Phone identities, triphone context
L9	Transformer Block 9	88.00	Sub-word tokens, lexical context
L10	Transformer Block 10	88.00	Lexical / syntactic context
L11	Transformer Block 11	89.00	High-level semantic abstractions
L12	Final Transformer Output	91.00	Masked unit prediction objective

The probing trajectory reveals three distinct phases. (i) Shallow Peak (Layers 1–2, 94.00%): Performance surges from 86.00% at the CNN output to a peak of 94.00% at Transformer Layers 1 and 2. Consistent with Pasad et al. [10, 11], early Transformer layers retain the highest mutual information with low-level acoustic properties, pitch, and individual vocal tract characteristics. (ii) Intermediate Stability (Layers 3–7, 93.00%): These blocks remain stable at 93.00%, encoding systematic phonological substitutions characteristic of regional accents (e.g., retroflex vs. dental articulation). (iii) Deep Attenuation (Layers 9–10, 88.00%): Representations become increasingly invariant to surface acoustic realization, abstracting toward lexical units to satisfy the masked cluster prediction objective, reducing linear accent separability by 5–6 percentage points. This analysis suggests that multi-layer weighted aggregation [12] targeting Layers 1–7 would maximize accent-specific information retention for future experiments.

D. Temporal Duration Sensitivity Analysis

To quantify how much speech context is required for reliable regional classification, we evaluated the hybrid Logistic Regression model on full-length utterances versus severely truncated sub-utterance segments restricted to the first 0.7 seconds.

TABLE V
CLASSIFICATION ACCURACY UNDER TEMPORAL DURATION TRUNCATION

Input Temporal Scope	Avg. Duration (s)	Feature Representation	Test Accuracy
Sub-Utterance Segment	0.70	Hybrid (MFCC + HuBERT)	64.85%
Full Continuous Sentence	4.25	Hybrid (MFCC + HuBERT)	90.33%

Restricting audio to 0.7 seconds causes test accuracy to drop from 90.33% to 64.85% (a 25.48% absolute degradation). This confirms that regional Indian English accent characteristics are not instantaneously identifiable from isolated phonetic bursts; they rely on suprasegmental prosody, vowel duration patterns, rhythmic cadences, and articulatory co-articulation patterns distributed across multi-syllable phrase contexts.

VI. INTERACTIVE SYSTEM DEPLOYMENT

To demonstrate practical real-world utility, the trained Logistic Regression classification pipeline was deployed as an interactive application using Gradio [16]. The system allows users to upload a WAV file or record speech directly through a microphone interface. The audio is processed through the complete 794-dimensional hybrid feature extraction pipeline and classified in real time (<150 ms inference latency on CPU).

Beyond accent identification, the application maps each predicted regional accent class to an authentic regional culinary recommendation profile—for example, predicting 'Andhra' triggers recommendations for Pulihora (tamarind rice), Gongura Pachadi, Gutthi Vankaya Curry, and Ragi Sangati with Natukodi Pulusu. This cultural-interface layer demonstrates how automated accent recognition can serve as an adaptive context signal for regional personalization in consumer voice applications, digital tourism, and language education platforms.

VII. CONCLUSION AND FUTURE WORK

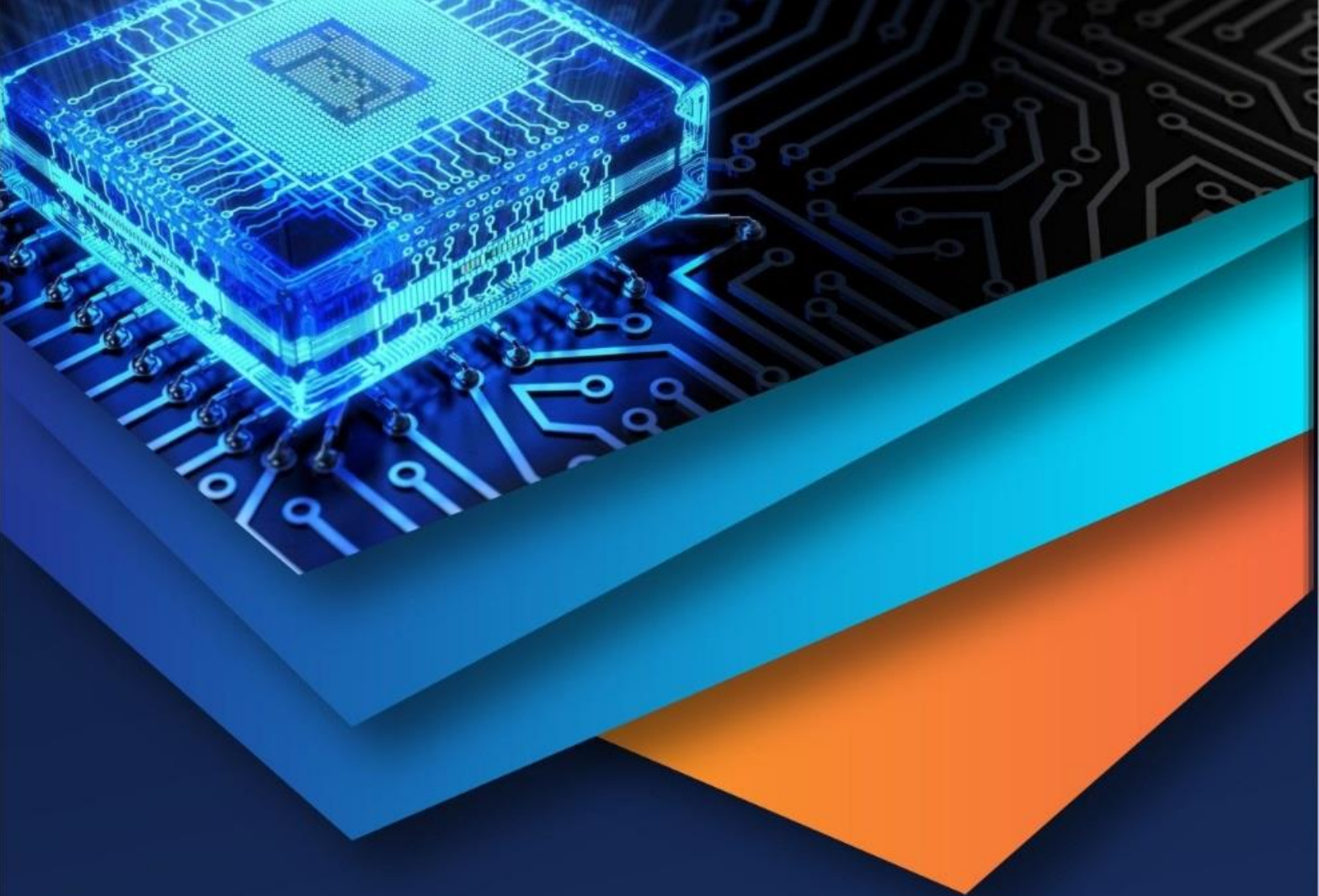
This paper presented a hybrid speech representation framework fusing 26-dimensional MFCC temporal statistics with 768-dimensional HuBERT-Base contextual embeddings into a 794-dimensional representation for six-class regional Indian English accent identification. Under stratified evaluation on 8,116 IndicAccentDB utterances, the framework achieves 90.33% accuracy with multinomial Logistic Regression and 90.83% with a 1D CNN. Hierarchical layer-wise probing across 13 HuBERT representations indicates that shallow layers (Layers 1–2) yield a peak linear separability of 94.00% by capturing low-level vocal tract and acoustic cues, before attenuating in deeper contextual layers. Temporal duration analysis demonstrates a 25.48% accuracy degradation on 0.7-second clips, confirming that regional accent discrimination relies critically on suprasegmental acoustic-prosodic context.

Future work will focus on speaker-disjoint evaluation on expanded corpora [3, 13], multi-layer HuBERT representation aggregation [12], parameter-efficient fine-tuning with LoRA [6], and training on larger, speaker-diverse datasets.

REFERENCES

- [1] S. Gaikwad, B. Gawali, and K. V. Kale, "Accent recognition for Indian English using acoustic feature approach," *International Journal of Computer Applications*, vol. 63, no. 7, pp. 25–32, 2013.
- [2] I. Patel, R. Kulkarni, and Y. S. Rao, "Automatic non-native dialect and accent voice detection of South Indian English," *Advances in Image and Video Processing*, vol. 5, no. 1, pp. 15–26, 2017. DOI: 10.14738/aivp.51.2749.
- [3] A. Ahamad, A. Anand, and P. Bhargava, "AccentDB: A database of non-native English accents to assist neural speech recognition," in *Proc. 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France, 2020, pp. 5353–5360.

- [4] S. Darshana, H. Theivaprakasham, G. Jyothish Lal, B. Premjith, V. Sowmya, and K. P. Soman, "MARS: A hybrid deep CNN-based multi-accent recognition system for English language," in Proc. 1st International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), IEEE, 2022, pp. 1–6. DOI: 10.1109/ICAITPR51569.2022.9844177.
- [5] S. Darshana, "IndicAccentDb," Hugging Face Datasets, Repository ID: DarshanaS/IndicAccentDb, 2022. [Online]. Available: <https://huggingface.co/datasets/DarshanaS/IndicAccentDb>.
- [6] R. Jairam, G. Jyothish Lal, and B. Premjith, "A few-shot multi-accented speech classification for Indian languages using transformers and LLM's fine-tuning approaches," in Proc. 4th Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech), Malta: ACL Anthology, 2024, pp. 1–7. DOI: 10.18653/v1/2024.dravidianlangtech-1.1.
- [7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.
- [8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 12449–12460.
- [9] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022.
- [10] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in Proc. IEEE ASRU Workshop, 2021, pp. 914–921.
- [11] A. Pasad, B. Bowen, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3855–3868, 2023.
- [12] S.-W. Yang et al., "SUPERB: Speech processing universal performance benchmark," in Proc. INTERSPEECH, 2021, pp. 1194–1198.
- [13] G. Zhao et al., "L2-ARCTIC: A non-native English speech corpus for accented speech research," in Proc. INTERSPEECH, Hyderabad, India, 2018, pp. 2783–2787.
- [14] K. Prahallad et al., "The IIIT-H Indic speech databases," in Proc. INTERSPEECH, Portland, OR, USA, 2012, pp. 2546–2549.
- [15] S. Chopra, A. Sharma, and S. Sakti, "Svarah: Evaluating Indian English speech recognition across diverse accents and domains," in Proc. INTERSPEECH, Dublin, Ireland, 2023, pp. 2018–2022.
- [16] A. Abid et al., "Gradio: Hassle-free sharing and testing of ML models in the wild," arXiv:1906.02569, 2019.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)