Real-Time Speaker Identification Using ECAPA-TDNN and Cosine Similarity for Classroom Engagement Monitoring

Mutiara Cantika
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
mutiaracantika@student.telkomuniversity.ac.id

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

Bambang Setia Nugroho
Fakultas Teknik Elektro
Telkom University
Bandung, Indonesia
bambangsetianugroho@telkomuniversity.ac.id

***Abstrak* — Identifikasi pembicara merupakan komponen kritis dalam sistem pemantauan keterlibatan mahasiswa di kelas. Penelitian ini mengimplementasikan modul identifikasi pembicara berbasis ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network) yang diintegrasikan ke dalam sistem Student Voice Detection (SVD) secara real-time. Model ECAPA-TDNN dari SpeechBrain digunakan untuk mengekstraksi voice embedding berdimensi 192 dari sinyal audio mahasiswa, kemudian dibandingkan dengan profil suara terdaftar menggunakan cosine similarity dengan threshold 0,50. Proses enrollment dilakukan dengan merata-ratakan 10 vektor embedding per mahasiswa menjadi satu profil suara berformat .npy. Pengujian dilakukan terhadap 11 mahasiswa terdaftar dan 4 pembicara tidak dikenal dengan total 54 sampel uji. Hasil pengujian menunjukkan akurasi identifikasi sebesar 98,15% dengan False Acceptance Rate (FAR) 0% dan False Rejection Rate (FRR) 2,50%, melampaui target awal 80%. Sistem berhasil berjalan secara real-time dengan latensi rata-rata 2,60 detik untuk keseluruhan pipeline, menunjukkan kelayakan implementasi ECAPA-TDNN dalam lingkungan kelas nyata.**

***Kata kunci* — identifikasi pembicara, ECAPA-TDNN, cosine similarity, real-time, student voice detection.**

I. PENDAHULUAN

Perkembangan teknologi pengenalan suara berbasis kecerdasan buatan dalam beberapa tahun terakhir membuka peluang baru dalam berbagai bidang, termasuk pendidikan. Salah satu tantangan nyata yang dihadapi dalam proses pembelajaran tatap muka adalah keterbatasan dosen dalam memantau keterlibatan seluruh mahasiswa secara objektif, terutama di kelas besar. Evaluasi yang dilakukan secara manual bergantung pada persepsi subjektif pengajar dan tidak menghasilkan data partisipasi yang terukur dan dapat dianalisis secara berkelanjutan [1].

Kondisi ini menunjukkan bahwa proses evaluasi pembelajaran masih menghadapi keterbatasan dalam memperoleh data partisipasi mahasiswa secara objektif. Salah satu penyebabnya adalah belum tersedianya sistem yang mampu merekam, mengidentifikasi, dan menganalisis aktivitas mahasiswa secara otomatis dan real-time.

Teknologi identifikasi pembicara (speaker identification) berbasis deep learning berpotensi menjadi solusi atas permasalahan tersebut karena mampu mengenali identitas pembicara secara otomatis berdasarkan karakteristik akustik suara [2]. Salah satu arsitektur yang banyak digunakan untuk tugas ini adalah ECAPA-TDNN, pengembangan dari TDNN (Time Delay Neural Network) yang meningkatkan kemampuan model dalam menangkap ciri khas suara tiap orang melalui mekanisme channel attention, dan telah divalidasi pada aplikasi kondisi audio dunia nyata seperti forensic voice recognition [3]. Karakteristik inilah yang menjadi dasar pemilihan ECAPA-TDNN pada penelitian ini, dipadukan dengan pengukuran cosine similarity terhadap profil suara terdaftar untuk melakukan identifikasi pembicara secara real-time selama sesi perkuliahan berlangsung.

Namun demikian, penerapan speaker identification pada konteks ruang kelas nyata dengan audio non-terkontrol dan banyak pembicara dalam satu sesi masih relatif terbatas dibahas dalam literatur. Salah satu studi bahkan melaporkan tingkat kesalahan identifikasi sekitar 34% pada rekaman kelas yang bising, jauh lebih tinggi dibanding hasil pada dataset uji yang terkontrol [4], menegaskan bahwa sebagian besar sistem speaker identification yang ada belum banyak diuji pada skenario pembelajaran aktual.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengembangkan dan mengevaluasi sistem identifikasi pembicara real-time berbasis ECAPA-TDNN dan cosine similarity untuk memantau keterlibatan mahasiswa di kelas, dengan mencatat identitas pembicara serta frekuensi partisipasi setiap mahasiswa secara otomatis selama proses pembelajaran berlangsung. Hasil penelitian diharapkan dapat memberikan gambaran mengenai tingkat akurasi identifikasi pembicara pada kondisi kelas nyata, serta menjadi dasar pengembangan sistem pemantauan keterlibatan mahasiswa yang lebih objektif dan berkelanjutan.

II. KAJIAN TEORI

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

A. Identifikasi Pembicara (Speaker Identification)

Identifikasi pembicara adalah tugas pengenalan pola yang bertujuan menentukan identitas seseorang secara otomatis berdasarkan karakteristik akustik suaranya, tanpa bergantung

pada kata-kata yang diucapkan (text-independent). Berbeda dengan verifikasi pembicara yang hanya menjawab "apakah ini orang yang sama?", identifikasi pembicara menjawab "siapa yang berbicara ini?" dari sekumpulan pembicara terdaftar [2].

Pendekatan berbasis *deep learning* bekerja dengan mengekstraksi representasi vektor berdimensi tetap dari sinyal suara, yang disebut *speaker embedding*. Embedding ini berfungsi sebagai "sidik jari" akustik yang merangkum karakteristik suara unik setiap individu dalam bentuk vektor kompak. Sistem kemudian membandingkan embedding dari audio masukan dengan profil suara terdaftar untuk menentukan identitas pembicara. Meskipun sistem berbasis *deep learning* mencapai akurasi tinggi pada kondisi akustik terkontrol, penerapan di lingkungan nyata tetap menjadi tantangan karena variasi noise dan kondisi rekaman yang tidak terprediksi [2].

B. ECAPA-TDNN

ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network) merupakan arsitektur deep learning yang dikembangkan untuk meningkatkan performa speaker identification melalui pembentukan speaker embedding yang lebih representatif. Arsitektur ini merupakan pengembangan dari Time Delay Neural Network (TDNN) dengan penambahan mekanisme Squeeze-and-Excitation, Res2Net, dan channel-dependent attentive statistics pooling sehingga mampu menangkap karakteristik suara secara lebih efektif [3].

Penelitian ini menggunakan model ECAPA-TDNN yang telah dilatih sebelumnya (*pre-trained*) menggunakan dataset VoxCeleb2. Karena telah mempelajari karakteristik suara dari data pelatihan yang beragam, model ini digunakan sebagai dasar dalam proses ekstraksi speaker embedding untuk identifikasi pembicara. Kemampuannya juga telah divalidasi dalam penelitian forensik sehingga menunjukkan performa yang baik pada berbagai kondisi perekaman [3].

C. Cosine Similarity

Cosine Similarity merupakan metode yang digunakan untuk mengukur tingkat kemiripan antara dua vektor berdasarkan sudut yang terbentuk di antara keduanya. Metode ini menghitung nilai cosinus dari sudut antar vektor, di mana nilai yang semakin mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi, sedangkan nilai yang mendekati 0 menunjukkan kemiripan yang rendah. Pada sistem speaker recognition, Cosine Similarity digunakan untuk membandingkan speaker embedding yang dihasilkan oleh model ECAPA-TDNN sehingga dapat menentukan tingkat kesamaan antara suara masukan dan data suara yang tersimpan dalam basis data [4].

Nilai yang dihasilkan berkisar antara -1 hingga 1. Pada penelitian ini, nilai Cosine Similarity digunakan sebagai dasar untuk menentukan identitas pembicara. Apabila skor kemiripan antara embedding suara uji dengan salah satu profil pada basis data mencapai atau melebihi nilai ambang batas yang telah ditentukan, maka pembicara dinyatakan teridentifikasi sebagai pemilik profil tersebut.

D. Proses Enrollment (Pendaftaran Suara)

Agar sistem dapat mengenali identitas seseorang, terlebih dahulu dilakukan proses enrollment, yaitu pendaftaran suara setiap pembicara. Pada tahap ini, setiap rekaman suara diproses oleh model speaker encoder, seperti ECAPA-TDNN, untuk menghasilkan speaker embedding yang merepresentasikan karakteristik unik suara pembicara [5]. Pada kondisi yang menggunakan beberapa rekaman (multiple enrollment utterances), speaker embedding dari setiap rekaman dapat dirata-ratakan (mean operation) untuk membentuk satu vektor representatif sebagai profil pembicara [5]. Selanjutnya, pada proses identifikasi, speaker embedding dari suara masukan dibandingkan dengan profil pembicara yang telah terdaftar menggunakan cosine similarity untuk menentukan tingkat kemiripan antar suara [5].

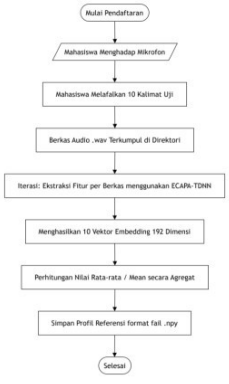
III. METODE

A. Pengumpulan Data Suara (Enrollment)

Data suara mahasiswa dikumpulkan melalui tahap pendaftaran (enrollment) menggunakan mikrofon dalam kondisi lingkungan yang tenang untuk menjaga kualitas audio. Setiap mahasiswa yang terdaftar diminta melafalkan sepuluh kalimat uji akademis dengan 10 kalimat yang berbeda-beda, sehingga diperoleh sepuluh berkas rekaman berformat .wav per mahasiswa sebagai representasi karakteristik suaranya.

B. Ekstraksi Fitur dan Pembentukan Profil Suara

Setiap berkas rekaman diproses menggunakan model pra-terlatih ECAPA-TDNN untuk menghasilkan speaker embedding, yaitu representasi numerik suara berdimensi 192. Karena setiap mahasiswa merekam sepuluh kalimat, maka dihasilkan sepuluh vektor embedding untuk satu orang. Kesepuluh vektor tersebut kemudian dirata-ratakan (mean) menjadi satu vektor tunggal yang mewakili profil suara mahasiswa tersebut. Cara ini dipilih agar profil suara tidak mudah berubah akibat variasi kecil pada suara sehari-hari, misalnya saat suara sedikit serak atau tempo bicara berubah. Vektor profil hasil rata-rata inilah yang kemudian disimpan sebagai berkas referensi berformat .npy.

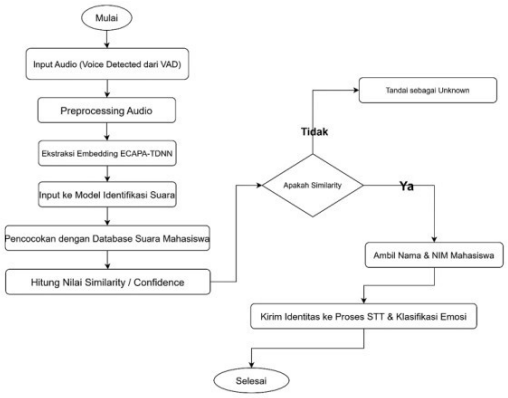


GAMBAR 1
Alur proses pendaftaran suara

Sebagaimana ditunjukkan pada Gambar 1, seluruh proses pendaftaran berjalan secara berurutan mulai dari perekaman kalimat uji hingga penyimpanan profil akhir. Vektor profil

hasil rata-rata tersebut disimpan sebagai berkas referensi berformat .npy, dan digunakan sebagai basis pencocokan pada tahap identifikasi pembicara yang dijelaskan pada bagian selanjutnya.

C. Proses Identifikasi Pembicara



GAMBAR 2
Alur proses identifikasi pembicara secara real-time

Pada tahap identifikasi, audio uji diekstraksi menjadi speaker embedding dengan model yang sama, kemudian dibandingkan dengan seluruh profil suara yang tersimpan menggunakan metrik cosine similarity, dihitung sebagai:

$$Cos(\theta) = \frac{A \,.\, B}{||A||\,||B||} \quad (1)$$

dengan A dan B masing-masing merepresentasikan vektor embedding audio uji dan vektor profil suara terdaftar. Profil dengan nilai kemiripan tertinggi dipilih sebagai kandidat identitas. Apabila nilai kemiripan tersebut memenuhi ambang batas (threshold) sebesar 0,50, sistem menetapkan identitas mahasiswa berupa nama dan Nomor Induk Mahasiswa (NIM); jika tidak memenuhi threshold, suara dikategorikan sebagai "Tidak dikenal".

D. Skema Pengujian

Pengujian dilakukan terhadap 54 sampel suara, mencakup suara mahasiswa yang telah terdaftar maupun suara pihak yang tidak terdaftar dalam sistem. Setiap hasil identifikasi diklasifikasikan ke dalam empat kategori:

a. True Acceptance (TA): suara terdaftar yang berhasil diterima dan diidentifikasi benar (similarity ≥ threshold).
b. False Rejection (FR): suara terdaftar yang keliru ditolak sistem (similarity < threshold).
c. True Rejection (TR): suara tidak terdaftar yang benar ditolak sistem.
d. False Acceptance (FA): suara tidak terdaftar yang keliru diterima sebagai pengguna terdaftar.

E. Metrik Evaluasi

Performa sistem identifikasi pembicara diukur menggunakan tiga metrik: akurasi keseluruhan, False Rejection Rate (FRR), dan False Acceptance Rate (FAR), dihitung sebagai berikut:

a. Akurasi Keseluruhan

$$\text{Accuracy} = \frac{TA+TR}{TA+TR+FA+FR} \times 100\% \quad (2)$$

b. False Rejection Rate (FRR)

$$\text{FRR} = \frac{FR}{TA+FR} \times 100\% \quad (3)$$

c. False Acceptance Rate (FAR)

$$\text{FAR} = \frac{FA}{FA+TR} \times 100\% \quad (4)$$

FRR mengukur seberapa sering sistem menolak suara mahasiswa yang sebenarnya sudah terdaftar, sedangkan FAR mengukur seberapa sering sistem keliru menerima suara pihak yang tidak terdaftar sebagai pengguna sah. Kedua metrik ini bersifat berlawanan terhadap nilai threshold cosine similarity yang digunakan: semakin ketat threshold, FAR cenderung mengecil namun FRR cenderung membesar, begitu pula sebaliknya.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Pengujian dilakukan terhadap 54 sampel suara, terdiri atas suara mahasiswa terdaftar maupun suara pihak yang tidak terdaftar dalam sistem. Hasil klasifikasi disajikan pada Tabel 1:

TABEL 1
Hasil Identifikasi Mahasiswa

Kategori	Jumlah	Istilah Teknis	Keterangan
Terdaftar dan Teridentifikasi Benar, nilai similarity dengan Threshold ≥ 50%	39	*True Acceptance* (TA)	Keputusan sistem Benar (Menerima pengguna sah)
Terdaftar tapi Tidak Teridentifikasi, nilai similarity dengan Threshold < 50%	1	*False Rejection* (FR)	Keputusan sistem Salah (Menolak pengguna sah)
Tidak Terdaftar & Tidak Teridentifikasi	14	*True Rejection* (TR)	Keputusan sistem Benar (Menolak pengguna tidak sah)
Tidak Terdaftar namun Teridentifikasi sebagai pengguna terdaftar	0	*False Acceptance* (FA)	Keputusan sistem Salah (Menerima pengguna tidak sah)

Berdasarkan data tersebut, diperoleh:

a. Akurasi Keseluruhan

$$\text{Accuracy} = \frac{39+14}{39+14+0+1} \times 100\%$$

$$\text{Accuracy} = \frac{39+14}{54} \times 100\%$$

$$\text{Accuracy} = 98.15\%$$

b. False Rejection Rate (FRR)

$$\text{FRR} = \frac{1}{39+1} \times 100\%$$

$$\text{FRR} = \frac{1}{40} \times 100\%$$

$$\text{FRR} = 2.50\%$$

c. False Acceptance Rate (FAR)

$$\text{FAR} = \frac{0}{14+0} \times 100\%$$

$$\text{FAR} = 0\%$$

Pada Tabel 1, terdapat 39 data True Acceptance (TA), 14 data True Rejection (TR), 1 data False Rejection (FR), dan tidak terdapat False Acceptance (FA). Hal ini menunjukkan

bahwa sistem mampu mengenali pengguna terdaftar dan menolak pengguna tidak terdaftar dengan baik.

B. Analisis Performa

Tingginya akurasi yang dicapai disebabkan oleh dua faktor utama. Pertama, model ECAPA-TDNN yang digunakan merupakan model pra-latih dari SpeechBrain yang dilatih pada dataset VoxCeleb1, berisi rekaman suara lebih dari 1.200 pembicara dari berbagai kondisi. Karena sudah dilatih pada data yang beragam, model ini mampu mengekstraksi ciri khas suara setiap mahasiswa dengan baik meskipun tanpa pelatihan ulang pada data kelas.

Kedua, profil suara setiap mahasiswa dibentuk dari rata-rata 10 vektor embedding hasil rekaman enrollment. Pendekatan rata-rata ini membuat profil lebih stabil karena variasi kecil yang terjadi antar rekaman seperti perbedaan volume atau intonasi dapat teredam. Hasilnya, profil yang tersimpan lebih merepresentasikan karakteristik suara asli setiap mahasiswa secara konsisten.

Nilai FA sebesar 0 menunjukkan bahwa sistem tidak pernah salah mengidentifikasi pembicara tidak terdaftar sebagai mahasiswa terdaftar. Hal ini penting agar data partisipasi yang tercatat di dashboard tetap akurat dan tidak terkontaminasi oleh suara dari luar sistem.

C. Analisis Kegagalan

Satu kasus kegagalan identifikasi terjadi pada salah satu sampel mahasiswa terdaftar yang menghasilkan nilai False Rejection. Kegagalan ini diduga disebabkan oleh durasi audio yang terlalu singkat setelah proses pemotongan bagian hening sehingga sinyal suara yang tersisa tidak cukup panjang bagi model untuk menghasilkan embedding yang stabil dan representatif. Akibatnya, nilai cosine similarity antara embedding uji dan profil terdaftar jatuh di bawah threshold 0,50 dan sistem mengkategorikan sampel tersebut sebagai pembicara tidak dikenal.

Temuan ini menunjukkan bahwa penetapan durasi audio minimum perlu dipertimbangkan sebagai syarat agar proses identifikasi berjalan optimal. Audio yang terlalu pendek tidak memberikan cukup informasi akustik bagi model untuk mengenali identitas pembicara secara akurat.

D. Perbandingan Target dan Realisasi

Tabel 2 menampilkan perbandingan antara target spesifikasi awal dengan hasil yang dicapai.

TABEL 2

Parameter	Target	Realisasi	Status
Akurasi identifikasi	≥ 80%	98,15%	Tercapai
False Acceptance Rate (FAR)	0%	0%	Tercapai
False Rejection Rate (FRR)	Seminimal mungkin	2,50%	Tercapai

Seluruh target spesifikasi berhasil dipenuhi. Akurasi 98,15% membuktikan bahwa kombinasi model ECAPA-TDNN dan cosine similarity layak digunakan sebagai modul identifikasi pembicara dalam sistem pemantauan keterlibatan mahasiswa di kelas secara real-time.

V. KESIMPULAN

Penelitian ini berhasil mengimplementasikan modul identifikasi pembicara real-time berbasis ECAPA-TDNN dengan cosine similarity ke dalam sistem Student Voice Detection. Pengujian terhadap 54 sampel menunjukkan akurasi 98,15%, FAR 0%, dan FRR 2,50%, melampaui target awal 80%. Arsitektur ECAPA-TDNN yang dilatih pada VoxCeleb2 terbukti mampu beradaptasi dengan baik pada domain kelas Indonesia tanpa fine-tuning, berkat kapasitas representasi embedding 192 dimensi yang kaya. Sistem berhasil berjalan secara real-time dengan RTF rata-rata 0,50, menunjukkan kelayakan implementasi di lingkungan kelas nyata.

Pengembangan ke depan dapat mencakup implementasi speaker diarization untuk menangani kondisi overlapping speech, penambahan bandpass filter sebagai preprocessing untuk meningkatkan kualitas sinyal di lingkungan bising, dan eksplorasi fine-tuning ECAPA-TDNN pada dataset suara mahasiswa Indonesia agar performa lebih optimal untuk domain target.

REFERENSI

[1] J. Wang, S. Dudy, X. He, Z. Wang, R. Southwell, and J. Whitehill, "Optimizing Speaker Diarization for the Classroom: Applications in Timing Student Speech and Distinguishing Teachers from Children," Journal of Educational Data Mining, vol. 17, no. 1, pp. 98–125, 2025.

[2] F. Sigona and M. Grimaldi, "Validation of an ECAPA-TDNN System for Forensic Automatic Speaker Recognition Under Case Work Conditions," Speech Communication, vol. 158, Art. no. 103045, 2024, doi: 10.1016/j.specom.2024.103045.

[3] N. F. Octavia and Berlilana, "Penerapan Algoritma K-Nearest Neighbor untuk Analisis Sentimen Ulasan Produk Elektronik pada Platform E-Commerce," Jurnal Algoritma, vol. 22, no. 2, pp. 2110–2121, Nov. 2025, doi: 10.33364/algoritma.v22i2.3083.

[4] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," IEEE Access, vol. 9, pp. 79236–79363, 2021, doi: 10.1109/ACCESS.2021.3084299.

[5] C. Zeng, X. Miao, X. Wang, E. Cooper, and J. Yamagishi, "Joint Speaker Encoder and Neural Back-End Model for Fully End-to-End Automatic Speaker Verification With Multiple Enrollment Utterances," Computer Speech & Language, vol. 86, Art. no. 101619, 2024, doi: 10.1016/j.csl.2024.101619.