Evaluasi Performa Faster-Whisper untuk Automatic Speech Recognition Bahasa Indonesia pada Lingkungan Perkuliahan

Nayla Nabila
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
naylanabila@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* — Sistem deteksi keterlibatan mahasiswa berbasis pemrosesan suara memerlukan modul Speech-to-Text (STT) yang mampu mentranskripsikan ucapan mahasiswa secara akurat pada kondisi kelas yang dinamis. Penelitian ini mengevaluasi performa model Faster-Whisper varian Small sebagai modul Automatic Speech Recognition (ASR) berbahasa Indonesia yang dijalankan secara lokal (offline) tanpa bergantung pada layanan Cloud API pihak ketiga, menggunakan tipe komputasi int8 pada CPU serta injeksi prompt kontekstual berisi kosakata akademik teknik telekomunikasi untuk membantu pengenalan istilah teknis. Pengujian dilakukan pada 54 sampel ucapan mahasiswa di lingkungan kelas nyata, meliputi kondisi relatif tenang maupun kondisi dengan gangguan suara latar, dan dievaluasi menggunakan metrik Word Error Rate (WER). Hasil pengujian menunjukkan sistem mencapai nilai total WER sebesar 13,04% atau Word Accuracy sebesar 86,96% dari total 759 kata referensi, serta Real-Time Factor (RTF) rata-rata sebesar 0,50 dengan latensi rata-rata 2,60 detik. Hasil ini membuktikan bahwa Faster-Whisper varian Small mampu beroperasi mendekati waktu nyata (real-time) dengan tingkat akurasi transkripsi yang memadai untuk mendukung evaluasi keterlibatan mahasiswa dalam lingkungan perkuliahan.**

Kata kunci: Faster-Whisper, Automatic Speech Recognition, Speech-to-Text, Word Error Rate, Bahasa Indonesia.

I. PENDAHULUAN

Perkembangan Keterlibatan mahasiswa dalam proses pembelajaran merupakan faktor krusial yang menentukan efektivitas pendidikan tinggi, dan dipengaruhi oleh berbagai faktor kompleks termasuk metode pengajaran, motivasi, serta kondisi emosional mahasiswa selama pembelajaran berlangsung [1]. Namun pada kenyataannya, dosen sering kali menghadapi kendala dalam menilai tingkat keaktifan mahasiswa secara objektif, terutama pada kelas dengan jumlah mahasiswa yang banyak dan aktivitas diskusi yang berlangsung secara dinamis, sehingga penilaian partisipasi yang dilakukan secara manual masih menghadapi berbagai isu terkait kriteria, konsistensi, dan objektivitas [2].

Salah satu pendekatan yang memungkinkan pemantauan keterlibatan mahasiswa secara lebih objektif dan terukur adalah dengan menganalisis suara mahasiswa selama proses pembelajaran, di mana ucapan mahasiswa perlu diubah menjadi teks secara otomatis melalui teknologi Speech-to-Text (STT) sebelum dianalisis lebih lanjut. Beberapa pendekatan yang umum digunakan untuk keperluan tersebut memiliki kekurangan masing-masing. Pendekatan berbasis layanan Cloud API pihak ketiga (seperti Google Cloud Speech-to-Text atau AWS Transcribe) memang terbukti memiliki akurasi transkripsi yang kompetitif antar penyedia dalam pengujian benchmarking [5], namun pendekatan ini memiliki ketergantungan mutlak pada koneksi internet yang stabil (latensi tinggi jika sinyal buruk), menimbulkan masalah privasi karena data suara mahasiswa dikirim ke server pihak ketiga, dan memerlukan biaya operasional berbayar berdasarkan durasi pemakaian. Padahal, berdasarkan Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi, data suara termasuk data pribadi yang memerlukan persetujuan eksplisit dan perlindungan yang memadai sebelum digunakan [3].

Selain itu, kondisi akustik ruang kelas turut menjadi tantangan tersendiri. Kelas dengan jumlah mahasiswa yang cukup banyak biasanya memiliki tingkat kebisingan latar belakang yang tinggi, ditambah pantulan suara di dalam ruangan dan suara yang saling bertumpuk ketika lebih dari satu mahasiswa berbicara hampir bersamaan. Kondisi ini membuat sinyal suara yang ditangkap mikrofon menjadi kurang bersih, sehingga berpotensi menurunkan akurasi sistem dalam mentranskripsikan ucapan. Oleh karena itu, diperlukan solusi STT yang dapat dijalankan secara lokal (offline) pada perangkat tanpa bergantung pada layanan cloud berbayar, sekaligus tetap mampu menghasilkan transkripsi ucapan yang akurat pada berbagai kondisi lingkungan kelas.

Model Speech-to-Text Faster-Whisper, yang dilatih menggunakan data audio multibahasa berskala besar, dipilih sebagai solusi karena mampu menghasilkan transkripsi ucapan secara akurat pada berbagai kondisi lingkungan tanpa memerlukan layanan cloud berbayar. Model ini merupakan hasil optimasi dari model Whisper yang dikembangkan oleh

OpenAI, melalui teknik kuantisasi yang menekan kebutuhan memori tanpa banyak mengorbankan akurasi. Penelitian ini secara khusus bertujuan untuk mengevaluasi performa model Faster-Whisper varian Small dalam mentranskripsikan ucapan mahasiswa berbahasa Indonesia pada lingkungan perkuliahan nyata, ditinjau dari aspek akurasi transkripsi (Word Error Rate) maupun efisiensi komputasi (Real-Time Factor dan latensi sistem), sebagai bagian dari evaluasi modul Speech-to-Text pada sistem deteksi suara mahasiswa yang lebih luas.

II. KAJIAN TEORI

Menyajikan dan menjelaskan teori-teori yang berkaitan dengan variabel-variabel penelitian. Poin subjudul ditulis dalam abjad.

A. Automatic Speech Recognition dan Speech-to-Text

Speech-to-text adalah proses mengubah ucapan atau suara manusia menjadi transkrip dalam bentuk teks, yang juga dikenal sebagai Voice-to-Text. Teknologi ini bekerja dengan memanfaatkan Automatic Speech Recognition (ASR) yang didukung oleh kecerdasan buatan untuk mengenali pola-pola suara dan menerjemahkannya ke dalam bentuk tulisan. Pada prosesnya, sistem terlebih dahulu menangkap sinyal suara yang masuk melalui mikrofon, kemudian menganalisis karakteristik suara tersebut menggunakan berbagai algoritma pemrosesan bahasa sehingga dapat dihasilkan kata, kalimat, maupun frasa dalam format digital [8]. Secara lebih teknis, ASR merupakan teknologi yang memungkinkan komputer mengenali, memahami, dan mengubah ucapan manusia menjadi bentuk teks secara otomatis. Sistem ASR bekerja dengan cara menerima sinyal audio sebagai input, mengekstraksi fitur akustik darinya (misalnya melalui teknik ekstraksi parameter fitur suara [4]), kemudian menghasilkan urutan kata atau teks yang paling mungkin sesuai dengan apa yang diucapkan [8]. Secara historis, sistem ASR dibangun menggunakan kombinasi Hidden Markov Model (HMM) dan Gaussian Mixture Model (GMM) sebagai model utama dalam pengenalan ucapan. Namun, pendekatan berbasis deep learning yang berkembang pesat dalam satu dekade terakhir telah secara signifikan meningkatkan akurasi sistem ASR [9].

B. Model Whisper dan Faster-Whisper

Whisper merupakan model ASR yang dikembangkan oleh OpenAI dan dilatih menggunakan data audio multibahasa dalam skala besar, termasuk bahasa Indonesia, sehingga mampu melakukan transkripsi ucapan menjadi teks secara otomatis dengan performa yang baik pada berbagai bahasa serta tetap memberikan hasil yang andal pada beragam kondisi lingkungan [8]. Secara arsitektural, Whisper dibangun menggunakan struktur encoder-decoder berbasis Transformer, di mana bagian encoder mengubah input spektrogram log-Mel menjadi representasi fitur kontekstual, sedangkan bagian decoder menghasilkan token teks secara bertahap dengan memanfaatkan mekanisme attention terhadap representasi dari encoder maupun token yang telah dihasilkan sebelumnya [8]. Faster-Whisper merupakan varian yang mengoptimalkan proses inferensi model Whisper melalui teknik kuantisasi, sehingga kebutuhan memori dapat ditekan tanpa banyak mengorbankan akurasi transkripsi. Selain modul transkripsi, Faster-Whisper juga menyediakan filter Voice Activity Detection (VAD) berbasis deep learning bawaan (Silero VAD) yang dapat dimanfaatkan pada tahap prapemrosesan untuk memisahkan segmen suara dan non-suara sebelum proses transkripsi dilakukan, sehingga potensi halusinasi teks pada segmen hening dapat ditekan.

C. Word Error Rate (WER)

Performa sistem ASR umumnya diukur menggunakan metrik akurasi dan Word Error Rate (WER), yaitu persentase kata yang salah dikenali terhadap total kata yang sebenarnya, dengan komponen kesalahan berupa Substitution (S), Deletion (D), dan Insertion (I) terhadap total jumlah kata referensi (N) [9]. Semakin rendah nilai WER, semakin baik performa sistem ASR tersebut [10]. Sebagai komplemen dari WER, dihitung pula nilai Word Accuracy yang menggambarkan tingkat ketepatan transkripsi secara keseluruhan.

III. METODE

Penelitian ini merupakan bagian dari evaluasi modul Speech-to-Text (STT) pada sistem Student Voice Detection, yaitu sistem deteksi suara mahasiswa berbasis pengolahan audio dan machine learning yang dijalankan secara lokal untuk mendukung analisis keterlibatan dan evaluasi pembelajaran di kelas. Model STT yang digunakan adalah Faster-Whisper dengan konfigurasi yang dijalankan secara lokal pada perangkat tanpa memerlukan layanan cloud berbayar.

A. Pengumpulan Data Uji

Sebelum tahap pengujian modul dilakukan, terlebih dahulu dilakukan analisis perbandingan terhadap dua metode pengambilan data suara mahasiswa beserta perangkat keras yang digunakan, guna menentukan pendekatan pengumpulan data yang paling optimal. Metode pertama adalah pengambilan data secara daring melalui Google Form, di mana mahasiswa merekam suaranya secara mandiri menggunakan mikrofon bawaan perangkat masing-masing (handphone atau earphone). Metode ini relatif cepat dan mudah didistribusikan secara masif, namun karena tidak adanya pemantauan langsung dari peneliti, banyak sampel suara yang direkam dengan tempo tidak teratur dan artikulasi kurang jelas, sedangkan perbedaan spesifikasi perangkat antar responden menghasilkan kualitas audio yang sangat bervariasi. Metode kedua adalah pengambilan data secara langsung dan terpusat (supervisi), di mana seluruh responden direkam menggunakan satu perangkat mikrofon omnidirectional yang sama di bawah pengawasan langsung peneliti, sehingga konsistensi jarak mikrofon, kejelasan artikulasi, serta kondisi akustik ruangan dapat dikendalikan selama sesi perekaman. Meskipun membutuhkan waktu, jadwal, dan tenaga yang lebih besar, metode ini terbukti menghasilkan karakteristik akustik dasar (baseline environment) yang jauh lebih konsisten, sehingga matriks fitur akustik yang diekstraksi menjadi lebih bersih dan turut memengaruhi tingginya akurasi speech-to-text yang diperoleh pada tahap pengujian akhir.

B. Konfigurasi Model

Rancangan awal sistem sempat mempertimbangkan penggunaan model Whisper varian Medium. Namun, berdasarkan observasi awal pada tahap pengembangan, model tersebut membutuhkan beban komputasi yang cukup besar pada perangkat keras lokal yang digunakan, sehingga

kurang ideal untuk skenario pemrosesan wicara yang membutuhkan respons cepat. Oleh karena itu, pada implementasi akhir, sistem menggunakan arsitektur Faster-Whisper dengan varian model Small yang dijalankan menggunakan tipe komputasi int8 pada CPU, agar modul STT dapat beroperasi dengan lebih ringan dan stabil mendekati waktu nyata (real-time) tanpa membebani sumber daya komputasi secara berlebih. Sistem STT mengimplementasikan model Faster-Whisper Small dengan parameter bahasa Indonesia (language="id") dan prompt kontekstual yang berisi kosakata akademik teknik telekomunikasi (seperti frekuensi, sinyal, modulasi) untuk membantu model mengenali istilah-istilah teknis yang umum diucapkan di kelas. Penggunaan prompt ini mengarahkan model untuk lebih memprioritaskan probabilitas kata-kata teknis dibandingkan kata umum yang memiliki kemiripan bunyi. Sebelum proses transkripsi dilakukan, sistem juga memanfaatkan filter *Voice Activity Detection* (VAD) bawaan Whisper yang dikombinasikan dengan tahap pra-pemrosesan sinyal untuk memisahkan segmen suara dan non-suara, sehingga mengurangi risiko halusinasi teks pada segmen sunyi maupun gangguan suara latar.

C. Lingkungan dan Skenario Pengujian

Pengujian modul STT dilaksanakan pada mata kuliah Robotic Process Automation (RPA) Program S1 Teknik Telekomunikasi Universitas Telkom, dengan melibatkan 15–20 mahasiswa sebagai responden yang mengucapkan kalimat secara bergantian menggunakan mikrofon omnidirectional milik tim peneliti, sehingga kondisi akustik dapat terpantau secara konsisten. Kondisi kebisingan lingkungan kelas selama pengujian berada pada rentang 55–60 dB, yang diukur menggunakan aplikasi decibel pada smartphone. Setiap mahasiswa mengucapkan dua kalimat berbeda dan masing-masing kalimat diucapkan sebanyak dua kali, baik pada kondisi kelas yang relatif tenang maupun kondisi kelas dengan gangguan suara latar (noise), sehingga diperoleh total 54 sampel ucapan untuk dievaluasi. Hasil transkripsi yang dihasilkan sistem kemudian dibandingkan dengan kalimat referensi yang telah ditentukan sebelumnya untuk mengetahui tingkat ketepatan pengenalan kata. Metrik utama yang digunakan adalah Word Error Rate (WER) yang mengukur tingkat kesalahan pada level kata melalui tiga operasi edit, yaitu Substitution (S), Deletion (D), dan Insertion (I), terhadap total kata referensi (N). Selain WER, dihitung pula nilai Word Accuracy sebagai komplemen dari WER. Batas maksimum nilai WER yang ditetapkan pada penelitian ini adalah ≤ 30%, yang menggambarkan bahwa sistem dapat melakukan transkripsi dengan tingkat kesalahan minimal dan tetap mempertahankan makna ujaran secara konsisten.

Persamaan yang digunakan untuk menghitung Word Error Rate dan Word Accuracy adalah sebagai berikut:

$$WER = \frac{S + I + D}{N} \times 100\% \quad (1)$$

$$Word\ Accuracy = 100\% - WER \quad (2)$$

dengan S adalah jumlah kata yang salah (substitution), D adalah jumlah kata yang hilang (deletion), I adalah jumlah kata yang ditambahkan (insertion), dan N adalah total jumlah kata referensi.

D. Pengujian Efisiensi Komputasi

Selain akurasi transkripsi, dilakukan pula pengujian latensi dan waktu komputasi untuk mengevaluasi efisiensi performa modul STT saat memproses aliran data secara langsung (real-time), menggunakan sampel data operasional dari berkas hasil rekaman simulasi kelas. Setiap durasi wicara mahasiswa (audio duration) dibandingkan secara langsung dengan total waktu pemrosesan sistem (computational time) menggunakan Python timer. Kriteria kelayakan operasional real-time dievaluasi menggunakan metrik Real-Time Factor (RTF), yaitu rasio antara waktu komputasi yang dibutuhkan sistem dengan durasi asli audio yang diproses, sebagaimana dirumuskan berikut:

$$RTF = \frac{Waktu\ Komputasi\ (C)}{Durasi\ Bicara\ (D)} \times 100\% \quad (3)$$

Latensi sistem diukur dari kolom waktu komputasi yang merepresentasikan waktu yang dibutuhkan sistem sejak audio diterima hingga hasil transkripsi selesai diproses, dengan batas spesifikasi latensi maksimum ditetapkan sebesar durasi audio ditambah 10 detik.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian Word Error Rate

Berdasarkan hasil pengujian, diperoleh total 54 sampel ucapan dengan jumlah kata referensi (N) sebanyak 759 kata. Dari hasil transkripsi yang dihasilkan sistem, terdapat total kesalahan yang terdiri dari Substitution (S), Insertion (I), dan Deletion (D) sebanyak 99 kesalahan. Rekapitulasi hasil pengujian disajikan pada Tabel 1.

TABEL 1
REKAPITULASI HASIL PENGUJIAN SPEECH-TO-TEXT

Jumlah Sampel	Total Kesalahan (S+I+D)	WER (%)	Word Accuracy (%)
54	99	13,04%	86,96%

$$WER = \frac{87 + 4 + 8}{759} \times 100\% = 13.04\%$$

$$Word\ Accuracy = 100\% - 13.04\% = 86{,}96\%$$

Dari hasil perhitungan tersebut, diperoleh nilai total WER sebesar 13,04%, sehingga Word Accuracy yang dicapai adalah 86,96%. Nilai ini menunjukkan bahwa modul Speech-to-Text berbasis Faster-Whisper memiliki tingkat akurasi transkripsi yang cukup tinggi dan berada jauh di bawah batas maksimum WER yang ditetapkan sebesar ≤ 30%.

B. Analisis Tingkat Keberhasilan per Sampel

Selain evaluasi akurasi pada tingkat kata, evaluasi pada tingkat kalimat (sampel) juga dilakukan untuk memberikan gambaran yang lebih jelas mengenai tingkat keberhasilan pengenalan sistem secara utuh. Dari total 54 sampel pengujian, sebanyak 13 sampel (sekitar 24,1%) berhasil ditranskripsikan secara sempurna tanpa kesalahan sama sekali (WER = 0%), sedangkan sisanya masih mengandung

kesalahan pada tingkat yang bervariasi, dengan nilai WER per sampel berkisar antara 6,67% hingga 40%.

Pengamatan terhadap pola kesalahan menunjukkan bahwa kesalahan transkripsi umumnya terjadi dalam bentuk substitusi kata yang memiliki kemiripan fonetik, misalnya kata “kebisingan” yang terbaca sebagai “kebesingan” atau “kebisangan”, maupun kata “menunjukkan” yang terbaca sebagai “menujukan”. Sebagian kesalahan lain muncul akibat tempo pengucapan yang terlalu cepat sehingga beberapa kata yang berdekatan tergabung atau terpotong, seperti pada kalimat yang memuat kata “implementasinya” dan “sistem”, di mana sistem sempat mentranskripsikannya sebagai rangkaian kata yang tidak utuh. Meskipun demikian, istilah teknis bidang telekomunikasi yang telah disematkan pada prompt kontekstual, seperti “sistem”, “sinyal”, dan “akurasi”, secara umum berhasil dikenali dengan baik pada sebagian besar sampel, yang mengindikasikan bahwa mekanisme prompt kontekstual cukup efektif membantu pengenalan kosakata teknis.

Kesalahan transkripsi yang tersisa terutama disebabkan oleh faktor gangguan noise pada kondisi kelas dengan suara latar, variasi pelafalan antar mahasiswa, serta keterbatasan model dalam mengenali kata-kata yang diucapkan dengan tempo cepat atau kurang jelas. Namun secara keseluruhan, sistem telah mampu bekerja dengan baik dan dapat digunakan untuk mendukung proses pengenalan suara mahasiswa dalam kondisi nyata.

C. Hasil Pengujian Efisiensi Komputasi

Pengujian efisiensi komputasi dilakukan dengan membandingkan durasi wicara mahasiswa (audio duration) dengan total waktu pemrosesan sistem (computational time) pada 30 sampel data operasional. Perhitungan Real-Time Factor (RTF) rata-rata menghasilkan nilai sebesar 0,496 atau dibulatkan menjadi 0,50, yang menunjukkan bahwa sistem hanya memerlukan waktu komputasi sekitar 0,50 detik untuk memproses setiap 1 detik durasi audio yang diterima. Seluruh nilai RTF pada 30 sampel pengujian berada secara konsisten di bawah ambang batas maksimal 1,0, dengan rentang terendah 0,39 hingga tertinggi 0,72, sehingga tidak ada satu pun proses inferensi yang melampaui durasi asli audio.

Dari sisi latensi, diperoleh nilai latensi minimum sebesar 2,39 detik, latensi maksimum sebesar 2,87 detik, dan latensi rata-rata sebesar 2,60 detik. Verifikasi pada sampel dengan latensi tertinggi (worst case, durasi audio 4,83 detik) menunjukkan batas spesifikasi latensi maksimum sebesar 14,83 detik (durasi audio ditambah 10 detik), sedangkan latensi aktual yang diperoleh hanya sebesar 2,87 detik, jauh berada di bawah batas tersebut dengan margin yang sangat besar. Hasil ini membuktikan bahwa modul Faster-Whisper mampu melakukan inferensi transkripsi secara efisien tanpa menimbulkan latensi yang mengganggu pengalaman pengguna selama sesi kelas berlangsung.

D. Pengaruh Kualitas Data dan Prapemrosesan

Analisis lebih lanjut menunjukkan bahwa metode pengumpulan data turut memengaruhi hasil pengujian STT. Penggunaan satu perangkat mikrofon omnidirectional yang seragam pada seluruh responden, sebagaimana diterapkan pada metode pengambilan data secara langsung dan terpusat, terbukti menghasilkan karakteristik akustik dasar yang jauh lebih konsisten dibandingkan perekaman mandiri menggunakan mikrofon bawaan smartphone yang heterogen. Konsistensi ini turut berkontribusi terhadap tingginya akurasi speech-to-text yang diperoleh pada tahap pengujian akhir. Selain itu, integrasi filter Voice Activity Detection (VAD) bawaan Whisper pada tahap prapemrosesan turut berperan penting dalam memurnikan sinyal sebelum ditranskripsikan, sehingga membantu menekan angka halusinasi pada teks hasil transkripsi maupun anomali pembacaan yang dapat muncul pada segmen hening atau bising.

E. Pembahasan

Pemilihan varian Faster-Whisper Small dengan komputasi int8 terbukti menjadi langkah optimasi yang tepat karena mampu memangkas latensi pemrosesan secara drastis tanpa membebani memori komputasi berlebih, dibandingkan varian Medium yang dipertimbangkan pada tahap awal pengembangan namun terlalu berat untuk skenario pemrosesan wicara secara real-time pada perangkat keras lokal yang digunakan. Secara keseluruhan, hasil pengujian menunjukkan bahwa model Faster-Whisper varian Small mampu memenuhi kebutuhan transkripsi ucapan mahasiswa berbahasa Indonesia pada lingkungan perkuliahan nyata, baik dari sisi akurasi (WER 13,04%, jauh di bawah ambang batas 30%) maupun efisiensi komputasi (RTF 0,50 dan latensi rata-rata 2,60 detik, jauh di bawah batas spesifikasi), sekaligus menghilangkan ketergantungan pada koneksi internet dan biaya operasional layanan cloud berbayar.

V. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan, model Faster-Whisper varian Small yang dijalankan secara lokal dengan komputasi int8 pada CPU terbukti mampu mentranskripsikan ucapan mahasiswa berbahasa Indonesia pada lingkungan perkuliahan dengan nilai Word Error Rate (WER) sebesar 13,04% atau Word Accuracy sebesar 86,96% dari total 759 kata referensi pada 54 sampel uji, jauh di bawah batas maksimum WER yang ditetapkan sebesar ≤ 30%. Sekitar 24,1% dari seluruh sampel berhasil ditranskripsikan secara sempurna tanpa kesalahan, sedangkan kesalahan yang tersisa pada sampel lain umumnya berupa substitusi kata akibat kemiripan fonetik, tempo pengucapan yang cepat, dan gangguan suara latar. Dari sisi efisiensi komputasi, sistem mencatatkan Real-Time Factor (RTF) rata-rata sebesar 0,50 dan latensi rata-rata sebesar 2,60 detik, yang membuktikan bahwa model dapat beroperasi mendekati waktu nyata (real-time) tanpa membebani sumber daya komputasi secara berlebih.

Dengan demikian, Faster-Whisper varian Small terbukti menjadi solusi Speech-to-Text yang layak digunakan untuk mendukung sistem evaluasi keterlibatan mahasiswa berbasis suara pada lingkungan kelas nyata, tanpa bergantung pada layanan Cloud API pihak ketiga. Penelitian selanjutnya disarankan untuk mengeksplorasi penerapan Bandpass Filter (BPF) pada tahap prapemrosesan audio guna mengisolasi rentang frekuensi fundamental suara manusia, serta memperluas variasi data pelatihan prompt kontekstual, sehingga kualitas transkripsi dapat lebih ditingkatkan

khususnya pada kondisi kelas dengan gangguan suara latar yang tinggi maupun tempo bicara yang cepat.

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, no. January, pp. 1–6, 2023, doi: 10.3389/feduc.2023.1306568.

[3] Pemerintah Pusat, “Undang-undang (UU) Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi,” Pemerintah Pus., no. 016999, p. 1, 2022, [Online]. Available: https://peraturan.bpk.go.id/Download/224884/UU Nomor 27 Tahun 2022.pdf

[4] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, “Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm,” Sensors, vol. 22, no. 21, 2022, doi: 10.3390/s22218122.

[5] Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its Applications: A Review,” IEEE Access, vol. 10, no. November, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[6] 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.

[7] S. Ali, S. Tanweer, S. Khalid, and N. Rao, “Mel Frequency Cepstral Coefficient: A Review,” 2021, doi: 10.4108/eai.27-2-2020.2303173.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Proc. Mach. Learn. Res., vol. 202, pp. 28492–28518, 2023.

[9] S. Jurnal and P. S. Portabel, “(S I N T E K),” vol. V, no. 2, pp. 144–149.

[10] M. M. Or-Rashid, A. K. Nondi, A. Al Sadnun, M. A. H. Wadud, T. M. A. U. H. Bhuiyan, and M. S. Hossain, “Speech Emotion Recognition from Audio Data Using LSTM Model,” Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 7, pp. 247–254, 2025, doi: 10.14569/IJACSA.2025.0160726.