PENGEMBANGAN PENGENDALI KARAKTER PERMAINAN MENGGUNAKAN SUARA REAL-TIME BERBASIS TRANSFORMER

Adrik Fikhtiyaaril Amro[1], dan Oddy Virgantara Putra[2]
[1,2] Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Darussalam Gontor
Jl. Raya Siman, Dusun I, Demangan, Kec. Siman, Kabupaten Ponorogo, Jawa Timur 63472
Email: adrikfikhtiyaarilamro41@student.cs.unida.gontor.ac.id
Email: oddy@unida.gontor.ac.id

Abstrak

Kompleksitas interaksi pada permainan digital modern menuntut skema kendali alternatif untuk meminimalkan beban kognitif pemain. Penelitian ini mengembangkan dan menganalisis sistem pengendali karakter permainan menggunakan suara secara real-time berbasis Deep Learning. Studi ini bertujuan mengevaluasi trade-off performa arsitektur Transformer murni dibandingkan dengan BiLSTM dan Hybrid (BiLSTM-Transformer) saat dilatih dari awal (from-scratch) pada dataset berskala kecil. Kami menerapkan ekstraksi fitur Mel-Frequency Cepstral Coefficients (MFCC) dan evaluasi Stratified 5-Fold Cross-Validation dengan augmentasi noise sintetik. Hasil eksperimen menunjukkan bahwa model Hybrid mencapai rata-rata akurasi tertinggi sebesar 97,53%, mengungguli arsitektur BiLSTM (95,89%) dan Transformer murni (94,81%). Analisis membuktikan bahwa Transformer murni kurang efisien dalam pemanfaatan data pada skala kecil, sedangkan integrasi pemrosesan temporal lokal BiLSTM dan atensi global Transformer pada model Hybrid menghasilkan stabilitas terbaik serta ketahanan (robustness) tinggi terhadap gangguan lingkungan. Meskipun efisiensi komputasi dievaluasi melalui throughput, penelitian ini juga melakukan pengukuran latensi inferensi pada perangkat keras GPU. Hasil eksperimen menunjukkan model Hybrid mencapai akurasi tertinggi 97,53% dengan latensi inferensi rata-rata 2,46 ms (RTF 0,0024), membuktikan kapabilitasnya untuk beroperasi secara real-time dengan jeda perseptual yang minimal.

***Kata kunci**: BiLSTM, , Hybrid Model, MFCC, Pengenalan Suara, Transformer.*

1. PENDAHULUAN

Industri permainan digital modern menghadapi tantangan dalam desain interaksi, di mana kompleksitas gameplay sering kali membebani kapasitas kognitif pemain jika hanya mengandalkan pengendali fisik seperti keyboard atau gamepad. Dalam skenario permainan yang intens, otak pemain dipaksa mengontrol banyak tugas secara simultan, yang dapat menyebabkan kelelahan kognitif (Wang 2023). Untuk mengatasi hal ini, diperlukan modalitas interaksi paralel, seperti penggunaan perintah suara, yang dapat mendistribusikan beban input dan memperkaya pengalaman pemain. Namun, tantangan teknis utama dalam implementasi kendali suara adalah latensi dan ketahanan terhadap noise (noise robustness), mengingat game menuntut respons sepersekian detik (Waqar dkk. 2021).

Kemajuan *Deep Learning*, khususnya arsitektur *Transformer* yang diperkenalkan oleh (Vaswani dkk. 2017) , menawarkan potensi solusi untuk pemrosesan sekuensial yang lebih baik. Meskipun *Transformer* mendominasi tugas pengenalan suara pada dataset skala besar, performanya pada dataset skala kecil yang dilatih dari awal (*from-scratch*) masih menjadi pertanyaan kritis (Loubser dkk. 2024). Penelitian terdahulu mencatat tantangan efisiensi data pada *Transformer* dibandingkan arsitektur rekuren tradisional seperti *Long Short-Term Memory* (LSTM) atau BiLSTM (Zaman dkk. 2025).

Penelitian ini difokuskan pada skenario perangkat dengan sumber daya terbatas (*resource-constrained*). Oleh karena itu, penggunaan dataset berskala kecil dan ekstraksi fitur ringan seperti MFCC dipilih secara sadar dibandingkan metode *end-to-end* modern yang menuntut komputasi berat. Keputusan ini bertujuan untuk menyeimbangkan akurasi dan beban komputasi. Validasi empiris terhadap aspek latensi inferensi juga dilakukan untuk membuktikan klaim responsivitas sistem secara *real-time*.

2. METODOLOGI

2.1. Dataset dan Augmentasi

Penelitian ini menggunakan dataset 15.219 sampel audio (4 kelas: up, down, left, right) dengan rata-rata ~3.800 sampel per kelas, yang terbukti memadai untuk closed-set recognition jika didukung regularisasi ketat (Loubser dkk, 2024; Zaman dkk, 2025) . Untuk memitigasi overfitting dan meningkatkan robustness, diterapkan strategi komprehensif: (1) Augmentasi probabilistik pada 30% data latih menggunakan Pink/White Noise (SNR 5-20 dB), (2) Evaluasi Stratified 5-Fold Cross-Validation (Ünalan dkk, 2024) , dan (3) Validasi multi-seed (36, 38, 42).

2.2. Ekstraksi Fitur

Dalam penelitian ini, MFCC dipilih sebagai metode ekstraksi fitur utama karena efisiensi komputasinya yang tinggi, menjadikannya solusi ideal untuk skenario perangkat dengan sumber daya terbatas (*resource-constrained*). Berbeda dengan metode modern seperti *wav2vec 2.0* atau *HuBERT* yang menuntut beban komputasi masif, MFCC menawarkan representasi fitur yang jauh lebih ringan namun tetap efektif untuk dataset berskala kecil.

Proses ekstraksi dioptimalkan dengan menetapkan 30 koefisien MFCC. Keputusan ini mengadopsi temuan empiris dari (Yan dkk, 2025), studi tersebut membuktikan bahwa penggunaan koefisien di atas standar (biasanya 13) meningkatkan akurasi sistem secara signifikan. Selain itu, digunakan *framing* 25ms dan *hop* 10ms untuk menjaga resolusi temporal (Tirronen dkk, 2024). Parameter ini disesuaikan secara adaptif per *sampling rate* (16/8 kHz) menggunakan 40 *mel filterbanks*. Guna meningkatkan *robustness* terhadap gangguan lingkungan, diterapkan augmentasi *white noise* berbasis *waveform* dengan SNR acak 5-20 dB pada 30% data latih.

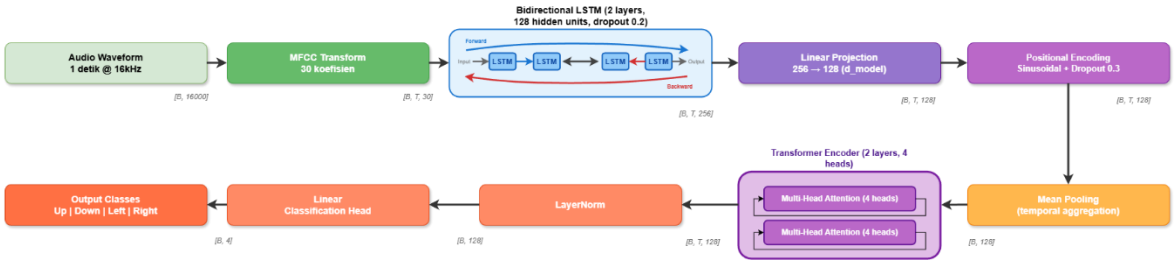
2.3. Arsitektur Model

Tiga konfigurasi arsitektur dievaluasi dalam penelitian ini:

2.3.1. **BiLSTM:** Menggunakan 2 layer BiLSTM dengan 128 hidden units per arah, diikuti oleh dropout 0.2 (Wang dkk. 2025).

2.3.2. **Transformer:** Arsitektur *encoder-only* dengan *Multi-Head Self-Attention* (Vaswani dkk. 2017). Input MFCC diproyeksikan ke dimensi model, ditambah *Positional Encoding* sinusoidal, dan diproses melalui blok *encoder* dengan normalisasi *layer* (Zaman dkk. 2025).

2.3.3. **Hybrid (BiLSTM-Transformer):** Menggabungkan keunggulan pemodelan temporal lokal dan global. Pada **Gambar 1** input diproses terlebih dahulu oleh 2 *layer* BiLSTM. Output BiLSTM kemudian diproyeksikan secara linear dan menjadi input bagi *Transformer Encoder* (Andayani dkk. 2022; Alsuwaylimi 2024).



Gambar 1. Hybrid Architecture (BiLSTM-Transformer)

2.4. Pengaturan Eksperimen (*Experimental Setup*)

Pengoptimalan model dilakukan menggunakan AdamW optimizer dengan gradient clipping dan Automatic Mixed Precision (AMP). Pelatihan dijalankan selama 10 epoch dengan batch size 64. Parameter MFCC disesuaikan per sampling rate (25ms window, 10ms hop), sementara arsitektur dioptimalkan secara spesifik: Transformer menggunakan pre-normalization dan aktivasi GELU (d_model=128, 4 heads, dropout 0.3), BiLSTM menerapkan inisialisasi orthogonal/xavier (hidden=128, 2 layers, dropout 0.2), dan model Hybrid menggabungkan konfigurasi keduanya untuk

stabilitas maksimal. Pengukuran inferensi dilakukan menggunakan lingkungan komputasi server dengan GPU NVIDIA Tesla P100. Pengukuran latensi dilakukan dalam mode evaluasi (batch size 1) untuk mensimulasikan skenario input tunggal *real-time*.

2.5. Evaluasi

Evaluasi kinerja model dilakukan menggunakan metode *Stratified 5-Fold Cross-Validation* untuk memastikan distribusi kelas yang seimbang pada setiap lipatan pengujian (Ünalan et al. 2024). Selain akurasi klasifikasi standar, penelitian ini menggunakan metrik khusus untuk mengukur stabilitas pelatihan terhadap inisialisasi bobot acak (*random seed*), mengingat arsitektur *Transformer* cenderung sensitif saat dilatih pada dataset berskala kecil (Loubser et al. 2024). Keseluruhan proses validasi silang diulang sebanyak tiga kali dengan *seed* yang berbeda.

Variabilitas kinerja diukur menggunakan Simpangan Baku Rata-rata (*Standard Deviation of Seed Means*), yang dinotasikan sebagai σ_{seed} dan dihitung menggunakan Persamaan (1):

$$\sigma_{\text{seed}} = \sqrt{\frac{\sum_{\text{i=1}}^{\text{N}}(\text{x}_\text{i} - \bar{\text{x}})^2}{\text{N} - 1}} \qquad (1)$$

Dimana x_i adalah rata-rata akurasi pada *seed* ke$-i$, $\bar{x}$ adalah rata-rata global dari seluruh *seed*, dan N adalah jumlah pengulangan eksperimen ($N = 3$).

Selanjutnya, untuk memberikan kuantifikasi tingkat kestabilan yang lebih intuitif (di mana nilai yang lebih besar menunjukkan performa yang lebih baik), didefinisikan metrik *Stability Score* sebagai invers dari simpangan baku, sebagaimana dirumuskan pada Persamaan (2):

$$\text{Stability Score} = \frac{1}{\sigma_{seed} + \epsilon} \qquad (2)$$

Nilai ϵ adalah konstanta kecil (10^{-8}) untuk menghindari pembagian dengan nol. Untuk analisis performa per kelas, digunakan kurva *Receiver Operating Characteristic* (ROC) yang memplot *True Positive Rate* (TPR) terhadap *False Positive Rate* (FPR), serta nilai *Area Under Curve* (AUC) untuk mengukur kemampuan separabilitas model secara keseluruhan.

Terakhir, untuk memvalidasi responsivitas sistem secara *real-time*, pengukuran latensi inferensi dilakukan sebagai rata-rata waktu eksekusi *forward pass* dari $N = 1000$ sampel uji acak dengan *batch size* 1. Pengukuran ini menggunakan sinkronisasi waktu CUDA (cuda.synchronize()) untuk memastikan akurasi pencatatan waktu proses pada perangkat keras GPU (3).

$$Latensi = \frac{\sum_{i=1}^{N}\left(t_{end\_i} - t_{start\_i}\right)}{N} \quad (3)$$

3. HASIL DAN PEMBAHASAN

3.1. Akurasi

Hasil eksperimen menunjukkan perbedaan performa yang signifikan antar arsitektur. Sebagaimana ditampilkan pada Tabel 1, model *Hybrid* (BiLSTM-Transformer) mencatat akurasi rata-rata tertinggi sebesar 97,53%, mengungguli BiLSTM murni (95,89%) dan *Transformer* murni (94,81%).

Tabel 1 perbedaan performa yang signifikan antar arsitektur

Model	Grand_MeanAcc	Weighted_MeanAcc
Hybrid	**97,53%**	**97,53%**
BiLSTM	95,89%	95,89%
Transformer	94,81%	94,81%

Analisis statistik mengonfirmasi bahwa keunggulan model Hybrid sangat signifikan $p < 0.001$. Hasil ini membuktikan bahwa integrasi pemrosesan sekuensial lokal oleh BiLSTM secara

efektif mengatasi inefisiensi data Transformer pada dataset kecil, menghasilkan performa superior yang konsisten dengan temuan (Mou dkk. 2026).

Selain mengukur akurasi rata-rata secara global, penelitian ini juga melakukan analisis mendalam terhadap kinerja per kelas untuk memastikan tidak terjadi bias pada perintah tertentu. Berdasarkan Tabel 2, dari nilai *Area Under the Receiver Operating Characteristic Curve* (ROC-AUC) menunjukkan kinerja diskriminasi yang sangat kuat dan seimbang di seluruh kelas perintah (*Up, Down, Left, Right*).

Terlihat bahwa nilai ROC-AUC secara konsisten berada pada rentang 99,60% hingga 100,00% baik pada *sampling rate* 8 kHz maupun 16 kHz . Nilai yang mendekati sempurna ini mengindikasikan bahwa model memiliki separabilitas kelas yang tinggi, di mana kesalahan klasifikasi yang terjadi cenderung bersifat acak akibat variabilitas *noise*, bukan karena kegagalan sistematis model dalam mengenali kata tertentu. Hal ini membuktikan bahwa arsitektur yang diusulkan memiliki kemampuan generalisasi yang tangguh (*robust*) untuk setiap perintah suara yang berbeda.

Tabel 2 Nilai ROC-AUC

Model	SR	Down	Left	Right	Up	Average
Hybrid	8000	99.90%	99.83%	99.89%	99.92%	99.89%
	16000	99.89%	99.86%	99.89%	99.92%	99.89%
BiLSTM	8000	99.81%	99.72%	99.80%	99.84%	99.79%
	16000	99.81%	99.73%	99.81%	99.83%	99.79%
Transformer	8000	99.90%	99.83%	99.89%	99.92%	99.89%
	16000	99.89%	99.86%	99.89%	99.92%	99.89%

3.2. Ketahanan terhadap Noise (*Robustness*)

Pengujian terhadap *noise* sintetik pada **Tabel 3** menunjukkan bahwa semua model memiliki ketahanan yang baik. Tidak ditemukan perbedaan performa yang signifikan antara gangguan *Pink Noise* dan *White Noise* (selisih ≈ 0.0 poin persentase). Konsistensi ini menunjukkan bahwa fitur MFCC yang dikombinasikan dengan arsitektur model berhasil mengisolasi konten akustik yang relevan terlepas dari jenis spektrum *noise* (Yan dkk. 2025).

Tabel 3 Pengujian terhadap noise sintetik

Model	Noise	SR	Accuracy				Avg Times	Total Times
			Mean	Std	Min	Max		
BiLSTM	pink	8000	95.79%	0.30%	95.11%	96.12%	88.52	26556.8
	pink	16000	95.99%	0.46%	95.34%	97.14%	66.56	19966.73
	white	8000	95.79%	0.30%	95.11%	96.12%	88.52	26556.8
	white	16000	95.99%	0.46%	95.34%	97.14%	66.56	19966.73
Transformer	pink	8000	94.74%	0.34%	94.22%	95.30%	37.15	371.53
	pink	16000	94.89%	0.54%	94.02%	96.29%	31.68	316.76
	white	8000	94.74%	0.34%	94.22%	95.30%	37.15	371.53
	white	16000	94.89%	0.54%	94.02%	96.29%	31.68	316.76
Hybrid	pink	8000	**97.49%**	0.19%	97.14%	**97.90%**	40.54	12161.74
	pink	16000	**97.56%**	0.19%	97.21%	**97.83%**	34.3	10289.55
	white	8000	**97.49%**	0.19%	97.14%	**97.90%**	40.54	12161.74
	white	16000	**97.56%**	0.19%	97.21%	**97.83%**	34.3	10289.55

3.3. Analisis Efisiensi dan Stabilitas

Meskipun *Transformer* murni pada **Tabel 4** memiliki jumlah parameter yang jauh lebih besar (1,69 juta) dibandingkan BiLSTM (560 ribu), akurasinya justru paling rendah pada skenario pelatihan *from-scratch*. Namun, dari segi efisiensi komputasi (*throughput*), arsitektur berbasis *Transformer* (termasuk *Hybrid*) menunjukkan kecepatan pelatihan sekitar 2,2 kali lebih tinggi dibandingkan BiLSTM karena kemampuan komputasi paralelnya (Karmakar dkk. 2024). Dalam hal

stabilitas, model *Hybrid* menunjukkan varians antar-*seed* yang rendah, menawarkan keseimbangan terbaik antara akurasi tinggi dan konsistensi pelatihan.

Tabel 4 Stability Score

Model	SR	Stability Score	Sd_Seed Means
Hybrid	8000	3905.671509	0.0002560379176
	16000	2784.223238	0.0003591666021
BiLSTM	8000	677.6650569	0.001475655252
	16000	466.2354928	0.002144838854
Transformer	8000	1096.160545	0.000912275127
	16000	8167.534563	0.0001224359679

3.4. Analisis Interferensi

Pada **Tabel 5** menunjukkan bahwa seluruh arsitektur beroperasi jauh di bawah ambang batas real-time (1000 ms untuk audio 1 detik). Model Hybrid mencatat rata-rata latensi 2,46 ms dengan Real-Time Factor (RTF) 0,0025, yang berarti sistem mampu memproses input 400 kali lebih cepat dari durasi aslinya.

Tabel 5 Interferensi

Model	Average Latency (ms)	Standard Deviation (ms)	Real-Time Factor (RTF)
Hybrid	2.46	0.02	0.0024
BiLSTM	2.23	0.10	0.0022
Transformer	0.90	0.12	0.0009

4. DISKUSI DAN KETERBATASAN PENELITIAN

Meskipun hasil eksperimen menunjukkan stabilitas tinggi, evaluasi pada dataset terkontrol ini memiliki keterbatasan dalam merepresentasikan variabilitas akustik dunia nyata (*real-world scenarios*), seperti keberagaman aksen pengguna dan gangguan *non-stationary noise*. Oleh karena itu, validasi generalisasi model di masa depan mutlak memerlukan pengujian lintas-korpus (*cross-dataset*) dengan diversitas pembicara yang lebih tinggi, khususnya untuk menjamin ketangguhan pada skenario *speaker-independent*.

Dari perspektif implementasi pada perangkat terbatas (*edge devices*), temuan kami menyoroti *trade-off* menarik antara efisiensi memori dan kecepatan. Meskipun BiLSTM unggul dalam efisiensi penyimpanan ($\approx$ 560 ribu parameter), evaluasi empiris membuktikan bahwa model Hybrid tetap mampu mempertahankan responsivitas superior dengan latensi inferensi rata-rata 2,46 ms per input. Dengan nilai *Real-Time Factor* (RTF) sebesar 0,0024, model ini terbukti memiliki *headroom* komputasi yang masif untuk beroperasi jauh di bawah ambang batas lag perseptual (100 ms), menjadikannya sangat layak untuk interaksi permainan *real-time*.

Namun, untuk menjembatani kendala ukuran model pada perangkat *low-end* dengan memori sangat ketat, strategi optimasi lanjutan seperti Kuantisasi (*Quantization*) atau *Knowledge Distillation* sangat disarankan. Pendekatan ini dapat mereduksi dimensi model Hybrid secara signifikan tanpa mengorbankan akurasi superior yang telah dicapai (97,53%). Terlepas dari performa komputasi tersebut, penelitian ini mengakui batasan pada cakupan kosakata (4 kelas), sehingga perluasan dataset menjadi langkah krusial untuk mewakili kompleksitas interaksi permainan modern.

5. KESIMPULAN

Berdasarkan hasil evaluasi dan diskusi, penelitian ini menyimpulkan bahwa arsitektur *Hybrid* (BiLSTM-Transformer) merupakan solusi paling optimal untuk sistem pengendali suara permainan pada dataset berskala kecil, dengan pencapaian akurasi tertinggi sebesar 97,53%. Model ini terbukti berhasil menggabungkan efisiensi ekstraksi fitur lokal dan pemahaman konteks global, mengungguli arsitektur *Transformer* murni yang cenderung *data-inefficient* jika dilatih dari awal (*from-scratch*) tanpa bantuan *Transfer Learning*.

Secara praktis, sistem yang dikembangkan menunjukkan kelayakan implementasi sebagai modalitas interaksi *hands-free* yang andal, ditandai dengan stabilitas performa yang tinggi terhadap gangguan lingkungan (*Pink* dan *White Noise*). Guna menyempurnakan sistem ini menuju adopsi pengguna secara luas, penelitian selanjutnya disarankan untuk memperluas evaluasi pada skenario *speaker-independent*. Selain itu, sangat krusial untuk memporting model ini ke dalam ekosistem perangkat seluler (Android). Integrasi ini bertujuan untuk memvalidasi ketangguhan generalisasi model terhadap variasi pengguna sekaligus menguji efisiensi manajemen daya di lingkungan sistem operasi seluler yang dinamis.

DAFTAR PUSTAKA

Alsuwaylimi, A.A. 2024. Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM. *Heliyon* 10(17). doi: 10.1016/j.heliyon.2024.e36280.

Andayani, F., Theng, L.B., Tsun, M.T. and Chua, C. 2022. Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access* 10, pp. 36018–36027. doi: 10.1109/ACCESS.2022.3163856.

Karmakar, P., Teng, S.W. and Lu, G. 2024. Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. *Intelligent Systems with Applications* 23. doi: 10.1016/j.iswa.2024.200406.

Loubser, A., De Villiers, P. and De Freitas, A. 2024. End-to-end automated speech recognition using a character based small scale transformer architecture. *Expert Systems with Applications* 252. doi: 10.1016/j.eswa.2024.124119.

Mou, H., Rong, H. and Teixeira, A.P. 2026. Detecting abnormal ship trajectory to avoid bridge collisions via a Transformer-BiLSTM model. *Ocean Engineering* 343. doi: 10.1016/j.oceaneng.2025.123232.

Tirronen, S., Kadiri, S.R. and Alku, P. 2024. The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection. *Journal of Voice* 38(5), pp. 975–982. doi: 10.1016/j.jvoice.2022.03.021.

Ünalan, S., Günay, O., Akkurt, I., Gunoglu, K. and Tekin, H.O. 2024. A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning. *Journal of Radiation Research and Applied Sciences* 17(4), p. 101080. doi: 10.1016/j.jrras.2024.101080.

Vaswani, A. et al. [no date]. *Attention Is All You Need*.

Wang, R. 2023. Research on human-computer interaction and game design. *Applied and Computational Engineering* 4(1), pp. 516–523. doi: 10.54254/2755-2721/4/2023316.

Wang, X., Zhang, P. and Liu, C. 2025. Acoustic signal-based identification of pipeline defects using optimized MFCC and LSTM. *Journal of Pipeline Science and Engineering*, p. 100355. doi: 10.1016/j.jpse.2025.100355.

Waqar, D.M., Gunawan, T.S., Kartiwi, M. and Ahmad, R. 2021. Real-Time Voice-Controlled Game Interaction using Convolutional Neural Networks. In: *2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2021*. Institute of Electrical and Electronics Engineers Inc., pp. 76–81. doi: 10.1109/ICSIMA50015.2021.9526318.

Yan, Y., Simons, S.O., van Bemmel, L., Reinders, L.G., Franssen, F.M.E. and Urovi, V. 2025. Optimizing MFCC parameters for the automatic detection of respiratory diseases. *Applied Acoustics* 228. doi: 10.1016/j.apacoust.2024.110299.

Zaman, K., Li, K., Sah, M., Direkoglu, C., Okada, S. and Unoki, M. 2025. Transformers and audio detection tasks: An overview. *Digital Signal Processing: A Review Journal* 158. doi: 10.1016/j.dsp.2024.104956.