Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Attention Is Not Always the Answer: Optimizing Voice Activity Detection with Simple Feature Fusion

Kumud Tripathi, Chowdam Venkata Kumar*, Pankaj Wasnik*

Media Analysis Group, Sony Research India

{kumud.tripathi, chowdam.kumar, pankaj.wasnik}@sony.com

Abstract

Voice Activity Detection (VAD) plays a key role in speech processing, often utilizing hand-crafted or neural features. This study examines the effectiveness of Mel-Frequency Cepstral Coefficients (MFCCs) and pre-trained model (PTM) features, including wav2vec 2.0, HuBERT, WavLM, UniSpeech, MMS, and Whisper. We propose FusionVAD, a unified framework that combines both feature types using three fusion strategies: concatenation, addition, and cross-attention (CA). Experimental results reveal that simple fusion techniques, particularly addition, outperform CA in both accuracy and efficiency. Fusion-based models consistently surpass single-feature models, highlighting the complementary nature of MFCCs and PTM features. Notably, our best-performing fusion model exceeds the state-of-the-art Pyannote across multiple datasets, achieving an absolute average improvement of 2.04%. These results confirm that simple feature fusion enhances VAD robustness while maintaining computational efficiency.

Index Terms: Voice activity detection, pre-trained model, feature fusion, light-weight model

1. Introduction

Voice Activity Detection (VAD) is the task of detecting speech segments within an audio signal [1]. It serves as a fundamental pre-processing step for various speech-related applications, including Automatic Speech Recognition (ASR), Speaker Recognition, Speaker Verification, and Speaker Diarization [2, 3, 4]. By accurately identifying speech and non-speech regions, VAD significantly enhances the performance of these systems by filtering out non-speech and noisy segments. As these speech-based technologies become more prevalent in applications such as virtual assistants, hearing aids, and telecommunications, improving VAD accuracy has become crucial for enhancing both user experience and system robustness. The problem of VAD has been an active research topic for several decades, typically approached as a frame-level classification task, distinguishing between speech and non-speech. Traditional approaches to VAD employ threshold-based or statistical machine learning methods using acoustic features such as energy, zero-crossing rate, pitch, and auto-correlation [5, 6, 7]. While, these methods perform well in clean environments, they often fail in real-world scenarios where background noise and varying acoustic conditions degrade their reliability.

Modern deep learning approaches for VAD, including Convolutional Neural Networks (CNNs) [8] and Recurrent Neural Networks (RNNs) [9], have demonstrated superior performance by effectively integrating frequency-domain filtering with temporal sequence modeling. These architectures enhance the robustness of VAD models in real-world noisy environments by jointly learning feature extraction and task modeling [10, 11]. However, their performance heavily relies on the availability of large-scale labeled datasets. In contrast, pre-trained models (PTMs) such as wav2vec 2.0 [12], HuBERT [13], and WavLM [14] utilize vast amounts of unlabeled speech data to learn generalized representations using CNNs and Transformers. wav2vec 2.0 and HuBERT efficiently capture phonetic structures, while WavLM enhances robustness [15]. UniSpeech and Massively Multilingual Speech (MMS) leverage multilingual learning [16, 17], while Whisper demonstrates strong performance in large-scale ASR tasks [18], including those involving low-resource languages [19]. Their diverse learning paradigms provide valuable insights for downstream speech processing tasks [20].

These PTM models have demonstrated success in several binary classification tasks, including DeepFake detection [21] and Violence Detection [22]. Given that VAD is also a binary classification task (Speech vs. Non-Speech), PTM-based approaches are particularly well-suited for this problem. Previous studies have employed PTM models for VAD by fine-tuning them on task-specific labeled datasets, demonstrating state-of-the-art results [23, 24]. However, there has been limited exploration of why PTM features perform well for VAD and how they compare to traditional hand-crafted features like MFCC [25]. Additionally, the potential benefits of combining PTM features with MFCC for VAD remain largely unexplored.

In this work, we systematically analyze the effectiveness of MFCCs and PTM-based speech representations for VAD. We explore wav2vec 2.0, HuBERT, WavLM, UniSpeech, MMS, and Whisper as PTMs due to their proven success in various speech-processing tasks. First, we compare the performance of VAD models trained separately with MFCCs and PTM representations and analyze their respective failure cases. Next, we explore different feature fusion techniques, including concatenation, addition, and cross-attention, to combine MFCC and PTM representations. Our experiments on publicly available datasets such as AMI [26], Callhome [27], and VoxConverse [28] reveal that both MFCC and PTM features contain complementary information, which, when effectively fused, enhances VAD performance. Surprisingly, we find that simple fusion techniques like concatenation and addition outperform cross-attention-based fusion, challenging the common assumption that complex attention mechanisms are always necessary for effective speech classification. The key contributions of this work are as follows:

1. Examines the role of attention mechanism in feature fusion for Voice Activity Detection (VAD) and shows that attention-based fusion is not always necessary for effective speech and

*These authors contributed equally.

10.21437/Interspeech.2025-2466

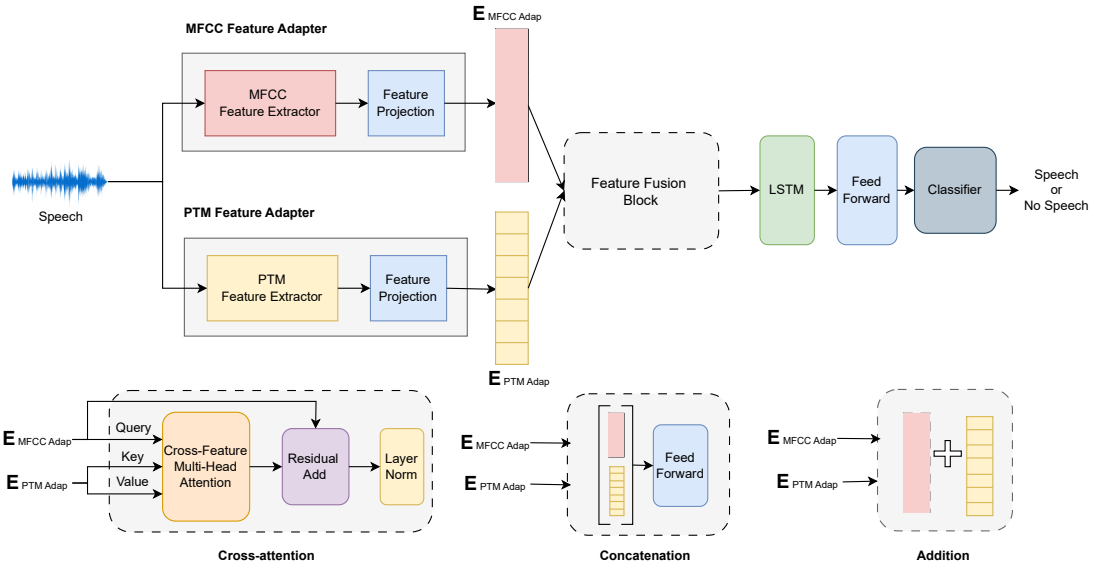


Figure 1: *Overview of the FusionVAD Framework with Different Feature Fusion Strategies.*

non-speech classification.

2. Introduces a simple yet effective feature fusion method that combines MFCC and PTM representations.
3. Conducts a comprehensive analysis of state-of-the-art PTMs to evaluate their effectiveness for VAD.
4. Demonstrates that addition-based feature fusion enhances both accuracy and computational efficiency.

2. Methodology

2.1. MFCC vs PTM Features

Pre-trained model based features have proven effective for various speech tasks, including VAD. These models leverage self-attention to capture long-range dependencies and generate contextual representations, which are particularly beneficial in noisy environments. In contrast, traditional hand-crafted features like spectrograms and MFCCs offer static time-frequency representations. While such information may be sufficient for speech detection in clean conditions, it becomes less effective in noisy settings where overlapping frequency components obscure speech cues. In these cases, PTM features provide more robust representations, as they are trained on diverse acoustic conditions and noise types. This robustness makes them valuable for improving VAD performance under challenging conditions. Understanding the strengths of PTM features relative to traditional features is essential, especially if they are found to encode complementary information. Combining both types of features could potentially enhance overall VAD performance by leveraging the contextual awareness of PTMs and the fine-grained spectral detail of hand-crafted features.

To analyze different feature types for VAD, we train models using both hand-crafted and PTM features separately. MFCCs represent hand-crafted features, while PTM features include speech encoders such as wav2vec 2.0, HuBERT, WavLM, UniSpeech, MMS, and Whisper. All speech encoders remain frozen during training, focusing on evaluating their effectiveness for VAD rather than fine-tuning them. For a fair comparison, we use the same architecture across all models. This architecture, referred to as FusionVAD, replaces the feature fusion block with a feedforward layer, allowing the use of one feature type at a time. The features first pass through two fully connected layers with a hidden size of 128 and GELU activation. Then, two bidirectional LSTM layers (hidden size 128) capture sequence information, which enhances VAD robustness against noise. Finally, two linear layers (hidden size 128, GELU activation) and a classification layer with sigmoid activation produce the output. Additionally, we visualize model predictions to identify failure cases and highlight the complementary nature of hand-crafted and PTM features, further reinforcing the importance of feature selection for improving VAD performance.

2.2. Feature Fusion Techniques

To integrate MFCC features with PTM representations, we explore three feature fusion techniques: concatenation, addition, and cross-attention. The block diagram (Figure 1) illustrates the FusionVAD pipeline, where both MFCC and PTM features are extracted, projected, and fused before being processed by an LSTM, followed by a feedforward network and classifier to determine speech or non-speech. The overall architecture remains consistent across all fusion methods, with differences only in the feature fusion block. Initially, MFCC and PTM features are passed through feature projection layers—fully connected layers that map both feature sets to a 128-dimensional space. Each fusion technique is implemented as follows:

1. **Concatenation:** The MFCC and PTM features are concatenated along the feature dimension and then passed through a fully connected layer to project them back to 128 dimensions before being input to the LSTM.
2. **Addition:** Element-wise addition is performed between the MFCC and PTM feature vectors, directly combining their information.
3. **Cross-Attention:** Projected MFCC features serve as queries, while the PTM features act as keys and values in a multi-head attention mechanism with a 128-dimensional hidden space and two attention heads. A residual connection adds the original projected MFCC features to the cross-attended output to retain important spectral information, followed by layer normalization for stable feature representation.

Table 1: *Performance (in %) of Voice activity detection with and without feature fusion. *Bold represents the best result.*

Feature Extractor	Base			Feature Fusion								
				Concatenation			Addition			Cross-attention		
	DER	FAR	MR	DER	FAR	MR	DER	FAR	MR	DER	FAR	MR
MFCC	6.79	3.23	3.56	-	-	-	-	-	-	-	-	-
wav2vec 2.0	6.70	1.77	4.92	5.95	1.82	4.14	6.74	1.96	4.78	6.12	2.08	4.04
HuBERT	7.51	1.80	5.71	4.95	2.29	2.66	6.82	0.92	5.89	6.39	2.01	4.38
WavLM	6.05	2.03	4.01	5.42	1.94	3.48	4.95	2.66	2.30	5.81	3.67	2.14
UniSpeech	6.25	2.22	4.03	5.58	2.19	3.38	5.55	3.34	2.21	6.44	2.26	4.18
MMS	6.33	1.45	4.88	5.11	2.91	2.19	4.87	2.32	2.55	5.93	3.38	2.55
Whisper	5.83	2.55	3.28	4.70	1.61	3.09	**4.50**	1.74	2.76	5.34	2.88	2.46

Table 2: *Comparison (in %) of best performing fusion model with baseline Pyannote.*

Model	AMI			Callhome			VoxConverse		
	DER	FAR	MR	DER	FAR	MR	DER	FAR	MR
Pyannote [29]	11.07	**1.70**	9.37	4.68	**0.54**	4.14	3.89	2.40	1.49
Whisper-MFCC-Addition	**7.25**	2.73	**4.53**	**3.28**	0.67	**2.61**	**2.97**	**1.82**	**1.15**

These fusion techniques aim to combine complementary information from spectral and learned representations to improve VAD performance.

3. Experiments

3.1. Dataset and Evaluation Metrics

We conducted all our experiments on three publicly available datasets, i.e., AMI, Callhome, and VoxConverse, to ensure domain diversity. We followed the dataset split methodology from [30]. Since VoxConverse lacks an official training set, we partitioned its development set into 144 training files and 72 development files. For the Callhome dataset, comprising unscripted English telephone conversations, we selected 139 files: 89 for training, and 25 each for development and testing, resulting in 22 hours of training data. The AMI Corpus was used with its official split, but we restricted training to the first 10 minutes of each file to maintain consistency in training durations across datasets, yielding 22 hours of training data. In total, we used approximately 75.4 hours of audio: VoxConverse contributed 19 hours (15 training, 2 development, 2 testing), AMI contributed 26 hours (22 training, 2 development, 2 testing), and Callhome contributed 30.4 hours (22 training, 4.2 development, 4.2 testing). Performance evaluation was conducted using standard VAD metrics: False Alarm Rate (FAR), Missing Rate (MR), and Detection Error Rate (DER), where DER is the sum of FAR and MR. These metrics provide a comprehensive measure of VAD performance across different datasets.

3.2. Experimental Setup

Training is configured for 50 epochs for all model training. Early stopping criteria with 5 epoch patience is used to avoid over fitting. Area under ROC on validation dataset is used for early stopping and also to select the best checkpoints. Models are trained on 2 seconds of chunks with a batch size of 32. All features are extracted with a stride of 20 ms. Pyannote toolkit [29] is used for training and testing the models and PTM speech encoders model checkpoints are obtained from huggingface. Base version checkpoints are considered for wav2vec 2.0[1], HuBERT[2], WavLM[3], UniSpeech [4]and Whisper[5]. One billion parameters checkpoints is used for MMS[6]. We consider baseline as Pyannote VAD [30] (official implementation in Pyannote toolkit is used for training) to compare with best performing model. Pyannote VAD follows same architecture except the initial feature extractor, which is Sincnet [31]. We train the Pyannote VAD with the same datasets used for other models.

4. Results and Analysis

4.1. MFCC vs PTM Features

First 3 columns in Table 1 shows the performance of VAD with individual features in terms of DER, FAR and MR. Whisper shows better performance than all other models. It is observed that MFCCs show high FAR than all other PTM features and lower MR than all PTM features except whisper. Also, it can be seen that all PTM features suffer from high MR than MFCC except Whisper. This pattern reveals that MFCC has information which helps in reducing MR, whereas PTM features has information which can reduce FAR. Through this pattern it can be hypothesized that MFCC might be detecting all high energy regions as speech including noisy, which results in high FAR. Whereas PTM features are correctly eliminating noise but also missing out many speech segments at the same time resulting in high MR. This shows that both features have complementary information, which can help to improve their worse counter parts if fused together.

4.2. Comparison of Feature Fusion Techniques

Table 1 presents the VAD performance using different feature fusion techniques applied to various PTM features. The results clearly indicate that all fusion-based models outperform their respective base models, demonstrating the effectiveness of feature fusion. The improvement in DER is mainly from a reduction in MR, which can be hypothesized because of MFCCs.

[1] https://huggingface.co/facebook/wav2vec2-base

[2] https://huggingface.co/facebook/hubert-base-ls960

[3] https://huggingface.co/patrickvonplaten/wavlm-libri-clean-100h-base-plus

[4] https://huggingface.co/microsoft/unispeech-sat-base-100h-libri-ft

[5] https://huggingface.co/openai/whisper-base

[6] https://huggingface.co/facebook/mms-1b-all

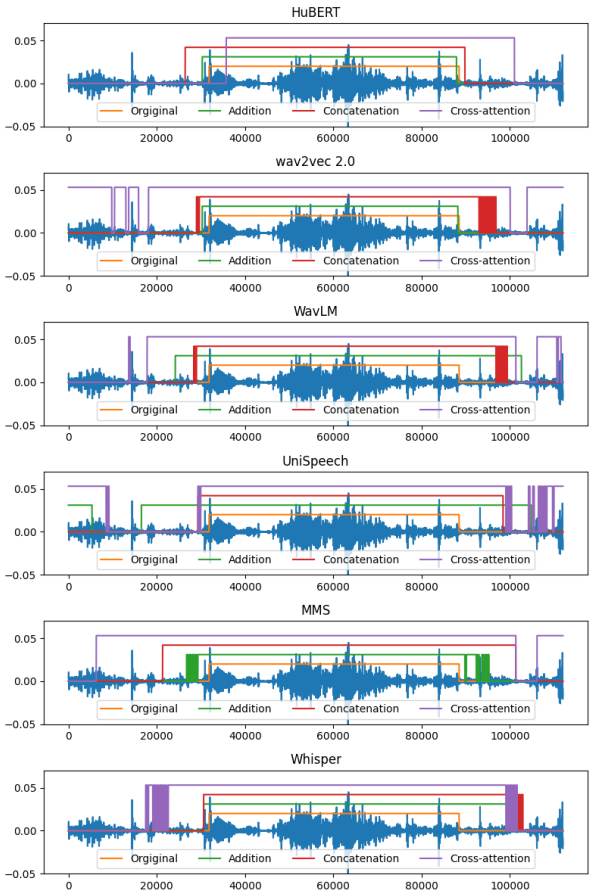


Figure 2: *Feature fusion outputs (Green: Addition, Red: Concatenation, and Purple: Cross-Attention) along with the original reference (Yellow) for all FusionVAD models on a single audio segment from the AMI file "EN2004a".*

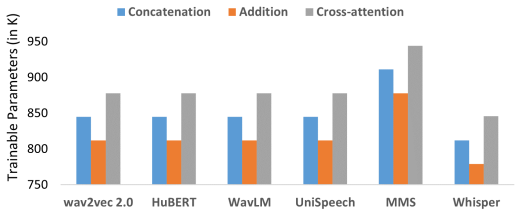


Figure 3: *Analysis of trainable parameters for all the FusionVAD models using different fusion techniques.*

This highlights the advantage of combining features that provide complementary information.

Additionally, cross-attention (CA) models consistently perform worse than concatenation and addition across all PTM features. This suggests that CA may not be optimal for this task. Among simpler fusion methods, addition outperforms concatenation in four out of six cases. Table 2 presents a dataset-wise comparison between the best fusion model i.e Fusion of MFCC and Whisper with addition and the SOTA Pyannote VAD. The fusion model consistently outperforms Pyannote across all three datasets, achieving an absolute average DER improvement of 2.04%. To further validate our approach, we experimented with the rVAD method [32] and used multi-resolution cochleagram (MRCG) features for comparison [33]. Our model outperformed rVAD by approximately 12%. However, incorporating MRCG features led to a performance drop of around 2% compared to using MFCC features alone in our best-performing model Whisper-MFCC-Addition.

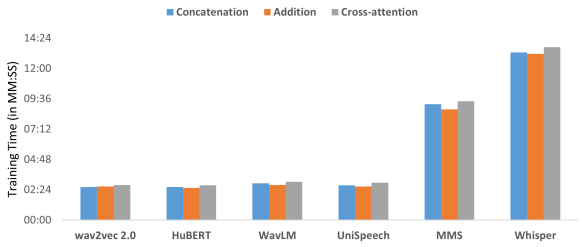


Figure 4: *Analysis of training time for all the FusionVAD models using different fusion techniques (M: minutes and S: seconds).*

Figure 2 presents the predictions obtained using different feature fusion techniques across various PTMs for a selected segment from the AMI corpus. The results indicate that concatenation and addition models produce boundaries that closely align with the ground truth, whereas CA fails to maintain this consistency in all cases. This trend aligns with the DER pattern observed in Table 1, reinforcing that simpler fusion techniques outperform the more complex CA method for VAD. We hypothesize that VAD is inherently a simpler task that does not require extensive contextual information, unlike more complex tasks such as ASR. Additionally, the cross-attention method introduces a higher number of trainable parameters, as evident from Figure 3. Due to this increased complexity, training time is also slightly longer for cross-attention compared to concatenation and addition, as shown in Figure 4.

Among the three fusion techniques, CA consistently demands the most computational resources, both in terms of trainable parameters and training time. Concatenation, while slightly heavier than addition, remains significantly more efficient than CA. Cross-attention requires up to 10% more training time than addition, making it the most computationally expensive fusion approach, while addition remains the most parameter-efficient option.

5. Conclusion

This study investigates the impact of different feature fusion techniques for Voice Activity Detection (VAD) by combining hand-crafted MFCC features with pre-trained model (PTM) features. Our experiments show that simple fusion methods like addition and concatenation consistently outperform the more complex cross-attention mechanism. The results indicate that VAD, being a relatively simple task, does not benefit from attention-based feature fusion, which adds unnecessary computational overhead. Addition emerges as the most effective fusion strategy in four out of six models, while concatenation also performs well. Furthermore, our best fusion model surpasses the state-of-the-art VAD model (Pyannote) with an absolute improvement of DER of 2. 04 % across datasets. These findings highlight that incorporating complementary features using lightweight fusion techniques enhances VAD performance while maintaining efficiency. Future work can explore extending these insights to other speech processing tasks, where the balance between complexity and effectiveness remains crucial.

6. References

[1] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5549–5553.

[2] T. Liu, C. Wang, Z. Li, M.-C. Huang, W. Xu, and F. Lin, "Wavoice: An mmwave-assisted noise-resistant speech recognition system," *ACM Transactions on Sensor Networks*, vol. 20, no. 4, pp. 1–29, 2024.

[3] D. Wang, X. Xiao, N. Kanda, M. Yousefi, T. Yoshioka, and J. Wu, "Profile-error-tolerant target-speaker voice activity detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 906–11 910.

[4] J. Thienpondt and K. Demuynck, "Speaker embeddings with weakly supervised voice activity detection for efficient speaker diarization," *arXiv preprint arXiv:2405.09142*, 2024.

[5] R. M. Patil and C. Patil, "Unveiling the state-of-the-art: A comprehensive survey on voice activity detection techniques," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. IEEE, 2024, pp. 1–5.

[6] S. Alimi and O. Awodele, "Voice activity detection: Fusion of time and frequency domain features with a svm classifier," *Comput. Eng. Intell. Syst*, vol. 13, no. 3, pp. 20–29, 2022.

[7] B. T. Nguyen, Y. Wakabayashi, K. Iwai, and T. Nishiura, "Analysis of derivative of instantaneous frequency and its application to voice activity detection," *Applied Acoustics*, vol. 181, p. 108116, 2021.

[8] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[9] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[10] N. Wilkinson and T. Niesler, "A hybrid cnn-bilstm voice activity detector," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6803–6807.

[11] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection." in *Interspeech*, 2016, pp. 3668–3672.

[12] 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.

[13] 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.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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.

[15] R. Shankar, K. Tan, B. Xu, and A. Kumar, "A closer look at wav2vec2 embeddings for on-device single-channel speech enhancement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 751–755.

[16] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[17] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[19] K. Tripathi, R. Gothi, and P. Wasnik, "Enhancing whisper's accuracy and speed for indian languages through prompt-tuning and tokenization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[20] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *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.

[21] O. C. Phukan, D. Singh, S. R. Behera, A. B. Buduru, and R. Sharma, "Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution," *arXiv preprint arXiv:2412.17796*, 2024.

[22] O. C. Phukan, D. Koshal, S. R. Behera, A. B. Buduru, and R. Sharma, "Multi-view multi-task modeling with speech foundation models for speech forensic tasks," *arXiv preprint arXiv:2410.12947*, 2024.

[23] B. Karan, J. J. van Vüren, F. de Wet, and T. Niesler, "A transformer-based voice activity detector," in *Proc. Interspeech 2024*, 2024, pp. 3819–3823.

[24] M. Kunešová and Z. Zajíc, "Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] S. Gupta, J. Jaafar, W. W. Ahmad, and A. Bansal, "Feature extraction using mfcc," *Signal & Image Processing: An International Journal*, vol. 4, no. 4, pp. 101–108, 2013.

[26] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[27] "2000 nist speaker recognition evaluation," " *https://catalog.ldc.upenn.edu/LDC2001S97*, 2000.

[28] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:2007.01216*, 2020.

[29] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[30] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," *arXiv preprint arXiv:2104.04045*, 2021.

[31] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[32] Z.-H. Tan, N. Dehak *et al.*, "rvad: An unsupervised segment-based robust voice activity detection method," *Computer speech & language*, vol. 59, pp. 1–21, 2020.

[33] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection." in *INTERSPEECH*, 2014, pp. 1534–1538.