Dimension 414 and Minimal Embedding Dimensions for Phonetic Feature Encoding in WavLM

Narthana Sivalingam* and Uthayasanker Thayasivam*
* Department of Computer Science and Engineering,
University of Moratuwa, Sri Lanka
E-mail: {narthanas, rtuthaya}@cse.mrt.ac.lk

***Abstract*—Self-supervised speech models like WavLM have advanced speech processing tasks, but the role of individual embedding dimensions from each transformer layer remains unclear. This work investigates how phonetic features, voicing, fricative, and nasality are encoded within WavLM representations across layers. By combining Integrated Gradients with a hybrid elbow-point method, we identify sparse subsets of embedding dimensions that are sufficient to retain classification performance. Interestingly, one dimension Dimension 414 consistently emerges as the most influential across layers and features, suggesting that WavLM concentrates phonetic information into compact and functionally distinct dimensions. Supporting correlation analysis with handcrafted acoustic descriptors reveals that this dominant dimension aligns with interpretable speech properties. Our findings demonstrate that WavLM embeddings encode phonetic features in a sparse, structured, and explainable manner, offering new pathways for transparent and efficient speech model analysis.**

I. Introduction

Speech processing involves transforming raw audio signals into compact, learned representations that capture the essential structure of speech. Recent advances in self-supervised models like Wav2Vec 2.0 [1], HuBERT [2], and WavLM [3] have replaced handcrafted features by learning rich contextual embeddings from unlabeled audio, achieving state-of-the-art performance in tasks such as ASR, speaker identification, and emotion recognition. However, despite their success, these models remain black boxes. Each layer outputs a high-dimensional vector per audio frame, yet the role of individual components—often referred to as *neurons* or *dimensions*—and their relationship to human-interpretable speech attributes remains unclear. This lack of interpretability hinders understanding of internal model behavior. Explainability has thus become essential: it enables debugging, improves deployment efficiency through pruning, reveals performance-affecting biases, supports compliance in critical domains, and enhances transferability across languages and acoustic conditions [4].

Motivated by this need for interpretability, we categorize prior work into three perspectives: global, layer, and neuron level. Global methods, such as block-dropping and ablation, offer high-level insights into which architectural components (e.g., encoders or attention modules) influence performance, but are too coarse to reveal how specific speech features are represented. At the layer level, probing classifiers have been widely used [5] to assess whether individual layers capture specific phonetic or prosodic information. However, these approaches typically treat each layer as a monolithic unit and do not analyze the roles of individual neurons or embedding dimensions. This leads to a critical gap at the neuron level, where fine-grained insight is lacking. Without neuron-level explanation, it remains difficult to precisely debug errors, identify redundant or biased representations, or safely adapt models to new domains making this level of analysis essential for deploying speech systems in real-world applications.

To advance the scope of neuron level interpretability in speech models, this study builds upon a previously established layerwise probing framework [6] by extending it toward fine-grained neuron level analysis. Rather than treating probing as a binary indicator of feature presence within a layer, we aim to identify which specific neurons contribute to the encoding of phonetic attributes and to assess the sparsity and consistency of their involvement. This transition from coarse-grained layer analysis to neuron-level attribution enables a more detailed understanding of how individual embedding dimensions participate in representing features such as voicing, fricative, and nasality, and whether shared neuron patterns emerge across these categories.

Our investigation integrates probing classifiers, IG-based attribution, and a hybrid elbow-point selection method to identify sparse, informative subsets of neurons. While prior studies have applied SHAP [7] for interpretability [8], its assumption of feature independence limits its suitability for this task. In contrast, Integrated Gradients better captures interactions among dimensions [9]. As summarized in Figure 1, our analysis reveals a particularly striking pattern where Dimension 414 consistently ranks among the most influential across nearly all probes. Furthermore, we find that a small, sparse subset of embedding dimensions including 414 can preserve classifier performance, suggesting that WavLM internally organizes phonetic information in a compact and interpretable manner. To further validate this finding, we correlate the most important neuron with standardized acoustic descriptors from the eGeMAPS set [10], uncovering strong links with interpretable formant-based features.

Despite these promising results, fine-grained interpretability remains challenging due to four factors: the high dimensionality of WavLM embeddings, nonlinear interactions and acoustic overlap between phonetic features. Still, our method successfully identifies sparse and consistent neuron subsets, advancing transparent speech model representations.

979-8-3315-7206-8/25/$31.00

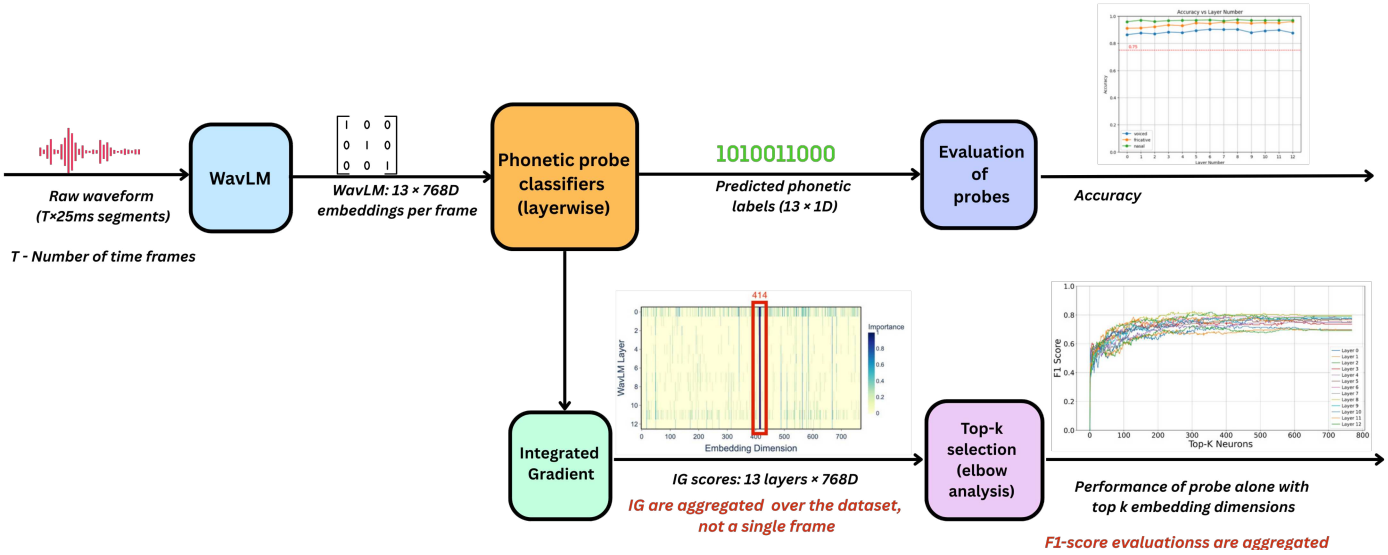


Fig. 1: **Overview of the proposed interpretability framework.** Raw audio is segmented and passed through WavLM to extract 13-layer embeddings per frame. Phonetic probe classifiers predict feature presence layerwise. Integrated Gradients (IG) are computed to assess neuron-level importance, aggregated across the dataset. A hybrid elbow-point method is used to select sparse top-k neurons while preserving probe performance. Notably, **Dimension 414 consistently emerges as the most influential across tasks**, Performance using only top-k dimensions reveals sparse and interpretable encoding.

II. RELATED WORKS

A. Deep Speech Representation Learning

Speech representation learning has evolved through distinct phases beginning with handcrafted acoustic features, progressing through supervised deep models, and culminating in modern self-supervised approaches [11] such as Wav2Vec 2.0 [1], HuBERT [2], and WavLM [3] learn expressive representations directly from raw audio. Among these, WavLM stands out by jointly optimizing masked prediction and denoising objectives, capturing both linguistic and paralinguistic traits. It further incorporates gated relative position bias and scales training to over 94,000 hours of audio, achieving strong performance on the SUPERB benchmark. Owing to its robust generalization, WavLM forms the foundation of our study on phonetic representation and neuron-level interpretability. Building on its architecture, we next explore how prior work has examined internal model structure, particularly via layerwise probing.

B. Layerwise Probing in Speech

Layerwise probing is widely used to examine how speech representations evolve across model depths. Prior studies [6], [12], [13] applied probing classifiers to hidden layers to predict phonetic or linguistic attributes, revealing a hierarchy: lower layers capture acoustic features (e.g., pitch, formants), while higher layers encode abstract semantics or prosody. While this has advanced understanding of layer roles, prior work treats each layer as a monolithic unit, overlooking contributions from individual embedding dimensions. We build on this foundation by shifting focus to the neuron level enabling more granular insight into how phonetic features are internally distributed.

C. Neuron-Level Attribution for Interpretable Speech Representations

Neuron-level interpretability has gained attention in NLP, most notably through the "sentiment neuron" discovery [14], but remains underexplored in speech. Dixit et al. [8] applied SHAP [7] to WavLM embeddings for emotion recognition, identifying salient dimensions. However, SHAP assumes feature independence and relies on extensive perturbations, making it computationally intensive and less stable in high-dimensional settings like 768D speech embeddings. Their study also did not examine whether a compact subset of neurons suffices for classification or if specific neurons consistently contribute across tasks.

In contrast, our work focuses on phonetic feature classification, aiming to identify minimal and consistent neuron subsets responsible for encoding voicing, fricative, and nasality. To this end, we adopt Integrated Gradients (IG) [9], which satisfies formal interpretability axioms and captures interdependencies among input dimensions. IG scales efficiently to high-dimensional data and quantifies neuron importance by integrating gradients from a baseline to the input. We use these scores to rank neurons and apply a sparsity-driven top-k selection strategy, enabling compact and interpretable subsets. This sets the stage for identifying such subsets via data-driven elbow-point analysis.

D. Elbow Point Analysis for Sparse Neuron Subset Selection

Elbow-point analysis, initially proposed for estimating cluster counts in unsupervised learning [15], is often applied via visual inspection of curve inflections. More recent variants, such as the maximum-distance method [16] and knee-detection

Algorithm 1 Hybrid Top-k Selection via Elbow Point Analysis

Require: IG importance scores $I \in \mathbb{R}^{768}$, sorted by descending importance
Require: F1-scores F_k for top-k subsets where $k = 10, 20, \ldots, 768$
Smooth F_k using a moving average (window size = w)
Normalize F_k curve to $[0, 1]$ range
Apply max-distance elbow method:
    Draw line from $(k_{\min}, F_{k_{\min}})$ to $(k_{\max}, F_{k_{\max}})$
    Compute perpendicular distance of each F_k point to this line
    Select k^* with maximum distance as Elbow_MaxDist
Apply pairwise linear fitting:
for each k_i in $[k_{\min}, k_{\max}]$ **do**
    Fit two lines: L_1 on $[10, k_i]$, L_2 on $[k_i + 1, 768]$
    Compute total squared error: $E = \text{MSE}(L_1) + \text{MSE}(L_2)$
end for
Select k^* minimizing E as Elbow_LinearFit
Compare both elbow points with full 768-D F1-score
Choose final k^* closest to full-dim F1-score
return Final top-k^* dimension indices from IG ranking

algorithms [17], offer automated alternatives. While widely used in clustering, these techniques have not been applied to neuron selection in speech models. We adapt these strategies to supervised probing, where performance curves exhibit diminishing returns. Our hybrid method—outlined in Algorithm 1—combines curve smoothing, max-distance detection, and piecewise linear fitting to select top-k neuron subsets that preserve classification performance.

E. Correlation-Based Interpretability of Speech Embeddings

To interpret neuron-level representations in speech models, we examine how top-k neuron activations correlate with known acoustic features. We use Spearman correlation for its robustness to non-linear, monotonic relationships and its resilience to outliers and scale variations—common in neural activations. This choice is supported by empirical comparisons across eight statistical methods and aligns with established interpretability guidelines [18]. As our interpretability target, we adopt the eGeMAPS feature set [10], an 88-dimensional collection of widely used paralinguistic descriptors spanning frequency, energy, spectral, and temporal traits. eGeMAPS offers a linguistically grounded bridge between learned neural embeddings and human-interpretable speech characteristics.

III. Experimental Methodology

This section outlines the end-to-end pipeline used in our study. We extract WavLM embeddings and align them with phonetic labels from the TIMIT dataset. Probing classifiers are then trained across 13 layers for three phonetic features: voicing, fricative, and nasality. Using Integrated Gradients, we compute per-dimension importance scores for each probe. A hybrid elbow-point detection method identifies sparse, high-importance neuron subsets. Finally, we analyze correlations between these neurons and interpretable acoustic features from the eGeMAPS set.

A. Dataset Preparation

We use the TIMIT corpus, a standard phonetic dataset with 5.4 hours of 16 kHz read speech and time-aligned phoneme labels. The standard train/test split is used, covering various American English dialects.

1) WavLM Embedding Extraction: We extract 13-layer embeddings (including the convolutional layer as Layer 0) from the WavLM Base model. Each utterance yields a tensor of shape [13 × N × 768], where N is the number of 25 ms frames (20 ms stride), and each 768D vector represents a contextual frame embedding.

2) Probing Dataset: Phone-level embeddings are created by averaging frame-level vectors within each phoneme segment in the training set, resulting in 4018 samples per layer. Each is labeled with binary tags for voicing, fricative, and nasality using standard phone-to-feature mappings. The test set yields 560 frame-level embeddings aligned to phones via timestamps, used for probe evaluation.

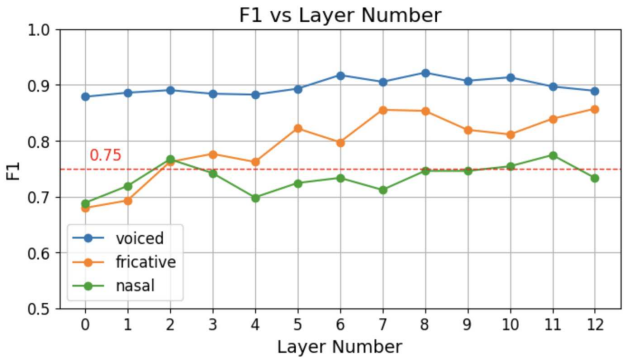
3) Correlation Dataset: For correlation, we compute mean utterance-level WavLM embeddings across all layers. In parallel, 88 eGeMAPS features are extracted using openSMILE. Each sample pairs a 768D mean embedding with its acoustic descriptor vector, enabling per-layer correlation analysis.

B. Probing Classifier Construction

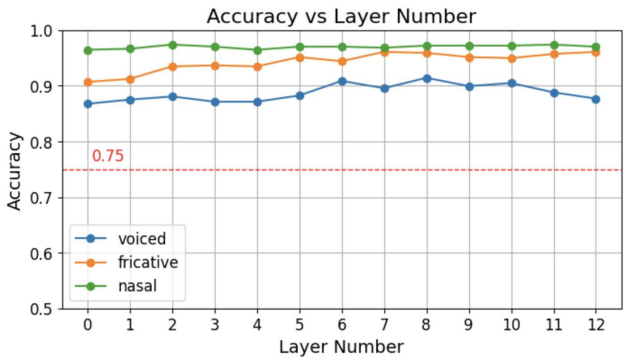
To examine whether specific WavLM layers encode phonetic features such as voicing, fricative, and nasality, we construct probing classifiers using a one-hidden-layer MLP implemented in PyTorch. Each model takes a 768-dimensional WavLM embedding, passes it through a hidden layer with 200 ReLU units, and outputs a sigmoid-activated binary prediction. This lightweight architecture is expressive enough to capture learnable patterns without overfitting. We train one probe per (layer, feature) pair, resulting in 39 models (13 layers × 3 features), using phone-averaged embeddings from the TIMIT training set with corresponding binary labels. For evaluation, probes are applied to frame-level embeddings from the TIMIT test set. As embedding dimensionality remains consistent (768D), models trained on averaged vectors transfer directly to time-step inputs. Performance is measured using accuracy and F1-score, with F1 prioritized due to potential label imbalance.

C. Attribution Score Computation

We compute Integrated Gradients (IG) for all trained probes using the Captum library to identify influential embedding dimensions. IG attributes importance by integrating gradients from a baseline to the input; we use a zero vector as the baseline, a standard choice in high-dimensional settings. For each (layer, feature) probe, we compute IG on the test set and average the resulting 768-dimensional scores across all samples to obtain stable per-dimension attributions. These profiles reveal consistently dominant neurons and guide top-k selection and correlation analysis. While averaging improves



(a) F1-score across layers for each phonetic feature.



(b) Accuracy across layers for each phonetic feature.

Fig. 2: Performance of probing classifiers across WavLM layers.

robustness, a full stability analysis across seeds and baselines is left for future work.

D. Top-k Dimension Selection via Hybrid Elbow Analysis

To identify sparse embedding subsets that preserve probe performance, we rank the 768 dimensions by absolute normalized average Integrated Gradients (IG) importance. For each probe (layer-feature pair), we evaluate F1-scores across increasing top-k subsets (e.g., $k = 1, 2, \ldots$) by masking the rest. We propose a hybrid strategy combining smoothing, geometric, and error-based techniques to locate the elbow point—where performance stabilizes with minimal dimensions:

1) *Smooth the F1 vs. top-k curve using moving average.*
2) *Apply the max-distance method: find the point farthest from the line connecting endpoints.*
3) *Use piecewise linear fitting: identify the split minimizing total error.*
4) *Select the elbow closest to full-dim F1 with the smallest k.*

This approach balances interpretability and performance, yielding compact and informative neuron subsets. The full procedure is detailed in Algorithm 1.

E. Correlation with Acoustic Features

To interpret the dominant neuron, we compute Spearman correlations between Dimension 414 and the 88 eGeMAPS features using utterance-level embeddings (see Section 4.1). For each layer, we assess whether Dimension 414 aligns with interpretable speech attributes such as pitch, formants, or energy.

Only strong spearman correlations($|\rho| \geq 0.7$) are retained, following statistical convention [19]. The results are visualized as heatmaps linking embedding dimensions to acoustic features (see Fig. 5). While these correlations suggest interpretability, we emphasize that they indicate association, not causation—neural encoding likely involves non-linear, multi-dimensional interactions.

IV. Results and Observation

A. Performance of Probes Across Layers

To assess how WavLM layers capture phonetic features, we trained 39 probing classifiers (13 layers × 3 features) using a one-hidden-layer MLP on 768D embeddings. Evaluation on TIMIT test frames uses accuracy and F1-score, with the latter emphasized due to class imbalance. Figure 2 confirms that phonetic encoding varies by depth, motivating our subsequent neuron-level analysis.

B. Importance of dimensions in the embedding across layers

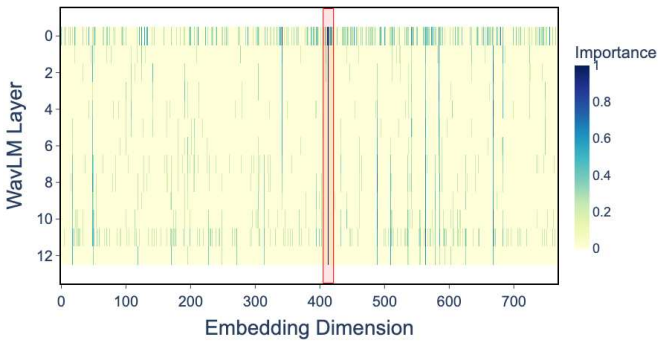
We used Integrated Gradients (IG) to compute per-dimension importance scores for all 39 probes. Averaging IG scores across test samples yields relevance profiles for each of the 768 embedding dimensions. As shown in Figure 3, importance is unevenly distributed across layers, but a few neurons consistently stand out. Dimension 414 ranks 1st in 38 probes and 2nd in one, making it a globally dominant dimension across phonetic features. Its IG values are positive for voicing and negative for fricative and nasality, suggesting a feature-specific modulation role. These results indicate that while many neurons contribute, a small subset—including Dim 414—plays a central role in phonetic representation.

C. Sparse subset selection through top-k analysis

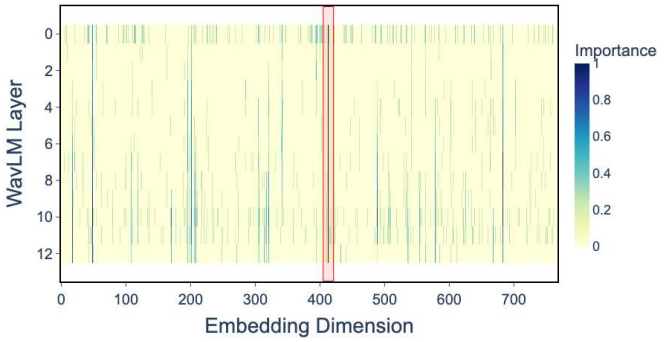
Using the hybrid selection strategy detailed in Algorithm 1, we evaluate how top-k IG-ranked neurons contribute to probing performance. As shown in Figure 4, F1-scores increase sharply in the early top-k range and stabilize gradually, though the number of neurons needed to reach saturation varies across layers and features.

D. Correlation of Dimension 414 with Handcrafted Acoustic Features

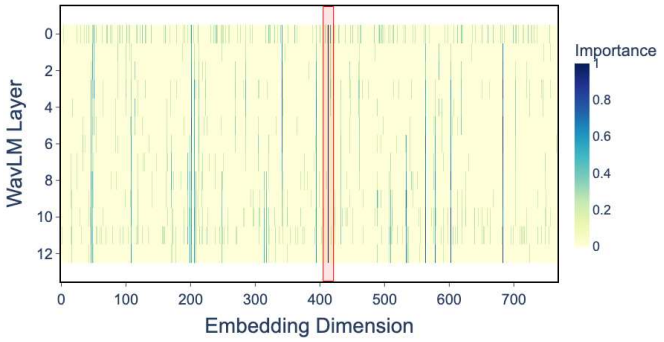
To interpret the role of Dimension 414, we computed its Spearman correlation with the 88 eGeMAPS acoustic features across layers. As shown in Figure 5, several strong correlations ($|\rho| \geq 0.7$) emerged. In Layer 1, it negatively correlated with F43 (Mean F1 Bandwidth). In Layer 12, it showed negative correlations with F45 and F51 (Mean F1/F2 Amplitude) and



(a) Voicing

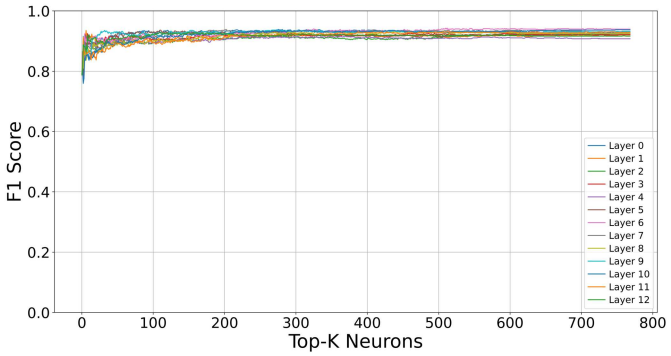


(b) fricative

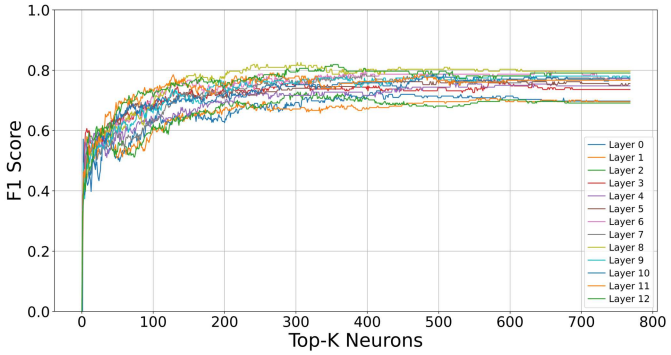


(c) Nasality

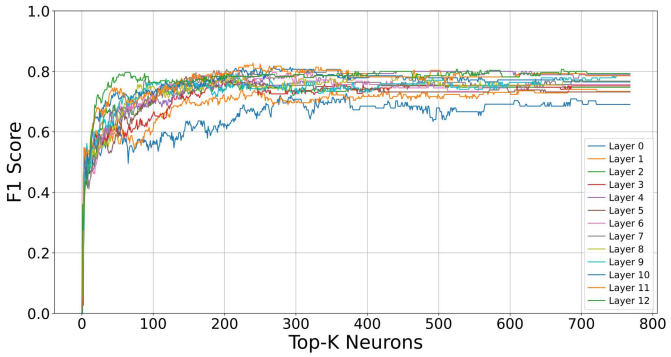
Fig. 3: Integrated Gradients heatmaps showing mean IG scores across 13 WavLM layers and 768 embedding dimensions for each phonetic feature.



(a) Voicing



(b) fricative



(c) Nasality

Fig. 4: Top-k neuron selection curves for all 13 layers. F1-score vs. number of top IG-ranked dimensions retained. Each curve represents one layer.

positive correlations with F46 and F52 (F1/F2 Amplitude Variability). These formant-related features reflect spectral energy and articulatory structure, particularly relevant to voicing and fricative. The consistency of these trends suggests that Dimension 414 encodes phonetic-relevant, interpretable acoustic cues linked to vocal tract resonances [20].

V. Conclusion

We enhance WavLM interpretability by moving from layer-wise probing to neuron-level analysis of phonetic feature encoding. Using Integrated Gradients for attribution and a hybrid elbow-point method for top-k selection, we identify sparse neuron subsets that retain classification performance while improving interpretability. Notably, Dimension 414 emerged as dominant, ranking first in 38 out of 39 probes and second in the other. Its attribution polarity varied: positive for voicing, but negative for nasality and fricative, suggesting sensitivity not just to voicing presence but to its acoustic quality. Correlation with eGeMAPS descriptors validated this, linking Dimension 414 to formant sharpness (F1 bandwidth) and dynamics (amplitude variability)—indicators of resonance clarity in voiced speech [20]. Overall, our findings show

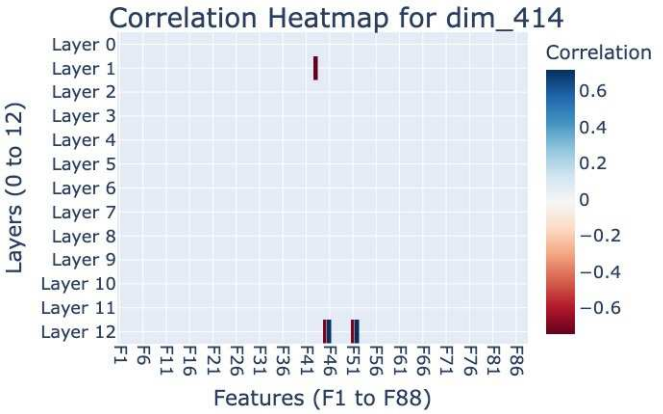


Fig. 5: Spearman correlation of Dimension 414 with 88 eGeMAPS acoustic features across all 13 WavLM layers. Bright regions indicate strong correlations ($|\rho| \geq 0.7$).

WavLM embeddings contain functionally meaningful, interpretable dimensions, paving the way for more transparent and linguistically grounded speech models.

Acknowledgment

We acknowledge the support received from the LK Domain Registry in publishing this paper for APSIPA ASC 2025.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, *et al.*, "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.

[3] S. Chen, Y. Wang, C. Wu, *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.

[4] R. Dwivedi, D. Dave, H. Naik, *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[5] T. Levy, O. Goldman, and R. Tsarfaty, *Is probing all you need? indicator tasks as an alternative to probing embedding spaces*, 2023. arXiv: 2310.15905 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.15905.

[6] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Discovering phonetic feature event patterns in transformer embeddings," in *INTERSPEECH*, 2023.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[8] S. Dixit, D. M. Low, G. Elbanna, F. Catania, and S. S. Ghosh, "Explaining deep learning embeddings for speech emotion recognition by predicting interpretable acoustic features," *arXiv preprint arXiv:2409.09511*, 2024.

[9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *International Conference on Machine Learning*, pp. 3319–3328, 2017.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[11] A. Mohamed, H.-y. Lee, L. Borgholt, *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[12] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 914–921.

[13] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis," *Computer Speech & Language*, vol. 83, p. 101 539, 2024.

[14] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," 2018.

[15] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953. DOI: 10.1007/BF02289263.

[16] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.

[17] C.-R. Huang, C.-S. Chen, and P.-C. Chung, "Contrast context histogram—an efficient discriminating local descriptor for object recognition and image matching," *Pattern Recognition*, vol. 41, no. 10, pp. 3071–3077, 2008, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2008.03.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320308000988.

[18] J. Cohen, *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[19] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[20] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2014.