Implementasi Hybrid Voice Activity Detection untuk Deteksi Aktivitas Suara Mahasiswa di Lingkungan Kelas

Andhika Chahya Syahputra
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
andhikacs@student.telkomuniversity.ac.id

Kris Sujatmoko, S.T., M.T.
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
Krissujatmoko@telkomuniversity.ac.id

Dr. Ir. Bambang Setia Nugroho, S.T., M.T.
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
bambangsetianugroho@telkomuniversity.ac.id

***Abstrak* — Keterlibatan mahasiswa dalam proses pembelajaran sulit dinilai secara objektif oleh dosen karena metode yang digunakan masih bersifat manual dan subjektif, terutama pada kondisi kelas dengan jumlah mahasiswa besar. Salah satu komponen krusial dalam sistem pemantauan keterlibatan berbasis suara adalah kemampuan memisahkan segmen ucapan dari segmen non-suara secara akurat sebelum data diproses lebih lanjut. Penelitian ini berfokus pada pengembangan dan evaluasi modul Voice Activity Detection (VAD) sebagai tahap pra-pemrosesan pada sistem deteksi suara mahasiswa berbasis kecerdasan buatan. Metode yang diterapkan adalah Hybrid VAD, yaitu kombinasi Silero VAD berbasis deep learning yang terintegrasi dalam Faster-Whisper dengan pendekatan energy-based threshold sebesar 20 desibel, untuk memisahkan aktivitas suara mahasiswa dari keheningan maupun gangguan suara latar (background noise) secara real-time tanpa membebani komputasi sistem. Pengujian dilakukan pada empat skenario, yaitu kondisi kelas normal (55–60 dB), kondisi kelas dengan background noise (65–80 dB), kondisi hening, dan suara berdurasi singkat (≤1 detik), masing-masing menggunakan 25 sampel audio. Hasil pengujian menunjukkan modul VAD mampu mendeteksi seluruh aktivitas suara dengan tepat pada keempat skenario, mencapai akurasi 100%, melampaui target minimum yang ditetapkan sebesar 75%. Temuan ini mengindikasikan bahwa pendekatan Hybrid VAD efektif diterapkan sebagai tahap penyaringan awal pada sistem pemantauan keterlibatan mahasiswa berbasis suara di lingkungan kelas nyata.*

***Kata kunci* — Voice Activity Detection, Hybrid VAD, Silero VAD, energy-based threshold, deteksi aktivitas suara**

I. PENDAHULUAN

Keterlibatan mahasiswa dalam proses pembelajaran merupakan faktor penting yang menentukan efektivitas pendidikan tinggi, mencakup dimensi perilaku, emosional, dan kognitif yang saling berkaitan [1]. Namun pada kenyataannya, dosen masih menghadapi kendala dalam menilai keterlibatan mahasiswa secara objektif, terutama pada kelas berukuran besar, karena metode penilaian yang digunakan saat ini masih bersifat manual dan rentan terhadap bias personal maupun institusional [2]. Kondisi ini mendorong berkembangnya pendekatan pemantauan berbasis audio, yang menganalisis aktivitas suara mahasiswa sebagai indikator partisipasi tanpa bergantung pada observasi visual dosen secara langsung.

penerapan sistem berbasis audio di lingkungan kelas nyata menghadapi tantangan akustik yang signifikan. Kelas dengan jumlah mahasiswa yang banyak umumnya memiliki tingkat kebisingan yang cukup tinggi, ditambah pantulan suara ruangan (*reverberation*) ucapan ketika lebih satu mahasiswa berbicara secara bersamaan [3]. Kondisi ini menyebabkan sinyal suara yang ditangkap mikrofon menjadi kurang bersih, sehingga berpotensi menurunkan akurasi sistem dalam mendeteksi maupun memisahkan segmen ucapan mahasiswa dari derau lingkungan, khususnya pada tahap paling awal pemrosesan sinyal [4]. Apabila tahap awal ini gagal memisahkan ucapan dari derau secara tepat, seluruh proses analisis lanjutan berpotensi ikut terganggu, karena kualitas input yang buruk akan terbawa ke setiap tahap berikutnya.

penerapan sistem berbasis audio di lingkungan kelas nyata menghadapi tantangan akustik yang signifikan. Kelas dengan jumlah mahasiswa yang banyak umumnya memiliki tingkat kebisingan latar belakang yang tinggi, ditambah pantulan suara ruangan dan ucapan ketika lebih dari satu mahasiswa berbicara secara bersamaan [3]. Kondisi ini menyebabkan sinyal suara yang ditangkap mikrofon menjadi kurang bersih, sehingga berpotensi menurunkan akurasi sistem dalam memisahkan segmen ucapan dari lingkungan, terutama pada tahap paling awal pemrosesan sinyal [4].

Voice Activity Detection (VAD) berperan sebagai tahap pra-pemrosesan yang menentukan keberhasilan seluruh pipeline pengolahan suara, karena bertugas memisahkan segmen yang mengandung ucapan dari segmen hening maupun *noise* sebelum data diteruskan ke tahap analisis selanjutnya [5]. Apabila segmen non-ucapan ikut diproses, tahap analisis berikutnya dapat menerima informasi yang tidak relevan sehingga menurunkan kualitas hasil secara keseluruhan [5]. Pendekatan VAD berbasis *energy threshold* konvensional unggul dari sisi kecepatan komputasi dan kemudahan implementasi lokal, tetapi cenderung kurang andal pada kondisi derau tinggi maupun ucapan berdurasi sangat singkat, karena ambang batas energi statis sulit membedakan derau bernada keras dari suara ucapan yang lemah. Sebaliknya, pendekatan VAD berbasis *deep learning* terbukti lebih tangguh terhadap variasi kondisi akustik

dibandingkan metode berbasis energi maupun statistik konvensional, meski umumnya membutuhkan beban komputasi yang lebih besar [6].

Berdasarkan hal tersebut, penelitian ini bertujuan merancang dan mengevaluasi modul VAD dengan pendekatan Hybrid VAD, yang menggabungkan Silero VAD dengan *energy-based threshold* untuk memisahkan aktivitas suara mahasiswa dari derau latar secara presisi pada kondisi kelas nyata, sekaligus mengevaluasi kontribusinya terhadap efisiensi pemrosesan real-time, sejalan dengan UU No. 27 Tahun 2022 tentang Pelindungan Data Pribadi [8].

II. KAJIAN TEORI

Kajian Teori membahas konsep yang mendasari penelitian Implementasi Hybrid Voice Activity Detection untuk Deteksi Aktivitas Suara Mahasiswa di Lingkungan Kelas. Meliputi Keterlibatan Mahasiswa Berbasis Audio, Voice Activity Detection, Energy-based VAD, Deep Learning-based VAD, dan Hybrid VAD.

A. Keterlibatan Mahasiswa Berbasis Audio

Keterlibatan mahasiswa merupakan faktor penting dalam keberhasilan pembelajaran, namun penilaian manual oleh dosen masih menghadapi kendala objektivitas, terutama pada kelas besar [1], [2]. Pemantauan berbasis audio menjadi solusi, karena aktivitas suara mahasiswa dapat dijadikan indikator partisipasi yang terukur secara otomatis.

B. Voice Activity Detection

VAD adalah proses mendeteksi bagian audio yang mengandung ucapan dan memisahkannya dari keheningan atau *noise* [5]. VAD merupakan tahap pra-pemrosesan krusial: jika segmen non-ucapan ikut diproses, kualitas analisis lanjutan (transkripsi, identifikasi pembicara, klasifikasi emosi) dapat menurun [5]. Pada lingkungan kelas, VAD menghadapi tantangan berupa kebisingan latar, pantulan ruangan.

C. Energy-Based VAD

Pendekatan ini menetapkan ambang batas energi sinyal (misalnya 20 dB) untuk memangkas segmen hening. Kelebihannya ringan secara komputasi dan mudah diimplementasikan lokal, tetapi kurang andal pada derau tinggi atau ucapan berdurasi singkat karena ambang statisnya sulit membedakan derau keras dari suara lemah.

D. Deep Learning-based VAD

Memanfaatkan *neural network* untuk mengenali pola akustik ucapan secara lebih adaptif [6]. Silero VAD, model ringan yang terintegrasi dalam Faster-Whisper, adalah contohnya [7]. Pendekatan ini lebih tangguh terhadap variasi akustik, namun performa tunggalnya belum tentu konsisten pada kondisi kelas yang dinamis

E. Hybrid VAD

Menggabungkan Silero VAD dengan energy threshold sebagai lapisan penyaring tambahan, sehingga kelemahan satu metode ditutupi metode lainnya. Model deep learning menangani variasi akustik kompleks, sementara energy threshold memangkas sisa keheningan secara ringan sebelum ekstraksi fitur menjaga sistem tetap responsif real-time dan lebih tangguh dibanding pendekatan tunggal

III. METODE

A. Batasan Penelitian

Data Penelitian ini dibatasi pada ruang lingkup berikut:

1. Pengujian dilakukan pada ruang kelas offline dengan jumlah mahasiswa sekitar 25–30 orang.
2. Input suara diperoleh melalui mikrofon yang terhubung ke laptop atau PC, tanpa melibatkan data visual sebagai pendukung.
3. Modul VAD diimplementasikan menggunakan bahasa pemrograman Python dan dijalankan secara lokal (*offline*) tanpa layanan *cloud* berbayar.
4. Evaluasi performa dibatasi pada parameter akurasi, *precision*, dan *recall* deteksi aktivitas suara, dan tidak mencakup tahap lanjut seperti transkripsi, identifikasi pembicara, atau klasifikasi emosi

B. Sumber Data dan Cara Perolehan Data

Data audio dikumpulkan secara mandiri melalui rekaman langsung suara mahasiswa di ruang kelas menggunakan mikrofon yang terhubung ke perangkat laptop/PC. Proses pengumpulan data diawali dengan mikrofon menangkap sinyal audio secara *real-time* selama sesi kelas berlangsung. Sinyal audio yang diterima kemudian diseragamkan frekuensi cuplikannya (*sample rate*) menjadi 16.000 Hz untuk menjaga konsistensi data sebelum diproses oleh modul VAD.

Pengujian dilakukan pada empat skenario yang mewakili kondisi penggunaan sistem di dalam kelas, yaitu: (1) kondisi kelas normal dengan tingkat kebisingan 55–60 dB, (2) kondisi kelas dengan derau latar (*background noise*) 65–80 dB, (3) kondisi hening tanpa aktivitas suara, dan (4) ucapan berdurasi singkat (≤ 1 detik). Setiap skenario diuji menggunakan 25 sampel audio, sehingga total data uji berjumlah 100 sampel.

C. Prosedur Hybrid VAD

Prosedur kerja modul VAD pada penelitian ini terdiri atas dua lapisan penyaringan yang bekerja secara berurutan:

1. Lapisan deep learning sinyal audio diproses menggunakan Silero VAD yang terintegrasi dalam Faster-Whisper) untuk mengidentifikasi segmen yang mengandung pola akustik ucapan manusia secara adaptif terhadap variasi kondisi lingkungan.
2. Lapisan energy-based threshold sebagai penyaring tambahan, segmen dengan level energi di bawah ambang batas 20 desibel dipangkas dari awal dan akhir sinyal, guna memastikan hanya segmen suara yang valid yang diteruskan ke tahap ekstraksi fitur

D. Persamaan

Pengujian Performa modul VAD dievaluasi menggunakan *confusion matrix*, yang membandingkan hasil deteksi sistem terhadap kondisi aktual pada setiap sampel uji, dengan kategori sebagai berikut:

a.True Positive (TP): sampel yang mengandung suara dan terdeteksi benar.

b.True Negative (TN): sampel hening yang dikenali benar sebagai tidak ada aktivitas suara.

c.False Positive (FP): sampel hening yang keliru terdeteksi sebagai suara.

d.False Negative (FN): sampel bersuara yang gagal terdeteksi.

Berdasarkan keempat nilai tersebut, dihitung tiga metrik evaluasi sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

Accuracy menunjukkan persentase prediksi benar secara keseluruhan, *precision* menunjukkan proporsi prediksi positif yang benar-benar tepat, sedangkan *recall* menunjukkan kemampuan sistem dalam menangkap seluruh sampel yang sebenarnya mengandung suara

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Pengujian Pengujian modul VAD dilakukan pada empat skenario akustik yang merepresentasikan kondisi penggunaan sistem di dalam kelas, masing-masing menggunakan 25 sampel audio, sehingga total data uji berjumlah 100 sampel. Ringkasan hasil pengujian ditampilkan pada Tabel 1.

TABEL 1

Hasil Pengujian Voice Activity Detection pada Empat Skenario

Kategori	Jumlah	Prediksi Benar	Akurasi
Kondisi kelas normal (55–60 dB)	25	*25*	100%
Kondisi derau latar (65–80 dB)	25	*25*	100%
Kondisi Hening	25	*25*	100%
Ucapan singkat (≤ 1 detik)	25	*25*)	100%
Total	100	100	100%

Pada skenario kondisi kelas normal, seluruh 25 sampel suara mahasiswa berhasil terdeteksi tanpa kegagalan, menunjukkan bahwa lapisan *energy-based threshold* mampu membedakan sinyal ucapan dari kondisi hening secara efektif pada tingkat kebisingan rendah. Pada skenario derau latar 65–80 dB, sistem tetap mendeteksi seluruh sampel dengan benar meskipun tingkat kebisingan meningkat signifikan, mengindikasikan bahwa lapisan Silero VAD berkontribusi menjaga ketahanan deteksi terhadap derau lingkungan yang tidak dapat ditangani optimal oleh ambang energi statis saja. Pada skenario kondisi hening, sistem tidak menghasilkan satu pun *false positive* dari 25 sampel uji, menunjukkan kemampuan menekan kesalahan deteksi ketika benar-benar tidak ada aktivitas suara. Pada skenario ucapan singkat (≤ 1 detik), seluruh sampel tetap terdeteksi tanpa *false negative*, mengindikasikan sensitivitas sistem terhadap durasi ucapan pendek yang umum terjadi saat mahasiswa memberi respons spontan di kelas.

B. Analisis Confusion Matrix

Distribusi hasil deteksi terhadap seluruh 100 sampel uji dirangkum dalam *confusion matrix* berikut:

TABEL 2

Confusion Matrix Hasil Pengujian VAD

	Prediksi: Suara	Prediksi: Hening
Aktual: Suara	TP = 75	FN = 0
Aktual: Hening	FP = 0	TN = 25

Berdasarkan Tabel 2, diperoleh nilai:

$$Accuracy = \frac{75 + 25}{75 + 25 + 0 + 0} \times 100\% = 100\%$$

$$Precision = \frac{75}{75 + 0} \times 100\% = 100\%$$

$$Recall = \frac{75}{75 + 0} \times 100\% = 100\%$$

Tidak ditemukannya *false positive* maupun *false negative* pada seluruh skenario menunjukkan bahwa kombinasi Silero VAD dan *energy-based threshold* pada skema Hybrid VAD bekerja secara saling melengkapi: lapisan *deep learning* menangani variasi pola akustik yang kompleks (derau tinggi, ucapan singkat), sementara lapisan *energy threshold* menyaring sisa keheningan secara ringan tanpa membebani komputas

C. Pembahasan

Nilai akurasi 100% yang diperoleh pada seluruh skenario pengujian melampaui target minimum yang ditetapkan sebesar 75%, dan mengonfirmasi bahwa kombinasi Silero VAD dengan energy-based threshold saling melengkapi sesuai dengan karakteristik masing-masing pendekatan: model deep learning menangani variasi pola akustik yang kompleks pada kondisi derau tinggi, sebagaimana ditunjukkan pada penelitian sebelumnya mengenai ketahanan VAD berbasis deep learning terhadap derau lingkungan [6], sementara energy threshold berfungsi sebagai lapisan penyaring tambahan yang ringan secara komputasi untuk memangkas sisa keheningan pada kondisi akustik yang lebih bersih.

Meskipun demikian, hasil ini diperoleh pada skala pengujian yang masih terbatas, yaitu 25 sampel per skenario dengan kalimat uji yang telah disediakan sebelumnya dan jumlah pembicara yang terbatas. Kondisi tersebut belum sepenuhnya merepresentasikan variasi kelas nyata yang lebih spontan, seperti ucapan saat diskusi berlangsung. Nilai ambang energi 20 desibel yang digunakan juga dikalibrasi

khusus untuk kondisi pengujian pada penelitian ini, sehingga penyesuaian ulang kemungkinan diperlukan apabila diterapkan pada ruang kelas dengan karakteristik akustik berbeda. Dengan demikian, hasil pengujian ini menunjukkan kelayakan awal pendekatan Hybrid VAD pada kondisi yang diuji, dan membuka ruang bagi pengujian pada skala yang lebih luas

V. KESIMPULAN

Penelitian ini berhasil merancang dan mengevaluasi modul Voice Activity Detection dengan pendekatan Hybrid VAD, yang menggabungkan Silero VAD berbasis deep learning (terintegrasi dalam Faster-Whisper) dengan energy-based threshold sebesar 20 desibel, sebagai tahap pra-pemrosesan pada sistem deteksi suara mahasiswa berbasis kecerdasan buatan. Pendekatan ini dirancang agar kedua lapisan saling menutupi kelemahan satu sama lain: model deep learning menangani variasi pola akustik yang kompleks, sementara energy threshold memangkas sisa keheningan secara ringan tanpa membebani komputasi.

Hasil pengujian pada empat skenario akustik kelas — kondisi normal (55–60 dB), derau latar (65–80 dB), hening, dan ucapan berdurasi singkat (≤1 detik) — menunjukkan bahwa modul VAD mampu mendeteksi seluruh 100 sampel uji dengan benar, mencapai akurasi, precision, dan recall sebesar 100%, melampaui target minimum yang ditetapkan sebesar 75%. Temuan ini mengindikasikan bahwa pendekatan Hybrid VAD berpotensi efektif diterapkan sebagai tahap penyaringan awal pada sistem pemantauan keterlibatan mahasiswa berbasis suara di lingkungan kelas nyata, meskipun hasil tersebut masih diperoleh dari skala pengujian yang terbatas (25 sampel per skenario, kondisi ucapan terstruktur) sehingga generalisasi pada kondisi kelas yang lebih luas dan spontan masih perlu diuji lebih lanjut. Pengembangan selanjutnya diarahkan pada perluasan skala pengujian, evaluasi pada kondisi ucapan tumpang tindih, penerapan ambang energi adaptif, serta perbandingan dengan model VAD lain guna memperkuat keandalan dan generalisasi sistem.

REFERENSI

[1] R. Das and S. Dev, "Optimizing student engagement detection using facial and behavioral features," *Neural Comput. Appl.*, vol. 37, no. 23, pp. 19063–19085, 2025, doi: 10.1007/s00521-025-11317-z.

[2] P. D. Simon, L. K. Fryer, and K. Nakao, "Assessing class participation in physical and virtual spaces: current approaches and issues," *Front. Educ.*, vol. 8, pp. 1–6, 2023, doi: 10.3389/feduc.2023.1306568.

[3] R. Southwell *et al.*, "Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms," *Proc. 15th Int. Conf. Educ. Data Mining, EDM 2022*, pp. 302–315, 2022, doi: 10.5281/zenodo.6853109.

[4] A. S. Khan, T. Ógúnrèmí, A. Attia, and D. Demszky, "Multi-Stage Speaker Diarization for Noisy Classrooms," *Proc. Int. Conf. Educ. Data Min.*, pp. 473–481, 2025, doi: 10.5281/zenodo.15870278.

[5] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 8, pp. 2624–2632, 2011, doi: 10.1109/TASL.2011.2125953.

[6] C. M. Larsen, P. Koch, and Z.-H. Tan, "Adversarial Multi-Task Deep Learning for Noise-Robust Voice Activity Detection with Low Algorithmic Delay," *arXiv:2207.01691*, 2022.

[7] Silero Team, "Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier," GitHub repository, 2021. [Online]. Available: https://github.com/snakers4/silero-vad

[8] Pemerintah Pusat, "Undang-undang (UU) Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi," 2022. [Online]. Available: https://peraturan.bpk.go.id/Download/224884/UU Nomor 27 Tahun 2022.pdf