INOTEK, Vol. 10
ISSN: 2580-3336 (Print) / 2549-7952 (Online)
Url: https://proceeding.unpkediri.ac.id/index.php/inotek/

Analisis Performa Wav2Vec2.0 pada Transkripsi Nama Barang dalam Kondisi Audio Bersih dan Bising

[1*] Dhimas Taka Desfrika Perdana, [2]Daniel Swanjaya, [3]Resty Wulanningrum
[1,2,3] Teknik Informatika, Universitas Nusantara PGRI Kediri
E-mail: [*1]dhimastak@gmail.com, [2]daniel@unpkediri.ac.id, [3]restyw@unpkdr.ac.id
Penulis Korespondens : Dhimas Taka Desfrika Perdana

Abstrak— *Automatic Speech Recognition* (ASR) merupakan teknologi yang memungkinkan suara diubah menjadi teks secara otomatis. Penelitian ini bertujuan menganalisis performa model Wav2Vec2.0 pada transkripsi nama barang berbahasa Indonesia dalam kondisi audio bersih dan bising. Dataset terdiri atas 400 rekaman audio yang dikumpulkan dari empat pembicara dengan kosakata terbatas berupa 20 nama barang. Evaluasi dilakukan menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER) pada kondisi audio bersih, *babble noise*, *crowd noise*, dan *vehicle noise*. Hasil pengujian menunjukkan bahwa kondisi audio bersih menghasilkan performa terbaik dengan nilai WER sebesar 2,08% dan CER sebesar 0,35%, sedangkan *crowd noise* menghasilkan tingkat kesalahan tertinggi dengan WER sebesar 22,92% dan CER sebesar 9,44%. Hasil penelitian menunjukkan bahwa peningkatan tingkat kebisingan berpengaruh terhadap penurunan performa transkripsi model Wav2Vec2.0 pada tugas pengenalan kosakata terbatas.
Kata Kunci— kebisingan, pengenalan suara otomatis, transkripsi nama barang, Wav2Vec2.0

Abstract— *Automatic Speech Recognition (ASR) is a technology that automatically converts speech into text. This study aims to analyze the performance of the Wav2Vec2.0 model for Indonesian product-name transcription under clean and noisy audio conditions. The dataset consists of 400 audio recordings collected from four speakers covering a limited vocabulary of 20 product names. Performance was evaluated using Word Error Rate (WER) and Character Error Rate (CER) under clean audio, babble noise, crowd noise, and vehicle noise conditions. The results show that clean audio achieved the best performance with a WER of 2.08% and a CER of 0.35%, while crowd noise produced the highest error rates with a WER of 22.92% and a CER of 9.44%. These findings indicate that increasing noise levels reduce the transcription performance of the Wav2Vec2.0 model in limited-vocabulary speech recognition tasks.*
Keywords— *Noise, Automatic Speech Recognition, Product Name Transcription, Wav2Vec2.0*

I. PENDAHULUAN

Proses pencatatan dan identifikasi barang merupakan kegiatan yang umum dilakukan pada berbagai sektor, seperti perdagangan, pergudangan, dan manajemen inventaris [1]. Namun, proses tersebut masih banyak dilakukan secara manual sehingga memerlukan waktu dan berpotensi menimbulkan kesalahan pencatatan[2]. Salah satu alternatif yang dapat digunakan untuk meningkatkan efisiensi proses tersebut adalah pemanfaatan teknologi pengenalan suara

yang memungkinkan informasi yang diucapkan pengguna dikonversi secara otomatis menjadi teks. Teknologi ini dikenal sebagai *Automatic Speech Recognition* (ASR) dan telah banyak diterapkan pada berbagai bidang, seperti asisten virtual, layanan pelanggan, sistem transkripsi otomatis, serta berbagai aplikasi berbasis suara lainnya [3]. Seiring meningkatnya kebutuhan interaksi manusia dan komputer secara alami, akurasi sistem pengenalan suara menjadi aspek penting dalam mendukung keberhasilan implementasi teknologi tersebut [4].

Salah satu model ASR yang banyak digunakan saat ini adalah Wav2Vec2.0, yang memanfaatkan pendekatan *self-supervised learning* untuk mempelajari representasi suara langsung dari sinyal audio mentah [5]. Berbagai penelitian menunjukkan bahwa model ini mampu menghasilkan performa pengenalan suara yang baik dengan kebutuhan data berlabel yang relatif lebih sedikit [6]. Meskipun demikian, performa Wav2Vec2.0 dapat menurun ketika audio terkontaminasi oleh kebisingan lingkungan yang sering dijumpai pada penggunaan di dunia nyata [7].

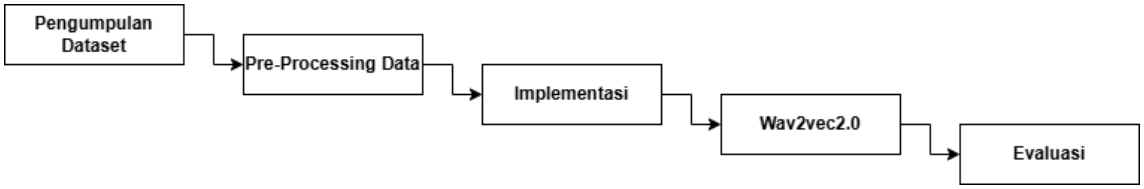
Beberapa penelitian terdahulu telah menunjukkan bahwa model Wav2Vec2.0 memiliki performa yang baik dalam tugas *Automatic Speech Recognition* (ASR) dan dapat ditingkatkan melalui berbagai pendekatan, seperti *preprocessing* data, *fine-tuning*, maupun teknik augmentasi audio [8]. Selain itu, penelitian lain menunjukkan bahwa keberadaan noise dapat menurunkan kualitas transkripsi sehingga diperlukan strategi tertentu untuk meningkatkan ketahanan model terhadap jenis kebisingan yang beragam [9]. Berbagai metode seperti *spectral masking*, augmentasi data bising, dan optimalisasi proses pelatihan telah dilaporkan mampu meningkatkan performa sistem ASR pada lingkungan yang tidak ideal [7], [10]. Meskipun demikian, penelitian yang membahas dampak kebisingan terhadap hasil transkripsi model Wav2Vec2.0 pada kosakata terbatas masih relatif terbatas. Oleh karena itu, penelitian ini difokuskan pada tugas pengenalan nama barang berbahasa Indonesia yang merepresentasikan skenario *limited vocabulary recognition*. Selain kondisi audio bersih, evaluasi juga dilakukan pada beberapa jenis kebisingan yang umum dijumpai pada penggunaan sehari-hari untuk mengetahui pengaruhnya terhadap performa transkripsi.

Berdasarkan tinjauan literatur tersebut, penelitian ini bertujuan untuk mengevaluasi kinerja model Wav2Vec2.0 pada kondisi audio bersih dan bising menggunakan kosakata terbatas berupa nama barang berbahasa Indonesia. Pengujian dilakukan pada beberapa jenis noise yang umum dijumpai, yaitu *babble noise*, *crowd noise*, dan *vehicle noise* [9], [11]. Evaluasi performa dilakukan menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER) untuk mengukur tingkat kesalahan transkripsi yang dihasilkan model [12]. Hasil penelitian diharapkan dapat memberikan gambaran mengenai pengaruh berbagai kondisi kebisingan terhadap performa Wav2Vec2.0 serta menjadi referensi bagi pengembangan sistem ASR berbasis kosakata terbatas dalam bahasa Indonesia.

II. METODE

Penelitian ini menggunakan metode eksperimen untuk menganalisis performa model Wav2Vec2.0 dalam melakukan transkripsi nama barang pada kondisi audio bersih dan bising.

Tahapan penelitian dilakukan secara sistematis mulai dari pengumpulan dataset, *preprocessing* data audio dan transkripsi, proses *fine-tuning* model Wav2Vec2.0, hingga pengujian dan evaluasi performa model. Pada tahap pengujian, model dievaluasi menggunakan data audio bersih serta data audio yang telah ditambahkan beberapa jenis noise, yaitu babble noise, crowd noise, dan *vehicle noise*. Kinerja model kemudian diukur menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER) untuk mengetahui tingkat kesalahan transkripsi yang dihasilkan. Alur penelitian yang digunakan pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Data penelitian diperoleh melalui perekaman suara empat pembicara yang terdiri atas dua laki-laki dan dua perempuan menggunakan gawai masing-masing pembicara pada lingkungan yang relatif tenang. Dataset terdiri dari 20 nama barang seperti pada Tabel 1 yang diucapkan oleh setiap pembicara dengan lima variasi perekaman, yaitu pengucapan normal, pengucapan lebih cepat, pengucapan lebih lambat, intonasi seperti kalimat tanya, dan pengucapan dengan jarak mikrofon yang lebih jauh dari mulut.. Dengan empat pembicara, 20 nama barang, dan lima variasi perekaman untuk setiap nama barang, diperoleh total 400 data audio yang digunakan dalam penelitian.

Tabel 1. Nama Barang

No	***Nama Barang***
1	Air Mineral
2	Antiseptik
3	Beras
4	Deterjen
5	Garam
6	Gula
7	Keramik
8	Kopi
9	Margarin
10	Mayones
11	Minyak Goreng
12	Mozzarella
13	Sabun Mandi
14	Sampo
15	Sereal

16	Susu
17	Teh
18	Telur
19	Tepung Terigu
20	Tisu

B. *Preprocessing*

Tahap *preprocessing* dilakukan untuk menyeragamkan data audio dan transkripsi sebelum digunakan pada proses pelatihan model Wav2Vec2.0. Seluruh rekaman suara yang semula berformat M4A dikonversi ke format WAV agar sesuai dengan kebutuhan pengolahan audio. Selanjutnya, audio diseragamkan pada *sampling rate* 16 kHz sesuai spesifikasi masukan model Wav2Vec2.0 [13]. Pada tahap transkripsi, seluruh teks dinormalisasi dengan mengubah huruf menjadi huruf kecil (*lowercase*) dan menghilangkan spasi yang tidak diperlukan. Setelah proses *preprocessing* selesai, data audio dan transkripsi disusun dalam bentuk pasangan data yang disimpan dalam berkas CSV untuk digunakan pada proses pelatihan dan pengujian model.

C. *Fine-tuning* Wav2Vec2.0

Proses *fine-tuning* dilakukan menggunakan model Wav2Vec2.0 Base yang tersedia pada pustaka Hugging Face Transformers [13]. Dataset yang telah melalui tahap *preprocessing* kemudian dibagi menjadi data pelatihan, validasi, dan pengujian, di mana data pelatihan digunakan untuk menyesuaikan parameter model, sedangkan data validasi digunakan untuk memantau proses pelatihan dan mengurangi risiko *overfitting*. Pada tahap ini, model mempelajari hubungan antara sinyal suara dan transkripsi teks yang sesuai sehingga mampu menyesuaikan representasi yang telah dipelajari sebelumnya dengan karakteristik dataset penelitian. Model hasil *fine-tuning* selanjutnya digunakan untuk melakukan transkripsi pada data uji dan dievaluasi pada kondisi audio bersih maupun bising.

D. Evaluasi

Pengujian dilakukan untuk mengevaluasi performa model Wav2Vec2.0 dalam mentranskripsikan nama barang pada kondisi audio bersih dan bising. Model yang telah melalui proses *fine-tuning* diuji menggunakan data uji yang tidak digunakan selama proses pelatihan. Pengujian dilakukan pada empat kondisi audio, yaitu audio bersih (*clean audio*), audio dengan *babble noise* (percakapan beberapa orang secara bersamaan), audio dengan *crowd noise* (suara keramaian atau aktivitas banyak orang dalam suatu lingkungan), serta audio dengan *vehicle noise* (suara lalu lintas seperti kendaraan bermotor dan kebisingan jalan raya). Data audio bising diperoleh dengan menambahkan sinyal *noise* pada audio bersih sehingga setiap jenis kebisingan dapat diterapkan secara konsisten pada seluruh data uji.

Evaluasi performa dilakukan menggunakan dua metrik, yaitu :

1. *Word Error Rate* (WER)

   WER digunakan untuk mengukur tingkat kesalahan transkripsi pada tingkat kata dengan membandingkan hasil transkripsi model terhadap transkripsi referensi. Semakin kecil nilai WER, semakin baik performa model dalam mengenali kata yang diucapkan.

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

Keterangan :
S= jumlah kata atau nama produk yang salah dikenali (substitution)
D= jumlah kata atau nama produk yang tidak dikenali (deletion)
I= jumlah kata atau nama produk yang muncul padahal tidak disebutkan (insertion)
N= jumlah kata atau nama produk pada acuan

2. *Character Error Rate* (CER)

CER digunakan untuk mengukur tingkat kesalahan transkripsi pada tingkat karakter. Metrik ini memberikan evaluasi yang lebih rinci karena memperhitungkan kesalahan pada setiap huruf yang terdapat dalam hasil transkripsi.

$$CER = \frac{S + D + I}{N} \times 100\% \quad (2)$$

Keterangan:
S= jumlah karakter yang salah dikenali (substitution)
D= jumlah karakter yang tidak dikenali (deletion)
I= jumlah karakter tambahan yang muncul padahal tidak diucapkan (insertion)
N= jumlah karakter pada teks acuan

III. HASIL DAN PEMBAHASAN

A. Hasil *Fine-tuning*

Proses *fine-tuning* dilakukan menggunakan dataset yang terdiri dari 400 rekaman audio hasil perekaman empat pembicara yang terdiri atas dua laki-laki dan dua perempuan. Dataset mencakup 20 nama barang dengan lima variasi perekaman untuk setiap nama barang. Sebelum pelatihan, data dibagi menjadi data pelatihan, validasi, dan pengujian dengan rasio 80:10:10, sehingga diperoleh 320 data pelatihan, 40 data validasi, dan 40 data pengujian. Data pelatihan digunakan untuk menyesuaikan parameter model Wav2Vec2.0, sedangkan data validasi digunakan untuk memantau proses pelatihan sebelum model dievaluasi menggunakan data pengujian.

Epoch	Training Loss	Validation Loss
1	11.707764	0.288254
2	3.121946	0.059840
3	2.030224	0.038268
4	1.029601	0.012644
5	0.921386	0.002954
6	0.856701	0.006041

Gambar 2 *Training loss* dan *Validation loss*

Pada Gambar 2 terlihat bahwa *training loss* menurun dari 11,7078 pada epoch pertama menjadi 0,8567 pada epoch keenam, sedangkan *validation loss* menurun dari 0,2883 menjadi 0,0060. Penurunan kedua nilai loss tersebut menunjukkan bahwa proses *fine-tuning* berjalan dengan baik dan model berhasil mempelajari pola data secara efektif. Meskipun terjadi sedikit peningkatan *validation loss* pada epoch terakhir, nilainya tetap rendah sehingga model dinilai telah mencapai kondisi yang stabil untuk digunakan pada tahap pengujian.

B. Pengujian pada Audio Bersih

Sebagai acuan awal, model Wav2Vec2.0 terlebih dahulu dievaluasi menggunakan data audio bersih yang tidak mengandung gangguan kebisingan. Pengujian ini bertujuan untuk mengetahui kemampuan dasar model dalam melakukan transkripsi sebelum diberikan berbagai kondisi noise.

Audio diuji pada 40 data uji , sebanyak 39 data berhasil ditranskripsikan dengan benar dan hanya 1 data yang mengalami kesalahan seperti pada Tabel 2. Nilai WER yang diperoleh sebesar 2,08 % dan CER sebesar 0,35%. Ini menunjukkan bahwa model mampu mengenali sebagian besar kata dan karakter dengan baik.

Tabel 2. Kesalahan pada Audio Bersih

Ground Truth	***Prediksi***
tepung	tebung

C. Pengujian pada Babble Noise

Setelah memperoleh hasil pada kondisi audio bersih, model selanjutnya diuji menggunakan audio yang telah ditambahkan *babble noise*. Jenis *noise* ini merepresentasikan percakapan beberapa orang yang berlangsung secara bersamaan sehingga dapat mengganggu proses pengenalan ucapan.

Dari 40 data yang diuji terdapat 4 data yang mengalami kesalahan transkripsi seperti pada Tabel 3 dan mendapat WER sebesar 12,50% dan CER 3,15%. Hasil ini menunjukkan bahwa babble noise dapat mengganggu proses pengenalan ucapan sehingga model mengalami kesulitan dalam mengenali beberapa kata secara tepat.

Tabel 3. Kesalahan pada Babble Noise

Ground Truth	***Prediksi***
keramik	geramik
margarin	margaril
minyak goreng	monyekgorang
tepung terigu	tebun tirigu

D. Pengujian pada Crowd Noise

Untuk mensimulasikan kondisi lingkungan yang ramai, pengujian berikutnya dilakukan menggunakan audio yang mengandung *crowd noise*. Kebisingan ini terdiri dari berbagai suara aktivitas manusia yang terjadi secara bersamaan dan sering dijumpai pada area publik yang padat.

Hasil pengujian pada crowd noise menunjukkan penurunan yang signifikan dibanding dengan pengujian lainnya. Dari 40 data pengujian, ada 9 data yang mengalami kesalahan transkripsi terlihat pada Tabel 4 terutama pada kata yang susah diucapkan seperti antiseptik, minyak goreng dan juga kata berulang seperti susu. Nilai WER yang didapatkan sebesar 22,92% dan CER 9,44%. Hal ini menunjukkan bahwa crowd noise memberikan gangguan yang lebih kompleks, sehingga menyebabkan model lebih sulit membedakan karakteristik ucapan.

Tabel 4. Kesalahan pada Crowd Noise

Ground Truth	***Prediksi***
antiseptik	samiseptik
antiseptik	antisep
keramik	geramik
margarin	margari
minyak goreng	mamnyak gula
mozzarella	mozzarel
susu	susa
susu	skesu
tepung terigu	tebun atiragol

E. Pengujian pada Vehicle Noise

Selain gangguan yang berasal dari aktivitas manusia, model juga dievaluasi pada kondisi audio yang mengandung *vehicle noise*. Jenis *noise* ini merepresentasikan suara kendaraan bermotor yang umum ditemukan pada lingkungan perkotaan dan berpotensi memengaruhi kualitas transkripsi yang dihasilkan model.

Dari 40 data yang diuji, disini terdapat 5 kesalahan transkripsi seperti pada Tabel 5, dengan nilai WER sebesar 12,50% dan CER 5,24%. Hasil ini menunjukkan bahwa vehicle noise dapat mengganggu pengenalan ucapan, terutama pada kata-kata sulit seperti antiseptik dan mozzarella.

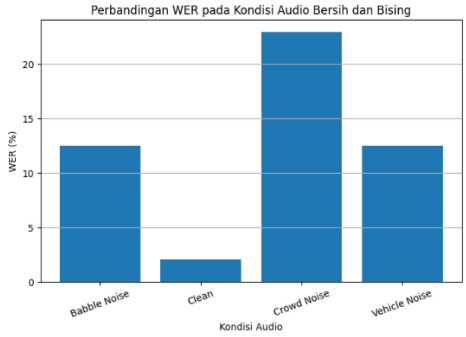
Tabel 5. Keslahan pada Vehicle Noise

Ground Truth	***Prediksi***

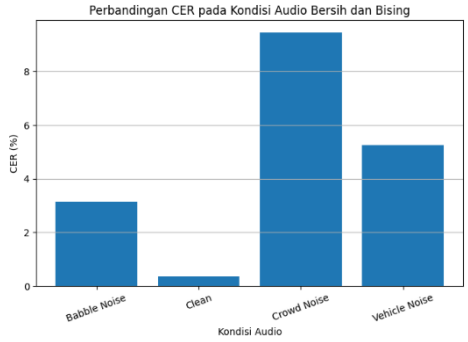
antiseptik	atusepti
garam	taram
keramik	geramik
minyak goreng	monekgola
mozzarella	mozzarel

F. Perbandingan Pengujian

Untuk memudahkan pembandingan mengenai performa model di berbagai kondisi, hasil evaluasi metrik WER dan CER pada setiap skenario pengujian divisualisasikan dalam bentuk grafik.



Gambar 3. Perbandingan WER



Gambar 4. Perbandingan CER

Gambar 3 dan Gambar 4 menunjukkan perbandingan nilai WER dan CER pada seluruh kondisi pengujian. Terlihat bahwa kondisi audio bersih menghasilkan performa terbaik dengan nilai WER sebesar 2,08% dan CER sebesar 0,35%. Setelah ditambahkan noise, nilai WER dan CER mengalami peningkatan pada seluruh skenario pengujian, yang menunjukkan adanya penurunan kemampuan model dalam melakukan transkripsi. Crowd noise menghasilkan tingkat kesalahan tertinggi dengan nilai WER sebesar 22,92% dan CER sebesar 9,44%, sedangkan babble noise dan vehicle noise menghasilkan nilai WER yang sama sebesar 12,50%, namun vehicle noise memiliki nilai CER yang lebih tinggi yaitu 5,24% dibandingkan babble noise sebesar 3,15%. Hasil ini menunjukkan bahwa keberadaan kebisingan berpengaruh terhadap performa Wav2Vec2.0, di mana crowd noise menjadi jenis kebisingan yang paling mengganggu proses transkripsi. Secara keseluruhan, semakin kompleks sumber suara pada lingkungan pengujian, semakin tinggi tingkat kesalahan transkripsi yang dihasilkan model.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa performa model Wav2Vec2.0 dalam mentranskripsikan nama barang berbahasa Indonesia dipengaruhi oleh kondisi audio yang digunakan. Pada kondisi tanpa gangguan kebisingan, model mampu melakukan transkripsi dengan baik, sedangkan keberadaan kebisingan menyebabkan penurunan kemampuan pengenalan ucapan dengan tingkat pengaruh yang berbeda pada setiap jenis *noise*. Temuan ini menunjukkan bahwa kondisi lingkungan audio merupakan faktor penting yang perlu diperhatikan dalam penerapan sistem *Automatic Speech Recognition* (ASR). Dengan demikian, penelitian ini memberikan kontribusi dalam memahami karakteristik performa Wav2Vec2.0 pada tugas transkripsi kosakata terbatas

serta dapat menjadi dasar bagi pengembangan sistem pengenalan suara yang lebih andal pada berbagai kondisi penggunaan.

UCAPAN TERIMAKASIH

Penulis menyampaikan terima kasih kepada Program Studi Teknik Informatika Universitas Nusantara PGRI Kediri atas dukungan yang diberikan selama pelaksanaan penelitian. Penulis juga mengucapkan terima kasih kepada dosen pembimbing atas bimbingan, arahan, dan masukan yang sangat membantu dalam proses penelitian serta penyusunan artikel ini. Ucapan terima kasih turut diberikan kepada seluruh pihak yang telah membantu, baik secara langsung maupun tidak langsung, sehingga penelitian dan penyusunan artikel ini dapat terlaksana dengan baik.

DAFTAR PUSTAKA

[1] R. Setiawan, N. P. Sugihartanti, and M. I. Ibadurrahman, "Sistem Manajemen Gudang Bebasis Web dengan Teknologi Barcode Scanner pada Industri Manufaktur: Sebuah Kajian Literatur," *Integrasi: Jurnal Ilmiah Teknik Industri*, vol. 9, no. 2, pp. 124–135, 2024, doi: https://doi.org/10.32502/integrasi.v9i2.181.

[2] P. Q. Izzatiy and S. Sundari, "IMPLEMENTASI ACCURATE DALAM SISTEM PENCATATAN KAS KECIL DAN HUTANG PADA PERUSAHAAN MANUFAKTUR DI SURABAYA," *Jurnal Ekonomi Bisnis Manajemen dan Akuntansi (JEBISMA)*, vol. 3, no. 3, 2026, doi: https://doi.org/10.70197/jebisma.v3i3.198.

[3] H. Azis, E. P. Indreswari, and R. Wisudawanto, "PEMANFAATAN TEKNOLOGI *AUTOMATIC SPEECH RECOGNITION* DALAM MENCIPTAKAN PEMBELAJARAN INKLUSIF: IMPLEMENTASI, EFEKTIFITAS DAN TANTANGAN," *GANESHA: Jurnal Pengabdian Masyarakat*, vol. 5, no. 1, pp. 27–33, 2025, doi: https://doi.org/10.36728/ganesha.v5i1.4171.

[4] H. Wijaya, "Teknologi Pengenalan Suara tentang Metode, Bahasa dan Tantangan: Systematic Literature Review," *bit-Tech*, vol. 7, no. 2, pp. 533–544, 2024, doi: https://doi.org/10.32877/bt.v7i2.1888.

[5] D. Ferdiansyah and C. S. K. Aditya, "Implementasi *Automatic Speech Recognition* Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper," *Jurnal Teknik Elektro Dan Komputer TRIAC*, vol. 11, no. 1, pp. 11–16, 2024, doi: https://doi.org/10.21107/triac.v11i1.24332.

[6] Abdullah Azzam, Ichsan Taufik, and Aldy Rialdy Atmadja, "Evaluating End-to-End ASR for Qur'an Recitation Using Whispers in Low Resource Settings," *Bulletin of Computer Science Research*, vol. 5, no. 4, pp. 778–787, Jun. 2025, doi: 10.47065/bulletincsr.v5i4.561.

[7] A. NOERCHOLIS, T. DWIANDINI, and F. S. MUKTI, "Optimasi Teknologi WAV2Vec 2.0 menggunakan Spectral Masking untuk meningkatkan Kualitas Transkripsi Teks Video bagi Tuna Rungu," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 12, no. 4, p. 877, 2024, doi: https://doi.org/10.26760/elkomika.v12i4.877.

[8] J. Cai, Y. Song, J. Wu, and X. Chen, "Voice Disorder Classification Using Wav2vec 2.0 Feature Extraction," *Journal of Voice*, 2024, doi: https://doi.org/10.1016/j.jvoice.2024.09.002.

[9] J. C. Duarte and S. Colcher, "Noise-Robust *Automatic Speech Recognition*: A Case Study for Communication Interference," *Journal on Interactive Systems*, vol. 15, no. 1, pp. 670–681, Jul. 2024, doi: 10.5753/jis.2024.4267.

[10] M. N. Althoff, A. Affandy, A. Luthfiarta, M. W. B. D. Satya, and H. Basiron, "Leveraging Label *Preprocessing* for Effective End-to-End Indonesian *Automatic Speech Recognition*," *Sinkron : jurnal dan penelitian teknik informatika*, vol. 9, no. 1, pp. 55–64, Jan. 2025, doi: 10.33395/sinkron.v9i1.14257.

[11] C. Mena, A. L. Padilla-Ortiz, and F. Orduña-Bustamante, "*Automatic Speech Recognition* in the presence of babble noise and reverberation compared to human speech intelligibility in Spanish," *Comput. Speech Lang.*, vol. 95, p. 101856, 2026, doi: 10.1016/j.csl.2025.101856.

[12] T. D. K, J. James, D. P. Gopinath, and M. A. K, "Advocating *Character Error Rate* for Multilingual ASR Evaluation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds., Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4941–4950. doi: 10.18653/v1/2025.findings-naacl.277.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for *self-supervised learning* of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020, doi: 10.48550/arXiv.2006.11477.