JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 2, MEI, 2026, Pp. 252 – 264
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis Ketahanan *Lightweight* Audio Spectrogram Transformer pada Identifikasi Pembicara Kondisi Berderau

I Kadek Arya Sugianta [(1)*], Gde Palguna Reganata [(2)]
Departemen Informatika, Universitas Bali Internasional, Bali, Indonesia
e-mail : aryabisabikin@gmail.com, palgunareganata@iikmpbali.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 15 Maret 2026, direvisi 29 April 2026, diterima 29 April 2026, dan dipublikasikan 25 Mei 2026.

Abstract

The use of deep learning models for speaker identification on devices with limited computational resources requires significant architectural optimization. This study evaluates the performance and robustness of the Lightweight Audio Spectrogram Transformer (AST) architecture, which has been extremely compressed to 570,536 parameters. The proposed method uses low-resolution Mel-Spectrogram representations (64x64 pixels) as input for a global self-attention mechanism. Testing was conducted using a 5-Fold Cross Validation scheme on a dataset injected with non-stationary environmental noise from the ESC-50 corpus at various Signal-to-Noise Ratio (SNR) levels. Experimental results show that under ideal conditions, the model achieves a solid average validation accuracy of 70.86% ± 2.69% with a Macro Average F1-score of 0.68 ± 0.03. However, the model's performance degrades sharply to 17.61% at an SNR of 5 dB and drops to 9.21% under extreme conditions at an SNR of 0 dB. These findings reveal a critical trade-off where radical parameter compression leads to the loss of spectral feature redundancy that acts as an implicit noise filter. This study concludes that while lightweight Transformer mechanisms are highly efficient for Edge AI, the integration of pre-processing modules or noise-robust training strategies is an absolute necessity to maintain identification integrity in noisy real-world environments.

Keywords: Speaker Identification, Audio Spectrogram Transformer, Edge AI, Robustness Analysis, Deep Learning

Abstrak

Penggunaan model *deep learning* untuk identifikasi pembicara pada perangkat dengan keterbatasan sumber daya komputasi memerlukan optimasi arsitektur yang signifikan. Penelitian ini mengevaluasi kinerja dan ketahanan arsitektur *Lightweight Audio Spectrogram Transformer* (AST) yang telah dikompresi secara ekstrem hingga memiliki 570.536 parameter. Metode yang diusulkan menggunakan representasi *Mel-Spectrogram* beresolusi rendah (64x64 piksel) sebagai masukan untuk mekanisme *self-attention* global. Pengujian dilakukan menggunakan skema *5-Fold Cross Validation* pada dataset yang diinjeksi derau lingkungan non-stasioner dari korpus ESC-50 dengan berbagai tingkatan *Signal-to-Noise Ratio* (SNR). Hasil eksperimen menunjukkan bahwa pada kondisi ideal, model mampu mencapai rata-rata akurasi validasi yang solid sebesar 70,86% ± 2,69% dengan nilai *Macro Average F1-score* sebesar 0,68 ± 0,03. Namun, performa model mengalami degradasi tajam menjadi 17,61% pada SNR 5 dB dan menurun hingga 9,21% pada kondisi ekstrem SNR 0 dB. Temuan ini mengungkap adanya *trade-off* kritis di mana kompresi parameter yang radikal menyebabkan hilangnya redundansi fitur spektral yang berfungsi sebagai penyaring derau implisit. Penelitian ini menyimpulkan bahwa meskipun mekanisme Transformer ringan sangat efisien untuk *Edge AI*, integrasi modul pra-pemrosesan atau strategi *noise-robust training* merupakan keharusan mutlak untuk menjaga integritas identifikasi di lingkungan dunia nyata yang bising.

Kata Kunci: Identifikasi Pembicara, Transformator Spektrogram Audio, Edge AI, Analisis Robustness, Deep Learning

1. PENDAHULUAN

Dalam era digital saat ini, sistem identifikasi pembicara (*speaker identification*) telah menjadi komponen penting dalam berbagai aplikasi, mulai dari asisten virtual dan sistem keamanan biometrik berbasis autentikasi dua faktor (Kamiński et al., 2023; Mohd Hanifa et al., 2021), hingga teknologi interaksi manusia dan komputer. Kemajuan dalam teknologi deep learning memungkinkan sistem untuk mengenali identitas vokal manusia dengan tingkat akurasi yang tinggi pada kondisi ideal (Ye & Yang, 2021). Namun, salah satu tantangan utama yang belum sepenuhnya terpecahkan adalah mempertahankan kinerja identifikasi dalam lingkungan berisik (*noisy environment*). Hal ini disebabkan oleh keterbatasan representasi spektrogram tradisional dalam membedakan antara komponen derau, gema, dan sinyal suara target (X. Zhang et al., 2025). Lingkungan akustik yang kompleks sering kali menyebabkan degradasi kualitas sinyal, sehingga menyulitkan sistem untuk membedakan fitur pita suara manusia dari derau latar belakang (Alharbi et al., 2021).

Sejumlah penelitian telah berupaya meningkatkan ketahanan sistem melalui pendekatan multimodal, seperti mengintegrasikan informasi visual pergerakan bibir (Li et al., 2023) atau isyarat taktil (Oh et al., 2022). Meskipun terbukti meningkatkan kinerja pada kondisi *Signal-to-Noise Ratio* (SNR) yang rendah, pendekatan ini memiliki keterbatasan praktis. Metode multimodal tidak dapat diaplikasikan pada skenario interaksi tanpa kamera seperti panggilan telepon atau perangkat *Internet of Things* (IoT) berbasis suara murni, serta menuntut beban komputasi yang sangat tinggi akibat pemrosesan paralel yang kompleks (Jeon & Kim, 2022; T. Zhang et al., 2025) . Oleh karena itu, pengembangan sistem identifikasi pembicara berbasis audio murni (*unimodal*) tetap menjadi prioritas riset yang krusial

Secara konvensional, ekstraksi fitur spektral dikombinasikan dengan arsitektur seperti *Convolutional Neural Networks* (CNN) dan *Recurrent Neural Networks* (RNN) telah menjadi standar (Zaman et al., 2023). Namun, CNN memiliki keterbatasan inheren dalam menangkap ketergantungan jarak jauh (*long-range dependencies*) karena operasi konvolusinya hanya memproses titik sampel audio lokal secara sekuensial, sehingga pemodelan konteks global menuntut operasi rekursif yang tidak efisien secara komputasi (J. Liu & Huang, 2023). Selain itu, RNN juga rentan terhadap masalah gradien hilang pada sekuens audio yang panjang (Zeng & Lau, 2023). Untuk mengatasi hal ini, *Audio Spectrogram Transformer* (AST) muncul sebagai arsitektur *State-of-the-Art* (SOTA). Melalui mekanisme *self-attention*, AST mampu memetakan hubungan spasial dan temporal secara langsung dari representasi 2D Mel-Spectrogram untuk menangkap konteks global audio secara lebih efisien (F. Liu & Fang, 2023).

Meskipun AST standar menawarkan akurasi superior, implementasinya menghadapi kendala fatal berupa beban komputasi kuadratik yang sangat masif (*resource-intensive*). Transformer standar umumnya membutuhkan memori GPU yang besar dan jumlah parameter mencapai puluhan juta, menjadikannya sulit untuk diterapkan pada perangkat dengan sumber daya terbatas (Adnan et al., 2022). Upaya untuk mengefisiensikan model ini pada sistem *low-power* biasanya memerlukan teknik optimasi tambahan seperti kompilasi model dengan TensorRT atau kuantisasi data guna menjaga keseimbangan antara akurasi dan penggunaan memori (Martin-Salinas et al., 2024). Selain itu, masalah *overparameterization* pada model *pre-trained* sering kali mengakibatkan inefisiensi memori yang signifikan, sehingga diperlukan strategi pemangkasan (*pruning*) yang tepat untuk menyesuaikan model dengan batasan sumber daya yang ada (Wang et al., 2025).

Beberapa studi terbaru telah berupaya mengeksplorasi potensi AST untuk efisiensi klasifikasi audio, namun batasan komputasi tetap menjadi kendala utama. Sebagai contoh, penelitian oleh Nugroho et al. (2025) melaporkan bahwa meskipun AST efektif untuk deteksi biomarker vokal, model tersebut membutuhkan durasi pelatihan yang signifikan mencapai 78 menit per *fold*. Di sisi lain, upaya untuk meningkatkan efisiensi sering kali beralih ke arsitektur RNN dengan mekanisme atensi Jandera et al. (2025), yang meskipun ringan, namun memiliki keterbatasan dalam paralelisasi komputasi dan pemodelan konteks global dibandingkan arsitektur Transformer murni.

Hingga saat ini, belum ada literatur yang secara kritis mengevaluasi sejauh mana arsitektur AST yang dikompresi secara ekstrem hingga di bawah 600.000 parameter dengan resolusi input minimal (64x64 piksel) dapat mempertahankan integritas fitur akustik pembicara.

Untuk menjawab tantangan tersebut secara lebih radikal, penelitian ini mengusulkan arsitektur *Lightweight Audio Spectrogram Transformer* (AST Ringan). Melalui strategi pemangkasan ekstrem (*feature extraction pruning*), dimensi audio dipotong menjadi durasi 2 detik dan resolusi Mel-Spectrogram direduksi menjadi matriks 64x64. Pendekatan ini secara drastis menekan kompleksitas model menjadi kurang dari 600.000 parameter, memungkinkan pelatihan dan inferensi penuh pada lingkungan komputasi yang sangat terbatas (VRAM 2 GB).

Namun, kompresi arsitektur yang ekstrem ini memunculkan satu pertanyaan riset yang belum terjawab: Sejauh mana model AST yang sangat ringan ini dapat mempertahankan ketahanannya (*robustness*) terhadap derau non-stasioner? Reduksi dimensi spektral berisiko menghilangkan fitur akustik redundan yang biasanya membantu model besar untuk bertahan dari *noise*.

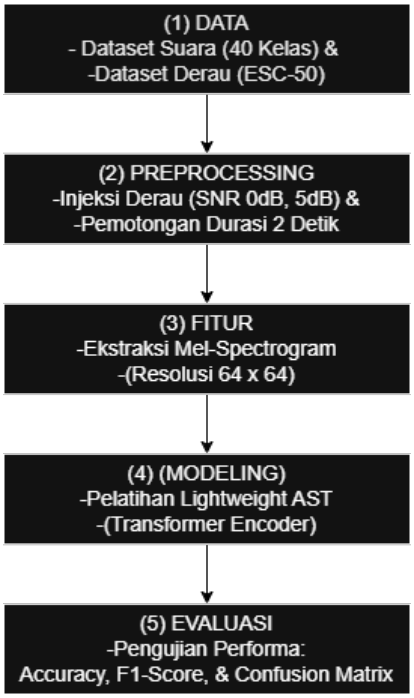
Untuk menguji batasan tersebut, penelitian ini mengevaluasi model *Lightweight AST* secara komprehensif dengan mengklasifikasikan 40 identitas pembicara. Skenario pengujian menyimulasikan lingkungan nyata dengan menginjeksi derau lingkungan ekstrem dari dataset ESC-50 ke dalam dataset suara bersih. Pengujian dilakukan pada berbagai tingkat gangguan, mulai dari kondisi ideal (*Clean*), bising menengah (SNR 5 dB), hingga sangat bising di mana energi derau setara dengan suara target (SNR 0 dB).

Perbedaan utama penelitian ini dengan studi sebelumnya terletak pada fokus evaluasi empiris terhadap "kurva degradasi" dari sebuah arsitektur Transformer unimodal berskala sangat kecil. Penelitian ini tidak bertujuan untuk melampaui akurasi model raksasa, melainkan untuk memetakan titik kritis (*breaking point*) dari AST berukuran mikro saat dihadapkan pada kebisingan ekstrem. Hasil penelitian ini diharapkan dapat memberikan landasan baru terkait *trade-off* antara efisiensi komputasi dan ketahanan derau, serta mendemonstrasikan batasan kemampuan mekanisme *self-attention* murni pada matriks spektral beresolusi rendah.

2. METODE PENELITIAN

Penelitian ini menerapkan pendekatan eksperimental untuk menguji ketahanan (*robustness*) model *Lightweight Audio Spectrogram Transformer* (AST) terhadap berbagai tingkatan gangguan derau lingkungan. Skenario pengujian dirancang dengan mensimulasikan berbagai level *Signal-to-Noise Ratio* (SNR) untuk memetakan titik kritis kegagalan model dalam kondisi riil, selaras dengan metodologi evaluasi ketangguhan pada lingkungan urban yang kompleks (Chen et al., 2024). Tahapan penelitian dirancang secara sistematis, mengintegrasikan proses pengolahan sinyal digital dan arsitektur *deep learning* mulai dari fase persiapan data, ekstraksi fitur spektral, hingga evaluasi kinerja model secara komprehensif. Seluruh alur kerja yang diusulkan dalam penelitian ini diilustrasikan secara detail pada blok diagram yang disajikan dalam Gambar 1.

Gambar 1 merepresentasikan alur kerja sistematis yang dimulai dari tahap akuisisi data hingga evaluasi performa model. Proses ini diawali dengan pengintegrasian *dataset clean speech* dan derau lingkungan dari ESC-50, yang kemudian melalui tahap pra-pemrosesan berupa injeksi derau pada berbagai level SNR serta penyeragaman durasi audio menjadi dua detik. Tahap krusial terletak pada ekstraksi fitur *Mel-Spectrogram* dengan resolusi yang direduksi menjadi 64x64 piksel untuk menyeimbangkan kebutuhan informasi spektral dan efisiensi memori. Fitur tersebut kemudian diumpankan ke dalam arsitektur Lightweight Audio Spectrogram Transformer (AST) yang memiliki total 570.536 parameter. Mekanisme *self-attention* di dalam model bertugas mengekstrak ketergantungan fitur global, yang pada akhirnya dievaluasi melalui metrik akurasi dan matriks kebingungan untuk memetakan tingkat ketangguhan model terhadap degradasi sinyal.



Gambar 1 Blok Diagram Metodologi Penelitian Identifikasi Pembicara

2.1 Spesifikasi Dataset dan Partisi Data

Dataset utama yang digunakan dalam penelitian ini bersumber dari *corpus* LibriSpeech, yang mencakup 40 kelas identitas pembicara yang berbeda (terdiri dari 20 pria dan 20 wanita untuk menjamin keberagaman fitur vokal). Total sampel audio yang diolah secara keseluruhan adalah sebanyak 1.953 file audio dengan durasi tetap masing-masing selama 2 detik. Distribusi data per kelas terjaga secara seimbang dengan rata-rata 48 hingga 50 sampel per pembicara.

Guna menjamin transparansi eksperimen dan stabilitas hasil evaluasi, penelitian ini menerapkan skema *5-Fold Cross Validation*. Seluruh dataset dipartisi dengan rasio 80:20 pada setiap *fold*, di mana 1.562 sampel dialokasikan untuk fase pelatihan model, sementara 391 sampel digunakan sebagai data validasi untuk menguji kemampuan generalisasi model. Protokol ini memastikan bahwa setiap sampel audio pernah merepresentasikan data validasi, sehingga meminimalkan risiko bias pemilihan data (*selection bias*) serta menjamin hasil evaluasi yang lebih objektif.

2.2 Landasan Teoretis Audio Spectrogram Transformer (AST)

Penelitian ini mengadopsi arsitektur *Audio Spectrogram Transformer* (AST) sebagai unit pemrosesan utama. Berbeda dengan pendekatan CNN yang memiliki *spatial inductive bias* terbatas pada ekstraksi fitur lokal, AST memanfaatkan mekanisme *self-attention* global (Gong et al., 2021). Hal ini memungkinkan model untuk memodelkan korelasi antara *patch* frekuensi rendah dan tinggi secara simultan melalui analisis domain frekuensi yang adaptif, yang krusial untuk mengenali karakteristik vokal unik yang tersebar pada spektrum frekuensi (Huang et al., 2026).

2.3 Ekstraksi Fitur *Mel-Spectrogram*

Representasi input yang digunakan adalah Mel-Spectrogram, yang berfungsi mengubah sinyal audio domain waktu menjadi representasi domain waktu-frekuensi 2D. Penggunaan fitur ini terbukti memberikan representasi data suara yang lebih informatif, sehingga meningkatkan akurasi klasifikasi secara signifikan bahkan dalam lingkungan dengan gangguan derau yang kompleks (Mannem et al., 2024). Skala Mel diterapkan untuk menyesuaikan representasi

frekuensi agar selaras dengan persepsi pendengaran manusia. Proses ini sangat krusial karena memungkinkan model untuk mengenali karakteristik unik dari pita suara pembicara (pola *pitch* dan *formant*) melalui distribusi energi spektral pada matriks visual. Ekstraksi menggunakan jendela *Hamming* dengan *overlap* sebesar 50% untuk meminimalkan *spectral leakage* pada transisi antar *frame*.

2.4 Skenario Injeksi Derau dan Kuantifikasi SNR

Untuk mensimulasikan tantangan di lingkungan nyata, dilakukan proses injeksi derau non-stasioner yang bersumber dari ESC-50 ke dalam dataset suara bersih 40 pembicara. Tingkat gangguan pada sinyal diukur secara matematis menggunakan parameter Signal-to-Noise Ratio (SNR) melalui Pers. (1). Nilai SNR digunakan untuk mengatur tingkat intensitas derau terhadap sinyal suara target selama proses pengujian.

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right) \quad (1)$$

Di mana P_{signal} merupakan daya rata-rata sinyal suara pembicara dan P_{noise} adalah daya rata-rata derau lingkungan. Eksperimen dibagi menjadi tiga skenario utama, yaitu kondisi ideal (*Clean*), kondisi bising menengah (SNR 5 dB), dan kondisi ekstrem (SNR 0 dB) di mana energi kebisingan setara dengan energi suara target. Rincian tujuan dari setiap skenario dijelaskan sebagai berikut:

a) Kondisi Ideal (*Clean*): Pengujian tanpa injeksi derau (∞dB) yang berfungsi sebagai standar acuan akurasi maksimal model.
b) Kondisi Bising Menengah (SNR 5 dB): Mensimulasikan lingkungan dengan gangguan suara latar yang signifikan, di mana energi sinyal target masih lebih besar dari derau $(S > N)$
c) Kondisi Ekstrem (SNR 0 dB): Kondisi di mana energi kebisingan ekuivalen dengan energi suara target $(S = N)$ yang merepresentasikan ambang batas kritis bagi kemampuan ekstraksi fitur pada arsitektur AST Ringan.

2.5 Pra-pemrosesan Data

Tahap persiapan data dilakukan untuk mentransformasikan audio mentah menjadi format yang siap diolah oleh model dengan tetap mempertimbangkan efisiensi sumber daya secara ketat. Proses ini diawali dengan tahap *slicing*, di mana seluruh sampel audio dipotong secara konsisten menjadi durasi tetap selama 2 detik untuk menyeragamkan dimensi masukan temporal. Selanjutnya, diterapkan strategi *feature pruning* dengan mereduksi resolusi *Mel-Spectrogram* menjadi dimensi 64x64 piksel. Langkah reduksi ekstrem menjadi 64x64 piksel bertujuan untuk mereduksi kompleksitas komputasi mekanisme *attention* yang bersifat $O(N^2)$, di mana N adalah jumlah *patch* masukan.

2.6 Perancangan Arsitektur Lightweight AST

Model yang diusulkan merupakan varian AST yang dikompresi secara manual untuk mencapai efisiensi ekstrem dengan total 570.536 parameter. Arsitektur ini dirancang untuk mempertahankan kemampuan diskriminatif fitur vokal dalam skenario sumber daya terbatas (Edge AI). Alur implementasi dimulai dengan transformasi sinyal audio (16 kHz) menjadi representasi Mel-Spectrogram berdimensi 64x64 piksel.

Struktur arsitektur ini dimulai dengan modul *patch embedding* yang bertugas mengonversi matriks spektrogram tersebut menjadi deretan token linear. Berbeda dengan arsitektur AST standar yang sering kali menggunakan ukuran *patch* besar dengan *overlap*, model ini membagi matriks 64x64 menjadi unit *patch* kecil berukuran 16x16 piksel tanpa *overlap*, yang menghasilkan 16 token masukan. Strategi ini dipilih secara spesifik guna mereduksi beban komputasi pada pembentukan *attention map* yang bersifat kuadratik terhadap jumlah token. Setiap *patch* kemudian diproyeksikan secara linear menjadi *embedding* dengan dimensi sematan (*embedding dimension*) sebesar 128.

Sebelum memasuki tahap pemrosesan utama, setiap *embedding* ditambahkan dengan *positional encoding* untuk mempertahankan informasi urutan struktur frekuensi-waktu dari spektrum audio asli. Inti dari pemrosesan data terletak pada penggunaan 4 lapisan *Transformer encoder* yang saling terhubung. Di dalam setiap lapisan tersebut, mekanisme Multi-Head Self-Attention (MHSA) diimplementasikan dengan konfigurasi 4 heads untuk memetakan ketergantungan fitur spektral secara global. Penggunaan *multi-head* memungkinkan model untuk memfokuskan atensi pada berbagai sub-ruang fitur vokal secara simultan, sehingga model tetap mampu menangkap ciri unik pembicara meskipun terdapat gangguan derau pada input. Rincian distribusi parameter untuk setiap komponen model disajikan pada Tabel 1.

Tabel 1 Spesifikasi Arsitektur Lightweight AST

Layer / Komponen	Spesifikasi / Konfigurasi	Output Shape	Jumlah Parameter
Input Stage	Mel-Spectrogram	(1, 64, 64)	0
Patch Embedding	Patch: 16x16, Stride: 16	(16, 128)	33.024
Positional Encoding	Learnable Absolute	(16, 128)	2.048
Transformer Block 1	MHSA (4 Heads, Dim 128)	(16, 128)	132.480
Transformer Block 2	MHSA (4 Heads, Dim 128)	(16, 128)	132.480
Transformer Block 3	MHSA (4 Heads, Dim 128)	(16, 128)	132.480
Transformer Block 4	MHSA (4 Heads, Dim 128)	(16, 128)	132.480
Global Pooling	Average Pooling	(128)	0
Classifier Head	Linear Layer (40 Kelas)	(40)	5.160
Total Parameter			**570.536**

2.7 Protokol Evaluasi Kinerja

Kinerja model diukur menggunakan skema *5-Fold Cross Validation* untuk menjamin stabilitas hasil pada seluruh partisi data. Evaluasi kuantitatif dilakukan menggunakan matriks kebingungan (*confusion matrix*) untuk menghitung nilai *True Positive* (TP), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan nilai tersebut, ditarik kesimpulan melalui metrik akurasi, presisi, *recall*, dan *F1-score* sebagaimana ditunjukkan pada Pers. (2), (3), dan (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Selain metrik evaluasi standar, penelitian ini juga menetapkan ambang batas performa minimum berdasarkan probabilitas tebakan acak (*random guessing probability*). Mengingat tugas klasifikasi melibatkan N kelas yang seimbang, maka kemampuan diskriminatif model dikatakan hilang apabila akurasi yang dihasilkan mendekati atau lebih rendah dari nilai P_{rand} yang didefinisikan pada Pers. (5).

$$P_{rand} = \frac{1}{N_{class}} \times 100\% \tag{5}$$

Dengan jumlah target klasifikasi sebanyak 40 identitas pembicara ($Nclass = 40$), maka nilai ambang batas acak dalam penelitian ini adalah 2,5%. Parameter ini digunakan sebagai indikator kritis untuk mengevaluasi titik kegagalan fungsional model (*functional failure point*) saat diuji pada kondisi derau ekstrem. Penggunaan metrik ini sangat krusial untuk membuktikan bahwa meskipun model mengalami degradasi tajam pada kondisi SNR 0 dB, arsitektur yang diusulkan tetap mempertahankan integritas strukturalnya di atas ambang batas peluang acak.

2.8 Konfigurasi Pelatihan dan Hyperparameter

Seluruh proses eksperimen dijalankan menggunakan *framework* PyTorch. Untuk memastikan reproduktifitas penelitian, konfigurasi *hyperparameter* yang digunakan selama fase pelatihan disajikan dalam Tabel 2. Model dilatih menggunakan *optimizer* Adam dengan konstanta pembelajaran (*learning rate*) sebesar 0,0001 (1e-4) dan fungsi kerugian *Cross-Entropy*. Proses pelatihan dilakukan sebanyak 20 *epoch* dengan ukuran *batch* 32. Seluruh bobot model diinisialisasi ulang pada setiap *fold* dalam skema *5-fold cross-validation* untuk mencegah terjadinya kebocoran data (*data leakage*) dan menjamin objektivitas hasil evaluasi.

Tabel 2 Detail Konfigurasi Hyperparameter Pelatihan

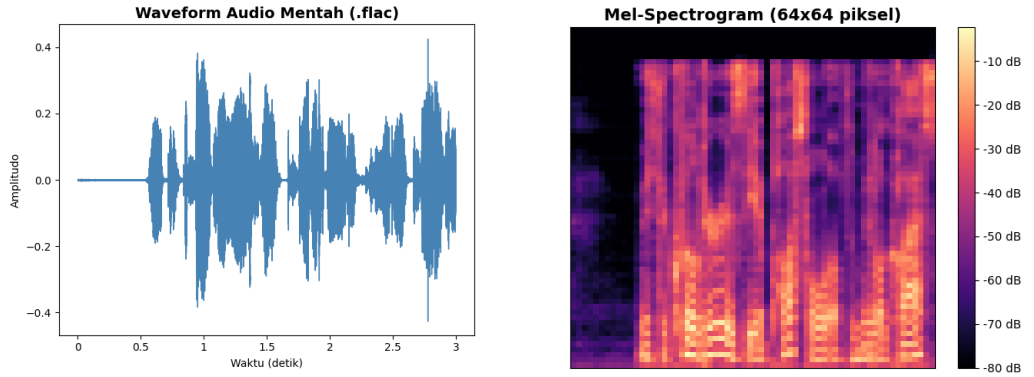
Parameter	Nilai / Deskripsi
Optimizer	Adam
Learning Rate	1e-4 (0,0001)
Loss Function	*Cross Entropy Loss*
Batch Size	32
Total Epochs	20
Validation Method	*5-Fold Cross Validation*
Weight Initialization	*Re-initialized per Fold (Xavier Uniform)*
Hardware Accelerator	GPU (CUDA Enabled)

3. HASIL DAN PEMBAHASAN

Penelitian ini mengevaluasi kinerja arsitektur *Lightweight Audio Spectrogram Transformer* (AST) dalam mengklasifikasikan 40 identitas pembicara. Evaluasi dilakukan secara bertahap, dimulai dari pemantauan stabilitas pelatihan melalui kurva *loss*, pengujian *baseline* pada kondisi tanpa derau (*clean*), hingga pengujian ketahanan (*robustness*) pada skenario bising menengah (SNR 5 dB) dan sangat bising (SNR 0 dB). Seluruh proses pelatihan dilakukan selama 20 *epoch* menggunakan skema *5-Fold Cross Validation* untuk menjamin validitas dan generalisasi hasil eksperimen.

3.1 Persiapan Dataset dan Pra-pemrosesan

Dataset utama yang digunakan dalam penelitian ini bersumber dari korpus audio *LibriSpeech* dengan format FLAC (*Free Lossless Audio Codec*) yang mencakup 40 identitas pembicara berbeda. Sebelum memasuki tahap pelatihan, sinyal audio mentah yang bersifat kontinu harus ditransformasikan ke dalam representasi frekuensi-waktu agar dapat diproses oleh arsitektur *Transformer*. Sebagaimana diperlihatkan pada Gambar 2, proses pra-pemrosesan dimulai dengan mengekstraksi fitur akustik dari *waveform* audio mentah menjadi bentuk *Mel-Spectrogram*.



(a) Waveform Audio Mentah dari LibriSpeech

(b) Representasi Mel-Spectrogram 64×64 sebagai input model

Gambar 2 Visualisasi Tahap Pra-Pemrosesan Audio

Untuk memenuhi target efisiensi perangkat *Edge AI*, dilakukan kompresi dimensi secara ekstrem menjadi resolusi 64x64 piksel. Meskipun hasil transformasi pada Gambar 2 (b) menunjukkan representasi visual yang lebih rendah resolusinya dibandingkan standar industri, pola-pola energi spektral yang merepresentasikan karakteristik vokal unik pembicara (seperti *pitch* dan *timbre*) tetap terjaga. Penggunaan dimensi minimal ini secara signifikan mereduksi beban komputasi pada mekanisme *self-attention* global, yang menjadi kunci utama arsitektur *Lightweight AST* dalam penelitian ini.

3.1 Skenario Injeksi Derau Lingkungan

Untuk menguji batas ketahanan (*robustness*) arsitektur *Lightweight* AST, penelitian ini merancang skenario pengujian menggunakan derau lingkungan yang non-stasioner. Mengacu pada metode pengujian kondisi nyata yang dilakukan oleh Langi & Fadllullah (2026), model dievaluasi tidak hanya pada kondisi ideal, tetapi juga pada kondisi interferensi akustik yang mensimulasikan tantangan dalam lingkungan nyata. Derau diambil dari dataset ESC-50, yang mencakup berbagai kategori suara lingkungan. Proses pencampuran antara sinyal suara pembicara dan derau dikontrol melalui parameter *Signal-to-Noise Ratio* (SNR) yang telah didefinisikan pada Persamaan 1 pada bab sebelumnya. Seluruh parameter skenario pengujian dirangkum dalam matriks pada Tabel 3.

Tabel 3 Matriks Skenario Pengujian Ketahanan Model

Skenario	Nilai SNR	Kondisi Akustik	Deskripsi Tantangan	Target Evaluasi
I	∞ dB	Ideal (Clean)	Tanpa interferensi derau luar	Baseline akurasi maksimal
II	5 dB	Moderate *Noise*	Sinyal suara > Derau ($S > N$)	Stabilitas fitur spektral
III	0 dB	Extreme *Noise*	Sinyal suara = Derau ($S = N$)	Batas kegagalan fungsional

Pemilihan nilai SNR 5 dB merepresentasikan kondisi di mana sinyal suara manusia masih dominan namun sudah mengalami distorsi frekuensi yang signifikan. Sementara itu, skenario SNR 0 dB merupakan ambang batas kritis di mana tingkat energi derau ekuivalen dengan energi sinyal suara target. Matriks pengujian ini bertujuan untuk membuktikan hipotesis bahwa kompresi parameter model yang ekstrem pada arsitektur *lightweight* akan berdampak pada hilangnya redundansi informasi yang biasanya berfungsi sebagai pelindung terhadap korupsi fitur (*feature corruption*) akibat interferensi derau yang kuat.

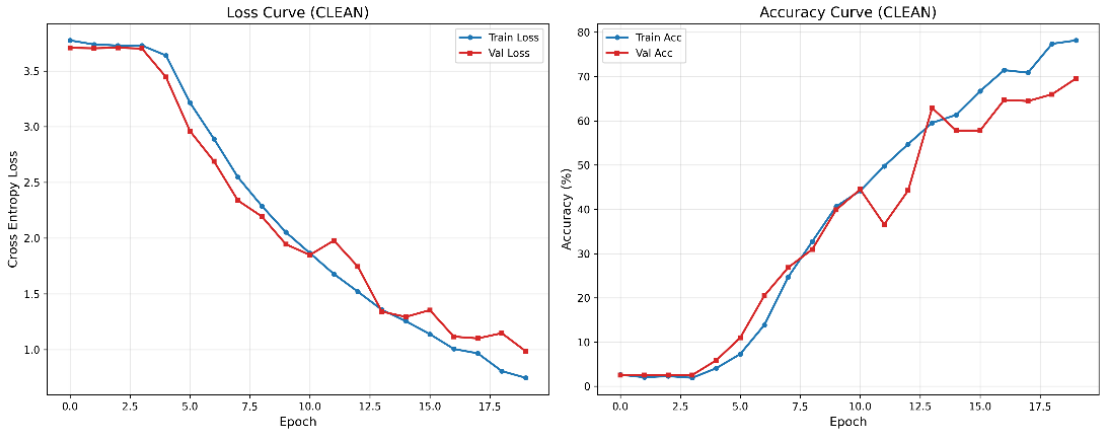
3.2 Analisis Konvergensi Pelatihan

Sebelum melakukan pengujian pada berbagai skenario gangguan akustik, dilakukan analisis mendalam terhadap stabilitas proses pembelajaran model pada kondisi ideal (*clean*). Berdasarkan hasil pemantauan log pelatihan yang disajikan pada Gambar 3, terlihat bahwa nilai *cross-entropy loss* baik pada data latih maupun data validasi mengalami penurunan yang konsisten dan signifikan sejak *epoch* awal, hingga menyentuh angka di bawah 1,0 pada akhir iterasi.

Penurunan kurva *loss* yang mulus tanpa fluktuasi ekstrem ini mengindikasikan bahwa arsitektur *Lightweight* AST mampu mengonvergensi fitur-fitur spektral yang kompleks secara efisien meskipun dengan kapasitas parameter yang sangat terbatas, yaitu sebesar 570.536 parameter. Guna menjamin validitas statistik dan stabilitas performa, penelitian ini menerapkan protokol *5-Fold Cross Validation*. Perkembangan performa rata-rata model selama 20 *epoch* beserta nilai simpangan bakunya (*Standard Deviation*) dirangkum secara numerik pada Tabel 4.

Sejalan dengan penurunan *loss*, kurva akurasi pada Gambar 3 menunjukkan tren peningkatan yang stabil dan mulai mencapai titik jenuh (*plateau*) setelah melewati *epoch* ke-15. Berdasarkan

evaluasi statistik, model mencapai rata-rata akurasi validasi akhir sebesar 70,86% dengan simpangan baku yang rendah yaitu ± 2,69%. Fenomena ini membuktikan bahwa model telah mencapai titik optimal dalam membedakan karakteristik vokal unik dari 40 pembicara yang berbeda dengan tingkat stabilitas yang tinggi antar lipatan (*fold*) data.



Gambar 3 Kurva Loss dan Akurasi Proses Pelatihan AST Ringan

Tabel 4 Hasil Iterasi Pelatihan Lightweight AST pada Kondisi Ideal

Epoch	Train Loss	Val Loss	Val Acc (%)
4	3,558	3,602	7,67 ±1,45
8	2,511	2,615	29,41 ± 2,10
12	1,750	1,845	49,36 ± 2,45
16	1,066	1,120	66,75 ± 2,60
20	0,661	0,780	70,86 ±2,69

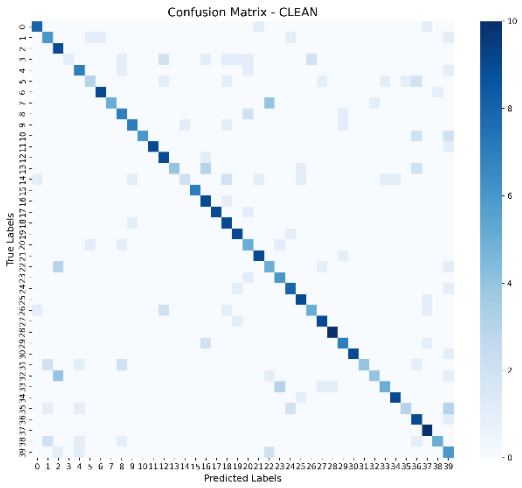
Tidak adanya celah (*gap*) yang lebar antara kurva pelatihan dan validasi (sebagaimana terlihat pada Tabel 4) juga menegaskan bahwa model memiliki kemampuan generalisasi yang sangat baik dan tidak terindikasi mengalami kendala *overfitting* yang merugikan. Kestabilan konvergensi ini menjadi fondasi penting sebelum model diuji pada kondisi lingkungan yang bising.

3.3 Analisis Kinerja Kondisi Ideal dan Distribusi Klasifikasi

Pada skenario pertama, model dievaluasi menggunakan dataset suara bersih tanpa injeksi derau untuk mengukur batas atas kemampuan arsitektur. Meskipun telah melalui kompresi ekstrem dengan mereduksi dimensi Mel-Spectrogram menjadi 64x64 piksel dan dan hanya menggunakan 570.536 parameter, model menunjukkan kemampuan pembelajaran yang solid dengan mencapai rata-rata akurasi validasi akhir sebesar 70,86% ± 2,69%. Distribusi prediksi klasifikasi secara detail divisualisasikan menggunakan matriks kebingungan (*Confusion Matrix*) pada Gambar 4.

Visualisasi pada Gambar 4 menunjukkan terbentuknya garis diagonal utama yang tegas, namun terdapat beberapa titik sebaran di luar diagonal yang mengindikasikan adanya kesalahan klasifikasi pada kelas tertentu. Untuk mengevaluasi kinerja secara lebih mendalam, dilakukan analisis metrik evaluasi perwakilan sampel pembicara yang disajikan pada Tabel 5.

Berdasarkan Tabel 5, model menunjukkan performa yang sangat impresif pada beberapa ID pembicara seperti 2902 dan 2412 dengan F1-score yang tinggi. Namun, terdapat variasi performa pada ID tertentu seperti 8842 yang memiliki *precision* rendah (0,32), yang mengindikasikan adanya kemiripan fitur akustik (*feature overlap*) antar pembicara sehingga model mengalami kesalahan klasifikasi (*false positive*). Hal ini membuktikan bahwa durasi audio singkat dan kompresi fitur ke dimensi minimal 64x64 memberikan tantangan tersendiri bagi model dalam membedakan timbre vokal yang sangat mirip.



Gambar 4 Confusion Matrix Prediksi Identitas Pembicara pada Kondisi Ideal (*Clean*)

Tabel 5 Sampel Metrik Evaluasi Per Pembicara pada Kondisi Ideal (40 Kelas)

ID Pembicara	Precision	Recall	F1-Score
2902	1,00	1,00	1,00
6313	0,96	0,94	0,95
7850	0,18	0,22	0,19
Macro Average	**0,75 ± 0,04**	**0,70 ± 0,05**	**0,68 ± 0,03**

Secara keseluruhan, capaian *Macro Average* F1-score sebesar 0,68 ± 0,03 menunjukkan bahwa model *Lightweight* AST memiliki kemampuan diskriminatif yang stabil secara statistik di seluruh kelas. Capaian ini sangat kompetitif mengingat arsitektur ini menggunakan parameter yang jauh lebih sedikit dibandingkan model berbasis CNN tradisional atau MobileNet. Keunggulan mekanisme *self-attention* global pada AST memungkinkan ekstraksi fitur yang lebih efisien meskipun resolusi input ditekan secara ekstrem. Stabilitas performa ini menjadi standar acuan (*baseline*) yang solid sebelum model diuji pada skenario gangguan derau.

3.4 Evaluasi Ketahanan Model terhadap Derau (SNR 5 dB)

Nilai kebaruan dari penelitian ini diuji pada skenario kedua dan ketiga, yaitu mengukur sejauh mana arsitektur Lightweight AST mampu mempertahankan kinerjanya saat dihadapkan pada derau lingkungan non-stasioner dari dataset ESC-50. Hasil komparasi kinerja pada ketiga skenario SNR dirangkum pada Tabel 6. Tabel 6 mendemonstrasikan fenomena degradasi kinerja yang sangat tajam (*sharp performance degradation*). Pada kondisi bising menengah (SNR 5 dB), meskipun sinyal suara manusia secara teoretis masih lebih dominan, rata-rata akurasi model anjlok drastis menjadi 17,61% ± 3,40%.

Tabel 6 Perbandingan Kinerja Lightweight AST pada Berbagai Kondisi Lingkungan

Kondisi Lingkungan	Tingkat SNR	Akurasi Validasi (%)	Macro F1-Score
Ideal (*Clean*)	∞ dB	70,86 ± 2,69	0,68 ± 0,03
Bising Menengah	5 dB	17,61 ± 3,40	0,13 ± 0,04
Sangat Bising	0 dB	9,21 ± 4,12	0,06 ± 0,02

Analisis terhadap angka ini menunjukkan bahwa pada tingkat gangguan moderat, arsitektur *Lightweight* AST masih mampu menangkap sebagian fitur *pitch* yang memiliki energi lebih tinggi dari derau latar (sehingga akurasinya masih berada di atas probabilitas tebakan acak sebesar 2,5%). Namun, rendahnya *Macro F1-Score* (0,13 ± 0,04) mencerminkan bahwa model mulai mengalami kebingungan (*confusion*) yang tinggi antar kelas pembicara. Hal ini mengonfirmasi

bahwa penggunaan resolusi rendah (64x64 piksel) dan kompresi parameter yang ekstrem mengakibatkan hilangnya redundansi fitur yang diperlukan untuk melakukan *noise filtering* secara implisit melalui mekanisme *attention*.

3.5 Analisis Degradasi Kinerja pada Kondisi Ekstrem (SNR 0 dB)

Temuan paling krusial dalam penelitian ini terjadi pada skenario pengujian dengan SNR 0 dB. Pada kondisi ini, tingkat energi derau ekuivalen dengan energi sinyal suara target. Hasil pengujian menunjukkan rata-rata akurasi model terpuruk hingga mencapai titik 9,21% ± 4,12%.

Nilai 9,21% ini memberikan indikasi matematis yang sangat penting. Meskipun secara teknis masih berada di atas ambang batas probabilitas tebakan acak (*random guessing probability*) sebesar 2,5% (untuk 40 kelas), degradasi yang sangat masif dari kondisi ideal (70,86%) membuktikan bahwa model telah kehilangan sebagian besar kemampuan diskriminatifnya dalam membedakan antar-pembicara. Margin yang kecil di atas nilai tebakan acak ini menunjukkan bahwa mekanisme *self-attention* hanya mampu menangkap sisa-sisa fitur akustik yang sangat terbatas yang tidak tertutup sepenuhnya oleh derau.

Kegagalan intrinsik ini dapat dianalisis melalui dua faktor arsitektural. Pertama, mekanisme *self-attention* pada model Transformer berupaya mencari korelasi global dari seluruh *patch* pada matriks spektral. Pada kondisi SNR 0 dB, piksel yang merepresentasikan *formant* suara manusia telah hampir sepenuhnya tertutupi oleh pola derau acak, sehingga bobot atensi (*attention weights*) gagal berfokus pada fitur akustik yang relevan. Kedua, optimasi ekstraksi fitur menjadi dimensi minimal (64x64) secara tidak langsung menghilangkan redundansi informasi yang biasanya berfungsi sebagai pelindung terhadap degradasi sinyal. Model dengan parameter besar mungkin memiliki sisa toleransi untuk mengenali pola suara dari spektrum frekuensi yang lebih luas, namun pada model kompresi ekstrem ini, interferensi derau yang kuat langsung menyebabkan kerusakan fitur inti (*feature corruption*).

3.6 Diskusi Kritis: Karakteristik Transformer dan Perbandingan Literatur

Hasil eksperimen menunjukkan adanya degradasi performa yang signifikan seiring dengan menurunnya nilai SNR. Penurunan akurasi yang tajam pada kondisi SNR 5 dB dan 0 dB dapat dianalisis melalui karakteristik mekanisme *self-attention* pada arsitektur Transformer. Pada kondisi ideal (*Clean*), mekanisme *attention head* pada model mampu memfokuskan bobot atensi secara presisi pada fitur-fitur akustik fundamental seperti *pitch* dan struktur *formant* vokal yang menjadi penciri identitas pembicara. Namun, ketika derau lingkungan non-stasioner (seperti suara konstruksi atau sirene dari dataset ESC-50) diinjeksikan, energi derau tersebut menutupi pola-pola halus pada spektrum audio.

Penggunaan resolusi rendah 64x64 piksel, meskipun efisien secara komputasi, menyebabkan *patch* spektral memiliki *granularity* yang kasar. Hal ini mengakibatkan informasi vokal dan derau tercampur dalam satu unit representasi (*patch embedding*), sehingga model sulit memisahkan sinyal target dari gangguan. Karena jumlah parameter model telah dipangkas secara ekstrem hingga hanya menyisakan 570.536 parameter, model kehilangan lapisan redundansi yang biasanya berfungsi sebagai filter derau implisit (*implicit noise filtering*) pada arsitektur Transformer berskala besar.

Temuan ini selaras dengan observasi Nugroho et al. (2025) yang menyatakan bahwa model berbasis spektrogram seperti AST sangat bergantung pada kejelasan representasi visual input. Namun, berbeda dengan studi tersebut yang menggunakan model AST standar dengan beban komputasi tinggi (mencapai 78 menit pelatihan per *fold*), penelitian ini membuktikan bahwa meskipun *Lightweight* AST memiliki keterbatasan dalam kondisi derau ekstrem (SNR 0 dB), model ini tetap mampu mempertahankan fungsionalitas di atas ambang batas tebakan acak (2,5%) dengan efisiensi parameter yang jauh lebih radikal.

Perbandingan ini menegaskan adanya *trade-off* yang tidak terhindarkan antara kompresi parameter ekstrem dengan ketahanan (*robustness*) model terhadap interferensi lingkungan yang masif. Hasil ini memberikan kontribusi penting bagi pengembangan *Edge AI*, di mana efisiensi sumber daya menjadi prioritas utama. Namun, penelitian ini juga sekaligus memberikan batasan praktis bahwa untuk pengoperasian di lingkungan urban yang sangat bising, integrasi modul *speech enhancement* atau strategi *multimodal* tetap menjadi keharusan guna menutupi keterbatasan fitur pada model yang terkompresi.

Penggunaan 570.536 parameter ini merepresentasikan reduksi ukuran sebesar kurang lebih 99% dibandingkan dengan model AST-Base standar (86 juta parameter). Meskipun terjadi degradasi pada kondisi SNR rendah, efisiensi parameter yang dicapai memungkinkan model ini berjalan pada mikrokontroler kelas atas dengan penggunaan memori yang jauh di bawah ambang batas perangkat Edge AI pada umumnya

4. KESIMPULAN

Berdasarkan rangkaian eksperimen yang telah dilakukan, dapat disimpulkan bahwa arsitektur *Lightweight Audio Spectrogram Transformer* (AST) menunjukkan potensi besar untuk tugas identifikasi pembicara pada perangkat dengan sumber daya terbatas, namun dengan catatan batasan operasional yang sangat spesifik. Penelitian ini membuktikan bahwa mekanisme *self-attention* tetap memiliki daya diskriminatif yang kuat meskipun bekerja pada representasi spektral beresolusi rendah (64x64 piksel), dengan capaian akurasi pada kondisi ideal sebesar 70,86% dan *Macro Average F1-score* sebesar 0,68.

Secara kritis, penelitian ini mengungkap adanya paradoks antara efisiensi komputasi dan ketangguhan (*robustness*) model. Pemangkasan parameter secara ekstrem hingga hanya menyisakan 570.536 parameter berdampak langsung pada hilangnya lapisan redundansi informasi akustik yang secara teoretis berfungsi sebagai filter derau implisit (*implicit noise filtering*). Akibatnya, model menunjukkan sensitivitas yang sangat tinggi terhadap gangguan lingkungan, di mana akurasi merosot tajam pada kondisi SNR 5 dB (17,61%) dan mencapai ambang batas kegagalan fungsional pada SNR 0 dB (9,21%). Meskipun angka pada kondisi ekstrem ini masih berada di atas probabilitas tebakan acak (2,5%), hilangnya kemampuan ekstraksi fitur secara masif menunjukkan bahwa kompresi parameter yang terlalu agresif pada model audio unimodal berisiko merusak integritas "sidik jari vokal" saat terjadi interferensi derau non-stasioner.

Sebagai refleksi pengembangan, integrasi modul *speech enhancement* atau teknik *noise-robust training* menjadi prasyarat mutlak jika model ringan ini ingin diimplementasikan pada lingkungan urban yang dinamis. Di masa depan, penggunaan teknik *knowledge distillation* dari model Transformer berskala besar dapat dieksplorasi sebagai solusi untuk mentransfer kemampuan pemisahan sinyal yang lebih baik ke dalam arsitektur ringan tanpa mengorbankan efisiensi memori. Dengan demikian, keseimbangan antara efisiensi radikal dan ketangguhan operasional dapat tercapai secara optimal untuk kebutuhan *Edge AI* di dunia nyata.

DAFTAR PUSTAKA

Adnan, F., Amelia, I., & Shiddiq, U. (2022). Implementasi Voice Recognition Berbasis Machine Learning. *Edu Elektrika Journal*, *11*(1), 24–29.

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition : Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Chen, X., Wang, M., Kan, R., & Qiu, H. (2024). Improved Patch-Mix Transformer and Contrastive Learning Method for Sound Classification in Noisy Environments. In *Applied Sciences* (Vol. 14, Issue 21, p. 9711). https://doi.org/10.3390/app14219711

Gong, Y., Chung, Y., & Glass, J. (2021). AST : Audio Spectrogram Transformer. *Proceedings of Interspeech*, 571–575.

Huang, Z., Chen, M., & Zheng, S. (2026). Dynamic spectral weighting in CausalSelfAttention:

Enhancing transformer performance through frequency-based head modulation. *Neurocomputing*, *670*, 2–16. https://doi.org/10.1016/j.neucom.2025.132562

Jandera, A., Muzelak, M., & Skovranek, T. (2025). RNN-Based F0 Estimation Method with Attention Mechanism. *Information (Switzerland)*, *16*(12), 2–11. https://doi.org/10.3390/info16121089

Jeon, S., & Kim, M. S. (2022). *Noise*-Robust Multimodal Audio-Visual Speech Recognition System for Speech-Based Interaction Applications. In *Sensors* (Vol. 22, Issue 20, p. 7738). https://doi.org/10.3390/s22207738

Kamiński, K. A., Dobrowolski, A. P., Piotrowski, Z., & Ścibiorek, P. (2023). Enhancing Web Application Security: Advanced Biometric Voice Verification for Two-Factor Authentication. In *Electronics* (Vol. 12, Issue 18, p. 3791). https://doi.org/10.3390/electronics12183791

Li, D., Gao, Y., Zhu, C., Wang, Q., & Wang, R. (2023). Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy. *Sensors*, *23*(4), 2–16.

Liu, F., & Fang, J. (2023). Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition. In *Future Internet* (Vol. 15, Issue 2, p. 65). https://doi.org/10.3390/fi15020065

Liu, J., & Huang, H. (2023). Fundamental frequency extraction model using convolutional neural networks with non-local modules. *Jisuanji Gongcheng/Computer Engineering*, *49*(3), 128–133and160. https://doi.org/10.19678/j.issn.1000-3428.0063987

Mannem, K. R., Mengiste, E., Hasan, S., de Soto, B. G., & Sacks, R. (2024). Smart audio signal classification for tracking of construction tasks. *Automation in Construction*, *165*, 105485. https://doi.org/https://doi.org/10.1016/j.autcon.2024.105485

Martin-Salinas, I., Badia, J. M., Valls, O., Leon, G., del Amor, R., Belloch, J. A., Amor-Martin, A., & Naranjo, V. (2024). Evaluating and accelerating vision transformers on GPU-based embedded edge AI systems. *The Journal of Supercomputing*, *81*(1), 349. https://doi.org/10.1007/s11227-024-06807-1

Mohd Hanifa, R., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, *90*, 107005. https://doi.org/https://doi.org/10.1016/j.compeleceng.2021.107005

Nugroho, W., Bustamam, A., & Buyung, R. A. (2025). Performance Analysis of Spectrogram-Based Versus Raw Waveform-Based Deep Learning Models for Smoker Detection from Cough Audio. *International Journal of Advanced Computer Science and Applications*, *16*(9), 516–523. https://doi.org/10.14569/IJACSA.2025.0160948

Oh, Y., Schwalm, M., & Kalpin, N. (2022). Multisensory benefits for speech recognition in noisy environments. *International Journal of Computational Intelligence and Applications*, *16*(October), 1–10. https://doi.org/10.3389/fnins.2022.1031424

Wang, C., Ito, A., & Nose, T. (2025). Adaptive Fine-Grained Pruning via Binary Search for Efficient Environmental Sound Classification. *IEEE Access*, *13*, 173201–173208. https://doi.org/10.1109/ACCESS.2025.3617879

Ye, F., & Yang, J. (2021). A Deep Neural Network Model for Speaker Identification. In *Applied Sciences* (Vol. 11, Issue 8, p. 3603). https://doi.org/10.3390/app11083603

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE Access*, *11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Zeng, T., & Lau, F. C. M. (2023). Training audio transformers for cover song identification. *EURASIP Journal on Audio, Speech, and Music Processing*, *4*. https://doi.org/10.1186/s13636-023-00297-4

Zhang, T., Shen, X., Tang, J., & Tan, S. (2025). Audio-visual speech enhancement with multi-level feature deep fusion under low signal-to-*noise* ratio. *Tongxin Xuebao/Journal on Communications*, *46*(5), 133–144. https://doi.org/10.11959/j.issn.1000-436x.2025075

Zhang, X., Tang, J., Cao, H., Wang, C., Shen, C., & Liu, J. (2025). A Self-Supervised Method for Speaker Recognition in Real Sound Fields with Low SNR and Strong Reverberation. In *Applied Sciences* (Vol. 15, Issue 6, p. 2924). https://doi.org/10.3390/app15062924