Voice Activity Detection: A Comparative Study of Traditional and State-of-the-Art Methods

Aryan Kumar M23CSA510
Abhilash Agarwal M23CSA502
IIT Jodhpur

February 2, 2025

Contents

Abstract

Voice Activity Detection (VAD) is the process of determining whether a given segment of an audio signal contains human speech. This report introduces VAD, explains its importance and applications, describes how it works, and then compares traditional models with modern state-of-the-art methods. Descriptions are provided for key techniques such as Gaussian Mixture Models, Hidden Markov Models, deep neural networks (including LSTM and Transformer models). Finally, the models are compared based on their performance, advantages, and limitations.

1 Introduction

Voice Activity Detection (VAD), also known as speech activity detection, is a technique that distinguishes segments of an audio signal that contain speech from those that do not. In most applications, VAD systems are optimized to detect human speech. However, the underlying idea is to separate the signal of interest (speech) from other background sounds. Although in theory one might adapt a VAD system to detect animal sounds or other noise types, standard VAD systems are trained on human speech and are therefore tailored to its unique features.

VAD is a critical front-end module in many speech processing systems such as:

- **Automatic Speech Recognition (ASR):** Only processing speech segments increases recognition accuracy.
- **Telecommunications (VoIP):** By transmitting only speech segments, bandwidth is saved.
- **Speaker Diarization and Emotion Recognition:** Identifying who is speaking or what emotion is conveyed requires an accurate segmentation of speech.
- **Hearing Aids and Embedded Devices:** Low-power devices rely on efficient processing to conserve battery.

2 How Does VAD Work?

In general, a VAD system follows these basic steps:

1. **Signal Acquisition and Preprocessing:** The continuous audio is divided into short overlapping frames (typically 20–30 milliseconds). A window function (such as a Hamming window) is applied to reduce edge effects.
2. **Feature Extraction:** Features that characterize the audio are computed for each frame. For example, one may calculate the short-term energy or spectral features (such as Mel-frequency cepstral coefficients, or MFCCs).
3. **Classification:** A decision rule (or a trained classifier) is applied to determine whether a frame contains speech. Often, this involves thresholding a computed value or using a statistical or neural model.
4. **Post-processing:** Finally, techniques such as smoothing or hangover schemes help produce continuous speech segments by avoiding rapid switches between speech and non-speech.

3 Traditional VAD Models

Traditional VAD approaches include energy-based and statistical methods. In this section, we describe these methods in detail.

3.1 Energy-Based Methods

Overview: These methods compute the energy of each frame and use a threshold to decide whether it contains speech.

Steps:

1. **Framing and Windowing:** The audio signal is split into frames using, for example, a Hamming window.
2. **Energy Computation:** The short-term energy E for a frame with samples $x[n]$ is computed as:
$$E = \sum_{n=1}^{N} x[n]^2$$
3. **Thresholding:** A predetermined or adaptive threshold is applied. If E exceeds the threshold, the frame is classified as speech.
4. **Post-Processing:** Techniques (such as hangover schemes) smooth the decisions to avoid sudden changes.

Pros: Simple, fast, and computationally inexpensive.

Cons: Sensitive to noise; may misclassify low-energy speech as silence.

3.2 Statistical Models: GMM-HMM

Overview: Statistical models use probability distributions to represent the speech and non-speech classes. The Gaussian Mixture Model (GMM) is often combined with a Hidden Markov Model (HMM) to also capture temporal information.

Gaussian Mixture Model (GMM):

- For a feature vector $\mathbf{x} \in \mathbb{R}^d$ (for example, MFCCs), a single Gaussian probability density function (pdf) is:
$$p(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$
- A GMM models the overall distribution as a weighted sum of K Gaussians:
$$p(\mathbf{x}) = \sum_{i=1}^{K} w_i \, p(\mathbf{x} \mid \mu_i, \Sigma_i)$$
where w_i are the mixture weights and $\sum_{i=1}^{K} w_i = 1$.

Hidden Markov Model (HMM):

- The HMM models the sequence of frames with states such as *speech* and *non-speech*.
- Each state has transition probabilities a_{ij} (the probability of moving from state i to state j).
- The emission probability from each state is modeled by a GMM.
- The forward algorithm computes the likelihood $\alpha_t(j)$ of being in state j at time t:

$$\alpha_t(j) = \left[\sum_i \alpha_{t-1}(i)\, a_{ij} \right] b_j(x_t)$$

  where $b_j(x_t)$ is the probability of the feature x_t in state j.

Pros: Improved robustness in moderate noise; incorporates temporal dynamics.

Cons: Requires manual feature engineering and careful tuning.

4 State-of-the-Art (SOTA) Models

Modern VAD systems use deep learning to automatically learn robust features from data. Here we discuss deep neural networks, including convolutional, recurrent, and transformer-based models.

4.1 Deep Neural Network (DNN) Based Models

Overview: These models use fully connected networks or convolutional neural networks (CNNs) to classify each frame based on features extracted from the audio.

Processing Flow:

1. **Feature Extraction:** The audio is segmented into frames, and features such as log-Mel spectrograms or MFCCs are computed.
2. **Neural Network Layers:** For each frame, a feature vector $\mathbf{x}_t$ is processed by layers of the network:

$$\mathbf{h}_t^{(1)} = f\Big(\mathbf{W}^{(1)}\mathbf{x}_t + \mathbf{b}^{(1)}\Big)$$

$$\mathbf{h}_t^{(l)} = f\Big(\mathbf{W}^{(l)}\mathbf{h}_t^{(l-1)} + \mathbf{b}^{(l)}\Big), \quad l = 2, \ldots, L-1$$

3. **Output Layer:** The final layer outputs scores $\mathbf{z}_t$ which are converted to probabilities using the softmax function:

$$y_{t,k} = \frac{\exp(z_{t,k})}{\sum_j \exp(z_{t,j})}$$

4. **Training:** The network is trained using a cross-entropy loss function:

$$\mathcal{L} = -\sum_t \sum_k \mathbf{1}\{y_t^{\text{true}} = k\} \log\big(y_{t,k}\big)$$

Explanation: The neural network learns to map the input features to a probability of speech presence. The function $f(\cdot)$ is a non-linear activation (e.g., ReLU), which helps the network learn complex patterns.

4.2 Recurrent Neural Networks (RNNs) and LSTMs

Overview: RNNs, particularly Long Short-Term Memory (LSTM) networks, capture the time dependencies between frames.

Key Equations (LSTM):

$$\begin{aligned}
\mathbf{i}_t &= \sigma\Big(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i\Big) && \text{(input gate)}\\
\mathbf{f}_t &= \sigma\Big(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f\Big) && \text{(forget gate)}\\
\mathbf{o}_t &= \sigma\Big(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{b}_o\Big) && \text{(output gate)}\\
\tilde{\mathbf{c}}_t &= \tanh\Big(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c\Big) && \text{(cell candidate)}\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t && \text{(cell state)}\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) && \text{(hidden state)}
\end{aligned}$$

Explanation: Here, σ denotes the sigmoid function, and $\odot$ is element-wise multiplication. These equations allow the LSTM to decide which information to keep or discard over time.

4.3 Transformer-Based Models

Overview: Transformers use self-attention to capture long-range dependencies without relying on recurrence.

Self-Attention Equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}\Big)V$$

Where:

- Q (queries), K (keys), and V (values) are linear projections of the input features.
- d_k is the dimension of the key vectors.

Feed-Forward Network in Transformer:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Explanation: Self-attention allows each frame to weigh information from other frames. This is especially helpful in noisy conditions where context helps decide if a frame contains speech.

Whisper: A fully supervised, transformer-based ASR model that also performs VAD implicitly. It uses 80-channel log-Mel spectrograms and an encoder-decoder architecture.

wav2vec2: A self-supervised model that learns speech representations from raw audio. It uses a convolutional feature encoder followed by a transformer network and is fine-tuned using Connectionist Temporal Classification (CTC) loss.

4.4 Personalized VAD

Overview: These systems combine standard VAD with speaker verification to detect whether the speech belongs to a target speaker.

1. A speaker embedding is generated (using, for example, a d-vector model).
2. The VAD system then combines the speech probability with a similarity score (often computed via cosine similarity) between the current frame's embedding and the target speaker's embedding.

Explanation: This additional step ensures that only speech from the desired (target) speaker is flagged as speech, which is useful in multi-speaker environments.

5 Comparison of VAD Models

The following table summarizes the main advantages and disadvantages:

Method	Advantages	Disadvantages
Energy-Based	Simple, fast, low computational cost	Sensitive to noise, may miss low-energy speech
GMM-HMM	Incorporates probabilistic modeling and temporal dynamics	Requires hand-crafted features, tuning complexity
DNN/CNN/RNN	Learns discriminative features automatically, high accuracy	Data-hungry, higher computational cost
Transformer	Models long-range dependencies effectively, robust in noisy conditions	Very high computational cost, complex
Personalized VAD	Distinguishes target speaker from background speech	Requires enrollment phase, additional complexity

Table 1: Comparison of VAD methods.

6 Performance Metrics for VAD

VAD systems are evaluated both at the overall level and on a per-frame basis. Common metrics include:

6.1 Overall Metrics

- **Accuracy:**
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$
where TP and TN are the true positives and true negatives, and FP and FN are the false positives and false negatives.

- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP},$$
which measures the proportion of frames predicted as speech that are actually speech.
- **Recall:**
$$\text{Recall} = \frac{TP}{TP + FN},$$
which measures the proportion of actual speech frames that were correctly detected.
- **F1 Score:**
$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

6.2 Frame-Level Metrics

Since VAD decisions are made on a frame-by-frame basis (e.g., every 20–30 ms), we also compute:

- **Detection Error Rate (DetER):**
$$\text{DetER} = \frac{N_{FA} + N_{Miss}}{N_{Total}},$$
where N_{FA} is the number of false alarms (non-speech frames labeled as speech) and N_{Miss} is the number of missed detections (speech frames labeled as non-speech).
- **False Alarm Rate (FAR):** The proportion of non-speech frames misclassified as speech.
- **Missed Detection Rate (Miss):** The proportion of speech frames that are missed.

7 Comparison of VAD Models Using Various Metrics

Table 2 summarizes approximate performance values reported in the literature for different VAD approaches. Note that these numbers are illustrative; actual values depend on test conditions (e.g., clean vs. noisy).

Discussion:

- *Energy-based methods* are very fast but suffer from lower accuracy, particularly in noisy environments.
- *GMM-HMM models* improve over energy-based approaches by modeling statistical distributions, yielding higher accuracy and lower detection error rates.
- *Deep learning models* (LSTM, CNN, Transformer) learn features automatically and achieve further improvements in precision and recall.
- Among SOTA approaches, *Transformer-based models* (as used in wav2vec2) and *Whisper* typically provide the best performance in terms of accuracy, precision, recall, and F1 score, along with low frame-level detection error rates.

Method	Accuracy	Precision	Recall	F1	DetER	FAR / Miss
Energy-Based	70%	72%	68%	70%	∼30%	15% / 15%
GMM-HMM	80%	82%	78%	80%	∼20%	10% / 10%
LSTM-Based	88%	90%	86%	88%	∼12%	8% / 10%
CNN-Based	90%	91%	89%	90%	∼10%	7% / 8%
Transformer (wav2vec2)	92%	93%	91%	92%	∼8%	5% / 7%
Whisper	93%	94%	92%	93%	∼7%	5% / 6%

Table 2: Approximate performance metrics of various VAD methods. Values are illustrative and depend on test conditions.

8 Open Problems & Future Directions

- *Real-time Low-Latency VAD* Reducing computation overhead for embedded devices.
- *Robustness to Noisy & Adverse Environments* Improving generalization across different recording conditions.
- *Low-Resource & Few-Shot Learning* Reducing dependency on large datasets for training VAD models.
- *Multilingual & Dialect Adaptability* Enhancing VAD performance for diverse linguistic datasets.

9 Conclusion

In this report, we have defined Voice Activity Detection (VAD) as the process of distinguishing speech from non-speech in audio signals—primarily focused on human speech. We discussed its applications, from speech recognition to telecommunication systems. Traditional methods such as energy-based approaches and statistical models (GMM-HMM) have been explained with corresponding mathematical formulas. We then covered modern state-of-the-art methods including deep neural networks, recurrent neural networks, transformer-based models, and personalized VAD systems, each with its mathematical details and an explanation of its operation.

While traditional methods are simple and computationally inexpensive, they often fall short in noisy environments. Deep learning methods, although more resource intensive, provide better robustness and higher accuracy. The choice of model thus depends on the specific application requirements such as real-time performance and available computational resources.

10 References

1. Wikipedia, "Voice activity detection," `https://en.wikipedia.org/wiki/Voice_activity_detection`.

2. Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1), 1–3.

3. Kumar, S., Buddi, S. S., Sarawgi, U. O., Garg, V., Ranjan, S., Rudovic, O., Abdelaziz, A. H., & Adya, S. (2024). Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness. *arXiv preprint arXiv:2406.09443*.

4. Bovbjerg, H. S., Jensen, J., Østergaard, J., & Tan, Z.-H. (2024). Self-Supervised Pretraining for Robust Personalized Voice Activity Detection in Adverse Conditions. *arXiv preprint arXiv:2312.16613*.

5. Wang, J.-Y., Zhang, J., & Dai, L.-R. (2023). Real-Time Causal Spectro-Temporal Voice Activity Detection Based on Convolutional Encoding and Residual Decoding. In *Interspeech 2023*.

6. Chakraborty, A. (2020). Voice Activity Detection Analysis. *International Research Journal of Modernization in Engineering Technology and Science*, 2(6).

7. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, 33, 12449–12460.

8. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *OpenAI Technical Report*.

9. Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2020). HuBERT: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460.