International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com OPEN ACCESS E-ISSN : 3049-2424
Volume 3 Issue 2 , Feb 2026
DOI: https://doi.org/10.71366/ijwos03022659307

An Explainable AI Framework for Voice Command Classification

Dr. S. Devibala[1], M S Chandhana Pandi[2]
[1] *Assistant Professor, PG and Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, India*
[2] *PG and Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore, India.*

Article Info

Article History:

Published:28 Feb 2026

Publication Issue:
Volume 3, Issue 2
February-2026

Page Number:
488-493

Corresponding Author:
Dr. S. Devibala

Abstract:

A feeling of discomfort tends to arise when working with something whose inner workings remain unclear. Voice-based systems have mostly escaped such scrutiny - since they function adequately most times, few tend to probe deeper. Yet those designing high-stakes applications must dig further; passive acceptance carries too much risk. It wasn't about proving neural nets can recognize spoken words correctly - we already had evidence for that - it was about peering into the model afterward, extracting reasoning clear enough for doctors, engineers, or regulators to act upon. Our approach used a convolutional network enhanced with attention mechanisms, fed with MFCC features pulled from Google Speech Commands v2, which holds over one hundred thousand utterances across thirty-five labels, achieving 97.3% precision. Floating above came SHAP, then LIME, followed by Grad-CAM - each serving as interpretability tools. A separate stage mapped timing-based relevance onto individual speech sounds. Results stayed true to model behavior while making sense in terms of spoken language. Yet beyond expectation emerged strong divergence among techniques when examples blurred category lines.

Keywords: Explainable AI, Voice Command Classification, SHAP, LIME, Grad-CAM, MFCC, Attention Mechanism, CNN, Phoneme Alignment, Interpretability, Deep Learning

1. INTRODUCTION

Chandhana spoke request slowly, expecting a reply. Silence followed instead. This quiet moment sparked everything written here. Progress in recognizing speech commands shows up plainly in test scores - models now handle complexity with ease. Still, understanding how decisions happen stays out of reach. Systems work without showing their steps. Correct answers pass unnoticed. Mistakes offer no clues about where things went off track.

What counts as acceptable clarity shifts depending on where it is used. In tools like speech-driven wheelchairs, medical interfaces, or crisis alerts, how clear things are becomes critical. Our work explored if accuracy and interpretability could coexist in one system. A classification model was developed first. Following that, techniques including SHAP, LIME, and Grad-CAM were tested for generating insights. To help users follow spoken patterns more easily, we wove in a phoneme alignment component. Then came checking whether those interpretations truly mirrored what the model relied on internally.

2. RELATED WORK

A. Speech Recognition: A Compressed History

At first, systems relied on clear phonetic patterns - accuracy was limited, yet decisions remained transparent. Following that, Hidden Markov Models brought stronger statistical methods into play: they captured how phonemes evolve over time, while Gaussian Mixture Models described sound features at every stage. Even then, understanding the outcome stayed possible - you could follow the chosen route and see exactly which sounds the model favored and what drove its choice. With deep learning came an architectural shift, abandoning much of the earlier clarity in favor of performance. End-to-end models such as Wav2Vec 2.0 and Whisper merged every stage into one unified process. Although precision improved, interpretability faded - yet this trade-off went mostly unquestioned until lately. For tasks like detecting keywords or classifying voice commands - the scope here - convolutional designs took hold due to their speed and effectiveness with time-frequency representations. Cross-model evaluations gained reliability once the Google Speech Commands dataset emerged as a common reference point.

B. Explainable AI and XAI for Audio

Starting with small tweaks to inputs, LIME builds quick, understandable linear models nearby - efficient and clear. Instead of shortcuts, SHAP uses game-theory-backed values that add up correctly and ignore useless traits, though it takes longer. From another angle, Grad-CAM tracks how class signals move backward through convolutions, giving rapid visual cues that align well with sound images. Not much has been done yet on explaining audio models. Some work by Mishra and team used altered versions of LIME on sounds and songs. Elsewhere, Haunschmid's group did similar things. In one case, Becker's analysis spread importance across layers in voice emotion systems, often spotlighting rhythm and pitch patterns - the results hint that gradient tools might uncover speech-like logic inside neural nets.

3. SYSTEM DESIGN AND METHODOLOGY

A. Dataset and Preprocessing

From Google Speech Commands v2, there come 105,829 audio clips divided into 35 word categories - some more frequent than others, so weighted sampling balanced their influence. Though recorded on different devices - like phones, laptops, or external mics - the variation was preserved intentionally, mirroring how people actually use voice interfaces. Each sound snippet gets processed using pre-emphasis with α set to 0.97, followed by slicing into 25 ms segments and transforming via a 40-band Mel filter system, producing a 40x98 MFCC representation. During learning phases, distortion comes from SpecAugment along with added background noise at signal-to-

noise ratios between 10 and 30 dB. Allocation across training, validation, and testing holds steady at 80%, 10%, and 10%, grouped evenly per label.

B. Classifier Architecture

A different path was taken by selecting a four-block CNN instead of ResNet-15 - skip connections tend to blur Grad-CAM outputs. Starting each block, dual 3x3 convolutions run (with filter counts stepping through 64, 128, 256, then 256), paired always with batch norm and ReLU activation. After those steps comes 2x2 max pooling, shrinking spatial dimensions down to 5x13. Across the time axis, attention emerges via eight-headed self-attention using 128-dimensional keys and values. This representation feeds into two dense layers, sized 256 and 128, where dropout at 0.4 stabilizes learning before reaching 35 output nodes. Optimization unfolds with Adam, guided by cosine decay, softened labels (epsilon = 0.1), plus early halting when progress stalls. Total learnable weights land near 2.1 million; full training wraps beneath four hours on one RTX 3080.

C. Explanation Methods and Phoneme Alignment

A single output from SHAP's DeepExplainer - using 500 prior cases as reference - delivers a 40 by 98 grid of influence scores; higher numbers show which time-frequency cells support the model's choice. Processing one instance takes about 2.3 seconds - feasible after the fact, yet too slow during live operation. With LIME, every MFCC layout was split into sixteen coarse blocks, then probed using 1,000 random patterns before fitting a regularized linear model where top five weights define the rationale. Based on activity in an intermediate filter layer, Grad-CAM generates heatmaps while extending prediction time by under five milliseconds. A single phoneme alignment stage was included, leveraging the Montreal Forced Aligner; from this, heatmap values were averaged across every phoneme span, resulting in one importance value per unit. This format yields a clear summary - something a speech therapist might quickly judge as reasonable. Low-confidence matches occurred in roughly 4% of test samples, leading to their removal.

4. EXPERIMENTAL RESULTS

A. Classification Performance and Explanation Findings

At 97.3%, the top-1 accuracy of the attention-enhanced CNN exceeds that of the standard CNN at 94.1%, the LSTM at 93.8%, and ResNet-15 at 96.2%, with consistent results over five different random seeds. Most notably, 32 out of the 35 classes achieve F1 scores beyond 0.95; only 'five', 'nine', and 'tree' fall short, contributing to 31% of total test mistakes because their vowel sounds share close acoustic features. Instead of blending together, the three interpretability approaches highlight parts of speech rich in spectral contrast - such as fricatives, sudden plosive releases, and shifts from vowels to consonants - a pattern fully aligned with established phonological principles. Take 'stop': Grad-CAM shows strongest activation during the initial /st/ cluster and the final /p/ burst.

In cases where 'stop' is incorrectly labeled as 'top' - happening seven times - the signal drops nearly to zero around the /s/ start, yet remains strongly positive at /t/ and /p/. When examining the critical distinction between 'yes' and 'no', each method produces clearly separate explanation patterns - and no mix-ups occurred throughout testing.

B. Faithfulness, Stability, and Ablations

Grad-CAM leads in both sufficiency and comprehensiveness when evaluated with ten superpixels - scoring 0.89 and 0.76 respectively. Though SHAP lags behind here, it shows stronger consistency, with a Jaccard index at 0.94, outperforming Grad-CAM's 0.87 and LIME's 0.79. When asked to rate interpretability, experts gave LIME the top score: 4.2 out of 5. In contrast, Grad-CAM received 3.6; SHAP only reached 3.1 - judged as overly detailed by reviewers. As inputs neared classification thresholds, SHAP and LIME pointed to distinct areas, suggesting each defines relevance in its own way. A sharp drop emerged when SpecAugment was removed - accuracy fell from 97.3% down to 95.6%. Yet divergence stood out in explanation maps under varying noise, even for identical utterances. This shift suggested reliance on both recording traits and speech sounds. When augmented, those maps stayed nearly unchanged despite differing background conditions.

5. DISCUSSION

Surprisingly, we entered the study expecting a trade-off between clarity and performance. Instead, results showed gains in both prediction precision and interpretable output. The attention component boosted correctness while also making explanations more consistent. Rather than competing, these improvements moved together. What seemed like separate outcomes actually reflected one core factor: reliance on meaningful patterns instead of accidental cues within the data.

Some people interact with this technology in distinct ways. Detailed SHAP maps serve device engineers better than anything else. Clinicians or regulators, for instance, gain more from summaries at the phoneme level - easier to check for reasonableness. Meanwhile, those simply trying to understand why a result occurred find little help across existing approaches. That space separating complex model outputs from usable explanations remains wide, barely touched so far. One reason many explanation evaluations fall short is their focus on the model itself rather than objective reality. Still, phoneme alignment offers a partial fix - when highlighted parts match sounds that stand out clearly, it suggests the model's logic lines up not only within itself but also with real-world patterns.

One major constraint lies in the need for a written transcript during phoneme alignment - an obstacle when handling open-vocabulary scenarios where such texts are unavailable. Our assessment relied solely on feedback from twelve scholars, limiting broader generalization. Processing SHAP values still demands too much time, making immediate deployment difficult on devices with limited computing power.

6. CONCLUSION

Beginning from a simple inquiry - whether a voice command classifier can clarify its decisions honestly yet understandably - we arrive at a conclusion less definitive than hoped, yet richer in insight. It functions. Explanations match the model's actual behavior, tied neatly to shifts between spoken sounds, making sense within language structure. Each of the three techniques enhances the others, achieving together what none achieves alone. Without guidance, merely through patterns in data, it focuses on parts of speech experts consider key.

Confidence rests on one core idea: transparency needn't come at a cost. Starting with intent - embedding clarity via attention mechanisms, data shaping decisions, while aligning phonemes - led to gains in openness plus stronger performance when compared to other models. What follows involves testing how everyday users interact with the system, refining SHAP estimates for live environments, adapting methods for diverse accents and second-language speakers, then crafting explanations people can grasp without relying solely on visual highlights or numerical ratings.

References

[1] P. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017.

[3] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, 2017.

[4] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," arXiv:1804.03209, 2018.

[5] S. Mishra et al., "Local interpretable model-agnostic explanations for music content analysis," in Proc. ISMIR, 2017.

[6] E. Haunschmid, M. Dorfer, and G. Widmer, "audioLIME: Listenable explanations using source separation," arXiv:2008.00582, 2020.

[7] S. Becker et al., "Interpreting and explaining deep neural networks for classification of audio signals," arXiv:1807.03418, 2018.

[8] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020.

[9] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023.

[10] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in Proc. Interspeech, 2019.

[11] M. McAuliffe et al., "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in Proc. Interspeech, 2017.