ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika

ISSN(p): 2338-8323 | ISSN(e): 2459-9638 | Vol. 12 | No. 4 | Halaman 877 - 891
DOI : http://dx.doi.org/10.26760/elkomika.v12i4.877 Oktober 2024

Optimasi Teknologi WAV2Vec 2.0 menggunakan *Spectral Masking* untuk meningkatkan Kualitas Transkripsi Teks Video bagi Tuna Rungu

ACHMAD NOERCHOLIS, TITANIA DWIANDINI, FRANSISKA SISILIA MUKTI

Institut Teknologi dan Bisnis Asia Malang, Indonesia
Email: anoercholis@asia.ac.id

Received 23 Agustus 2024 **|** *Revised* 30 September 2024 **|** *Accepted* 3 Oktober 2024

ABSTRAK

Teknologi Automatic Speech Recognition (ASR) telah berkembang pesat sebagai alat untuk meningkatkan aksesibilitas informasi bagi penyandang tuna rungu, terutama melalui video. WAV2Vec 2.0, salah satu teknologi ASR unggulan, efektif dalam transkripsi teks, namun kinerjanya menurun saat menghadapi noise. Penelitian ini bertujuan mengoptimalkan WAV2Vec 2.0 dengan menerapkan Spectral Masking untuk mengurangi noise tanpa mengorbankan kejelasan sinyal utama. Evaluasi dilakukan pada tiga jenis video: podcast, video dengan background noise, dan video dengan background music. Hasil menunjukkan penurunan Word Error Rate (WER) yang signifikan, sebesar 78.06% pada podcast dan 53.85% pada video dengan background noise. Hasil penelitian menunjukkan bahwa Spectral Masking efektif dalam meningkatkan akurasi transkripsi, menawarkan solusi inovatif untuk aksesibilitas tuna rungu dalam kondisi audio yang kompleks.

***Kata kunci**: noise reduction, spectral masking, tuna rungu, WAV2Vec 2.0*

ABSTRACT

Automatic Speech Recognition (ASR) technology has rapidly evolved as a tool to enhance information accessibility for the hearing impaired, particularly through video content. WAV2Vec 2.0, a leading ASR technology, is effective in text transcription, but its performance degrades in the presence of noise. This study aims to optimize WAV2Vec 2.0 by applying Spectral Masking to reduce noise without compromising the clarity of the main signal. The evaluation was conducted on three types of videos: podcasts, videos with background noise, and videos with background music. The results show a significant reduction in Word Error Rate (WER), with a 78.06% decrease in podcasts and a 53.85% decrease in videos with background noise. These findings demonstrate that Spectral Masking effectively enhances transcription accuracy, offering an innovative solution for improving accessibility for the hearing impaired in complex audio conditions.

***Keywords**: noise reduction, spectral masking, tuna rungu, WAV2Vec 2.0*

1. PENDAHULUAN

Disabilitas merujuk pada kondisi yang mempengaruhi fungsi fisik atau mental seseorang, yang dapat membatasi kemampuan mereka untuk melakukan aktivitas pada umumnya. Secara general, disabilitas yang paling banyak ditemui adalah gangguan penglihatan, pendengaran, dan mobilitas, yang umumnya disebabkan faktor genetik, penyakit, cedera, atau kondisi lingkungan. Badan Pusat Statistik Indonesia merilis hasil survey terhadap prevalensi penyandang disabilitas di Indonesia tahun 2022 didominasi oleh gangguan pendengaran, yang mencapai 38% dari total penyandang disabilitas di Indonesia **(Direktorat Analisis dan Pengembangan Statistik, 2022)**.

Pada dasarnya, penyandang disabilitas memiliki hak yang setara dalam berbagai aspek kehidupan. Seperti halnya individu lainnya, mereka berhak atas pemenuhan hak-hak dasar serta memiliki peluang untuk berkontribusi secara inklusif dalam pembangunan dan kemajuan masyarakat, tanpa memandang keterbatasan fisik, sensorik, kognitif, atau emosional. Oleh karena itu, penyandang disabilitas harus memiliki akses yang setara terhadap pendidikan, pekerjaan, layanan kesehatan, transportasi, informasi, dan kesempatan untuk berpartisipasi dalam kehidupan sosial. Hal ini secara khusus telah menjadi komitmen pemerintah dalam membangun masyarakat inklusif, melalui adanya UU No. 8 Tahun 2016 tentang Penyandang Disabilitas (PD) **(Undang - Undang Republik Indonesia Nomor 8 Tahun 2016 Tentang Penyandang Disabilitas, 2016)**.

Aksesibilitas informasi telah menjadi salah satu elemen penting dalam menciptakan lingkungan yang inklusif bagi semua individu, termasuk bagi komunitas disabilitas tuna rungu. Dalam beberapa dekade terakhir, teknologi informasi dan komunikasi telah berkembang pesat, dan video telah menjadi salah satu medium utama untuk berbagi pengetahuan, informasi, dan hiburan. Namun, komunitas tunarungu sering kali menghadapi tantangan signifikan dalam mengakses konten video, terutama karena tidak tersedianya teks atau *subtitle* yang akurat. Hal ini menjadi lebih kompleks ketika video mengandung banyak *noise* latar belakang atau suara yang tidak jelas, yang dapat menghambat kemampuan teknologi pengenalan ucapan untuk menghasilkan transkripsi teks yang berkualitas.

Teknologi pengenalan ucapan otomatis (*Automatic Speech Recognition*/ASR) telah berkembang secara signifikan dan menjadi alat yang potensial untuk mengatasi tantangan ini **(Tirta, dkk, 2022)**. Salah satu teknologi terkemuka dalam bidang ini adalah WAV2Vec 2.0, sebuah model berbasis *self-supervised learning*, yang memiliki kemampuan dalam menghasilkan representasi audio yang kuat dari sinyal suara dan dapat dimanfaatkan untuk berbagai aplikasi terkait pemrosesan bahasa alami. WAV2Vec 2.0 dirancang untuk bekerja dengan baik dalam kondisi akustik yang berbeda-beda dan telah digunakan secara luas dalam berbagai aplikasi, termasuk layanan transkripsi dan asisten virtual **(Ferdiansyah & Aditya, 2024)**. Lebih lanjut, model ini dapat dengan efisien mempelajari fitur yang berguna dari data suara tanpa label **(Gondi, 2022)(Tak, dkk, 2022)**, dan menunjukkan hasil yang menjanjikan untuk tugas pengenalan ucapan di berbagai bahasa **(Yi, dkk, 2020)**.

Sebagai salah satu model yang dianggap efektif dalam pengenalan ucapan, untuk pengenalan ucapan, WAV2Vec 2.0 memiliki beberapa keterbatasan, terutama saat menghadapi kondisi audio dengan tingkat kebisingan yang tinggi. Kinerja model ini cenderung menurun secara signifikan ketika digunakan dalam lingkungan dengan banyak *noise* atau gangguan audio lainnya, yang menyebabkan peningkatan tingkat kesalahan kata (*Word Error Rate*/WER)**(Gondi, 2022)**. Selain itu, dalam aplikasi *real-time,* terutama pada perangkat dengan sumber daya terbatas, WAV2Vec 2.0 sering kali menghadapi tantangan dalam mempertahankan performa yang baik karena kebutuhan komputasi yang tinggi dan kesulitan

dalam memproses input audio berisik dengan cepat dan efisien **(Ragano, dkk, 2022)**. Model ini juga sering memerlukan *fine-tuning* khusus ketika diterapkan pada tugas-tugas yang melibatkan banyak noise atau jenis audio yang berbeda seperti musik atau suara lingkungan yang kompleks **(Kang, dkk, 2024)**. Oleh karena itu, pengembangan lebih lanjut diperlukan untuk meningkatkan ketahanan WAV2Vec 2.0 terhadap gangguan audio dalam lingkungan dunia nyata.

Berbagai penelitian telah dilakukan dengan tujuan untuk mengembangkan teknologi ASR dalam menghasilkan hasil transkripsi yang lebih baik. Pendekatan menggunakan *Hidden Markov Model* yang dilengkapi fitur *Mel Frequency Cepstral Coefficients* digunakan dalam penelitian **(Tirta, dkk, 2022)** berhasil menghasilkan akurasi yang lebih baik pada video lagu (mencapai 81%). Pemanfaatan *deep learning* dilakukan oleh **(Andra & Usagawa, 2020)** untuk mentranskripsikan percakapan simultan dalam bahasa Indonesia menggunakan *Pitch-Aware Gain-Based Speech Separation* untuk membedakan suara antar pembicara dan *Recurrent Neural Network* untuk menghasilkan transkripsi pembicaraan yang lebih baik pada banyak pembicara. Model Wav2Vec 2.0 diaplikasikan untuk mengidentifikasi berbagai bahasa lokal, seperti yang dilakukan oleh **(Yi, dkk, 2020)** dan **(Kozhirbayev, 2023)**, maupun bahasa anak-anak **(Jain, dkk, 2023) (Getman, dkk, 2023)**. Pengembangan model Wav2Vec 2.0 juga dilakukan melalui proses *fine-tuning* untuk meningkatkan akurasi **(Cryssiover & Zahra, 2024) (Chen & Rudnicky, 2023)**, maupun untuk mengakomodir kebutuhan disabilitas **(Javanmardi, dkk, 2024)(Smolik, dkk, 2024)**. Penelitian-penelitian ini menunjukkan potensi besar model WAV2Vec 2.0 ini dalam menangani tugas-tugas pengenalan suara.

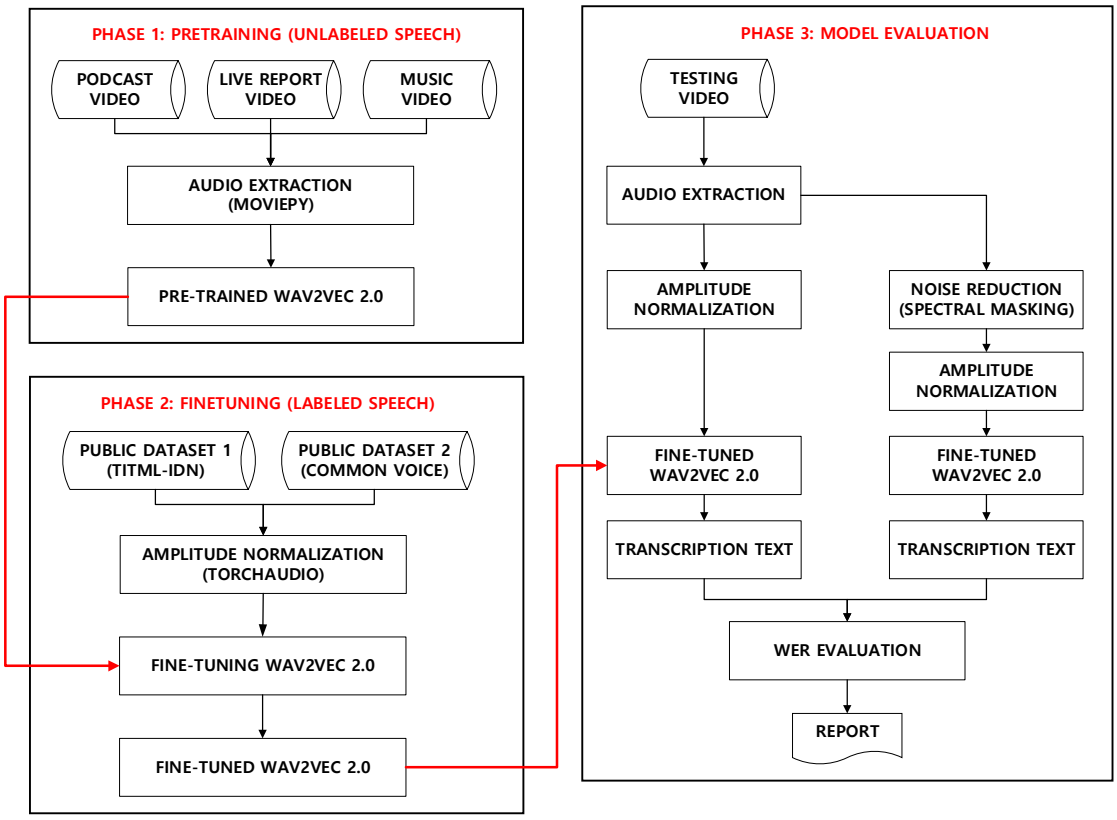
Walaupun berbagai penelitian telah berhasil mengembangkan model ASR dengan akurasi yang lebih tinggi, banyak dari model ini masih menghadapi keterbatasan signifikan dalam menangani kondisi audio yang penuh dengan *noise*—sebuah tantangan utama dalam aplikasi di dunia nyata. Upaya untuk memanfaatkan teknologi ini demi meningkatkan aksesibilitas, terutama bagi penyandang disabilitas seperti tuna rungu, masih belum mencapai hasil yang memadai. Masalah ini terutama tampak ketika model digunakan dalam lingkungan berisik, yang secara drastis dapat mengurangi akurasi transkripsi.

WAV2Vec 2.0, yang telah diakui potensinya dalam berbagai tugas pengenalan suara, juga tidak terlepas dari tantangan ini. Kendala dalam menangani noise tetap menjadi hambatan besar yang menghalangi model ini untuk mencapai akurasi yang optimal, terutama dalam kondisi audio yang tidak ideal. Oleh karena itu, penelitian ini bertujuan untuk mengoptimalkan teknologi WAV2Vec 2.0 dengan fokus khusus pada peningkatan kemampuannya dalam mengatasi *noise* menggunakan pendekatan *Spectral Masking*. *Spectral Masking* adalah teknik *noise reduction* yang secara efektif menyaring *noise* dari sinyal audio, memungkinkan model untuk lebih akurat mengekstraksi fitur suara yang relevan. Dengan penerapan teknik ini, diharapkan kualitas transkripsi dapat meningkat secara signifikan, terutama dalam kondisi audio yang tidak ideal **(Ravanelli, dkk, 2021)**.

Dengan tujuan akhir untuk mendukung aksesibilitas informasi melalui video bagi tuna rungu, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam memungkinkan penyandang disabilitas untuk mengakses konten audio-visual dengan lebih baik dan lebih mandiri. Pendekatan inovatif ini tidak hanya menjawab celah yang ada dalam penelitian sebelumnya, tetapi juga berpotensi membawa dampak positif yang luas bagi teknologi inklusif di masa depan.

2. METODE

Penelitian ini mengadopsi metodologi eksperimental yang terstruktur untuk mengevaluasi efektivitas teknik *Spectral Masking* dalam mengoptimalkan performa teknologi WAV2Vec 2.0. Metodologi ini terdiri dari tiga fase utama, yaitu *pretraining*, *finetuning*, dan evaluasi model, sebagaimana dijelaskan secara rinci dalam Gambar 1. Setiap fase dirancang dengan pendekatan sistematis yang mencakup pengumpulan data audio dari berbagai sumber video, pelatihan model menggunakan dataset bahasa Indonesia, serta penerapan teknik pengurangan noise dalam proses transkripsi teks.



Gambar 1. Diagram Alir Penelitian

Dalam penelitian ini, data video yang mengandung *noise* atau latar belakang musik akan ditranskripsi menggunakan model WAV2Vec 2.0 dengan dan tanpa penerapan teknik *Spectral Masking*. Model WAV2Vec 2.0 menggunakan 2 tahapan dalam melakukan ekstraksi representasi suara dari file audio mentah dalam skenario pembelajaran mandiri (*self-supervised learning*) dan memanfaatkan representasi tersebut untuk tugas transkripsi teks otomatis (ASR) **(Jain, dkk, 2023)**.

WAV2Vec 2.0 dipilih karena kemampuannya untuk mencapai hasil terbaik ketika dilatih dengan sejumlah besar data suara yang tidak berlabel, dan di-finetuning pada data berlabel dalam jumlah kecil **(Baevski, dkk, 2020)**. Hal ini sangat ideal untuk tujuan penelitian ini, yaitu meningkatkan kualitas transkripsi teks video bagi komunitas tuna rungu, karena lebih mudah untuk mengumpulkan data suara yang tidak berlabel dalam jumlah besar daripada data yang berlabel dengan akurasi tinggi.

Sementara itu, upaya peningkatan hasil transkripi teks pada WAV2Vec 2.0 dilakukan dengan menggunakan teknik *Spectral Masking*, sebagai salah satu teknik yang efektif dalam mengurangi *noise* latar belakang dengan memanipulasi spektrum frekuensi audio untuk memperjelas komponen suara utama, seperti vocal **(Pascual, dkk, 2017)**. Dengan menekan frekuensi yang dianggap sebagai *noise*, teknik ini memungkinkan model ASR seperti WAV2Vec 2.0 untuk fokus pada sinyal vokal yang lebih bersih, yang pada akhirnya meningkatkan akurasi transkripsi **(Wang & Chen, 2018)**. Ini sangat penting dalam konteks transkripsi bagi komunitas tuna rungu, di mana kejelasan teks yang dihasilkan sangat berpengaruh terhadap aksesibilitas informasi.

Selanjutnya, kinerja hasil transkripsi dievaluasi melalui pengukuran *Word Error Rate* (WER), yang kemudian dianalisis secara mendalam untuk menilai efektivitas metode yang diusulkan. Melalui pendekatan ini, penelitian diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan akurasi transkripsi teks video, yang secara langsung berdampak pada peningkatan aksesibilitas informasi bagi komunitas tuna rungu di Indonesia.

2.1 Pre-Training (*Unlabelled Speech*)

Pada fase pertama penelitian ini, dilakukan *pre-training* model WAV2Vec 2.0 menggunakan data suara yang tidak memiliki label transkripsi teks. Fase *pre-training* ini krusial karena bertujuan untuk melatih model dalam mengenali pola-pola dasar suara secara umum sebelum dilanjutkan ke fase *fine-tuning* yang menggunakan data berlabel. Data yang digunakan untuk *pre-training* diambil dari berbagai jenis video yang tersedia di platform YouTube, dengan tujuan untuk mendapatkan variasi konten suara yang luas dan beragam. Jenis video yang dipilih mencakup tiga kategori utama:

1) *Podcast Video,* dipilih karena format diskusi atau monologhnya menghadirkan suara jernih dengan SNR tinggi (di atas 20 dB), memastikan energi frekuensi terkonsentrasi pada 200–3.000 Hz, yang ideal untuk melatih model dalam mendeteksi pola suara manusia tanpa gangguan.
2) *Live Report Video,* dipilih karena mengandung noise lingkungan seperti suara angin atau keramaian, memberikan SNR bervariasi antara 5–15 dB dan distribusi energi frekuensi pada 100–5.000 Hz. Ini memperkaya model dalam mengenali variasi suara dan menangani kondisi audio dinamis di dunia nyata.
3) *Music Video*, mengombinasikan instrumen dan vokal, menuntut model untuk memisahkan komponen audio dengan presisi. SNR yang rendah hingga sedang (di bawah 10 dB) dan energi yang tersebar dari 20 Hz hingga 20.000 Hz menantang model untuk menangani audio kompleks dan memastikan akurasi transkripsi.

Setiap kategori berisi 10 video dengan durasi percakapan antara 2 hingga 3 menit. Selanjutnya, data audio diekstraksi menggunakan *MoviePy*, sebuah library Python yang mampu memisahkan audio dari file video dengan efisien. Proses ekstraksi ini menghasilkan data suara mentah, yang kemudian digunakan dalam tahap *pre-training*. Pada tahap ini, model belum dilatih untuk menghasilkan transkripsi teks yang spesifik. Sebaliknya, model *pre-trained* WAV2Vec 2.0 dilatih untuk memahami representasi akustik dari audio dengan cara berikut:

1) *Feature Extraction*: Model WAV2Vec 2.0 pertama-tama mengekstrak fitur dari data audio mentah. Fitur-fitur ini merepresentasikan berbagai aspek dari sinyal audio, seperti frekuensi dan amplitudo.
2) *Contextual Representation*: Setelah fitur diekstraksi, model melatih dirinya untuk memprediksi representasi kontekstual dari audio tersebut. Ini berarti model mencoba memahami konteks dari segmen audio, tanpa benar-benar memahami teks yang diucapkan.

3) *Learning Acoustic Patterns*: Model dilatih untuk mengenali pola-pola akustik tanpa supervisi langsung, artinya tidak menggunakan label teks yang sesuai. Ini dilakukan melalui proses pelatihan yang melibatkan memprediksi sebagian dari data audio berdasarkan konteks sekitarnya.

Output pada tahap *pre-trained* ini adalah representasi vektor yang menyandikan informasi akustik dari audio. Dengan kata lain, WAV2Vec 2.0 tidak langsung menghasilkan teks, tetapi menghasilkan vektor representasi yang mencerminkan karakteristik suara dalam audio. Vektor-vektor ini kemudian digunakan pada tahap *fine-tuning* untuk melatih model agar mampu menghasilkan transkripsi teks dari audio.

2.2 *Fine-Tuning* (*Labelled Speech*)

Pada fase kedua penelitian ini, dilakukan proses *fine-tuning* untuk menyempurnakan model WAV2Vec 2.0 yang telah melalui tahap *pre-training*. Fase *fine-tuning* ini bertujuan untuk meningkatkan kemampuan model dalam menghasilkan transkripsi teks yang lebih akurat dengan memanfaatkan dataset suara berlabel.

Dalam proses ini, dua dataset publik yang relevan digunakan, yaitu TITML-IDN **(Tokyo Institute of Technology Multilingual Speech Corpus (TITML), 2008)** dan Common Voice **(Mozilla, 2017).** Dataset TITML-IDN adalah kumpulan data suara dalam bahasa Indonesia yang mencakup berbagai dialek dan aksen, sehingga memberikan model paparan terhadap variasi linguistik yang umum di Indonesia. Sementara itu, *Common Voice* adalah dataset multibahasa yang mencakup berbagai gaya bicara dan latar belakang demografi penutur, yang dapat membantu model dalam mengenali dan memahami variasi suara dari populasi penutur yang beragam.

Sebelum model di-*finetune*, dilakukan proses normalisasi amplitudo pada data audio menggunakan *TorchAudio*. Proses ini penting untuk memastikan bahwa input audio yang diberikan kepada model memiliki rentang amplitudo yang konsisten, yang berarti mengurangi variabilitas yang disebabkan oleh perbedaan volume rekaman. Normalisasi amplitudo ini bertujuan untuk mengurangi *noise* yang mungkin timbul akibat perbedaan pengaturan volume pada saat perekaman, sehingga model dapat fokus pada fitur suara yang lebih relevan untuk proses transkripsi.

Setelah proses normalisasi, model WAV2Vec 2.0 yang telah melalui *pre-training* kemudian di-*finetune* menggunakan dataset yang telah dinormalisasi. Dalam tahap ini, label transkripsi teks ditambahkan pada data suara, memungkinkan model untuk mempelajari hubungan antara suara yang didengar dan representasi teksnya. *Fine-tuning* ini bertujuan untuk memperdalam pemahaman model terhadap pola-pola spesifik yang ada dalam data berlabel, sehingga model tidak hanya mengenali suara secara umum, tetapi juga mampu menghasilkan transkripsi yang tepat sesuai dengan konteks bahasa dan intonasi yang ada. Dengan proses ini, model diharapkan mampu melakukan transkripsi dengan akurasi tinggi, yang sangat penting untuk aplikasi yang ditargetkan, seperti meningkatkan aksesibilitas bagi komunitas tuna rungu di Indonesia.

2.3 Evaluasi Model

Pada fase ketiga penelitian ini, fokus utama adalah evaluasi model WAV2Vec 2.0 yang telah dioptimalkan dengan teknik *Spectral Masking*, untuk mengukur seberapa efektif model tersebut dalam melakukan transkripsi teks dari audio yang berasal dari video. Fase ini merupakan tahap kritis karena di sinilah performa model yang telah melalui tahap *pre-training* dan *fine-tuning* diuji dalam kondisi yang mendekati dunia nyata. Tahapan evaluasi model diuraikan sebagai berikut.

1) Pengumpulan data video untuk pengujian
   Tahap awal evaluasi ini melibatkan seleksi dan pengumpulan video sebagai data uji, dipilih berdasarkan kriteria yang mencakup spektrum luas skenario dunia nyata. Pemilihan video mempertimbangkan variasi kebisingan latar belakang, jenis dan kualitas suara, serta kompleksitas audio, untuk merepresentasikan situasi seperti percakapan dalam ruangan tenang, laporan luar ruangan dengan gangguan lingkungan, dan video dengan latar belakang musik. Video juga dipilih berdasarkan relevansi konten dengan aplikasi model, seperti akses informasi bagi komunitas tuna rungu. Tujuannya adalah menguji kemampuan model dalam mentranskripsikan audio kompleks dan beragam secara akurat dalam kondisi penggunaan nyata.
2) Ekstraksi Audio dan Normalisasi Amplitudo
   Setelah video pengujian dikumpulkan, langkah selanjutnya adalah mengekstraksi audio dari video-video tersebut. Proses ini menggunakan teknik yang sama dengan yang digunakan pada tahap sebelumnya untuk memisahkan komponen audio dari video. Setelah audio diekstraksi, dilakukan proses normalisasi amplitudo untuk memastikan bahwa tingkat volume dari semua audio yang dihasilkan konsisten dan sesuai dengan standar yang digunakan selama fase finetuning. Normalisasi ini penting untuk mengurangi variasi yang tidak diinginkan yang dapat mempengaruhi kinerja model dalam mengenali dan mentranskripsikan suara.
3) *Noise Reduction (Spectral Masking)*
   Salah satu inovasi utama dalam penelitian ini adalah penerapan teknik *Spectral Masking* pada audio hasil ekstraksi. Dalam konteks penelitian ini, *Spectral Masking* diterapkan untuk memperjelas suara utama dalam audio, dengan harapan dapat meningkatkan akurasi transkripsi yang dihasilkan oleh model WAV2Vec 2.0. Audio yang telah melalui proses ini kemudian akan digunakan untuk mengevaluasi perbedaan kinerja model dalam dua skenario: dengan dan tanpa pengurangan *noise*.
4) Transkripsi Teks
   Setelah proses pengurangan *noise*, model WAV2Vec 2.0 yang telah di-*finetune* digunakan untuk mentranskripsikan audio pengujian menjadi teks. Pada tahap ini, dua skenario evaluasi digunakan: (1) transkripsi tanpa pengurangan *noise*, di mana model menggunakan audio asli tanpa modifikasi, dan (2) transkripsi dengan pengurangan *noise*, di mana audio telah melalui proses *Spectral Masking* sebelum ditranskripsikan oleh model. Pendekatan ini digunakan untuk mengevaluasi dampak langsung dari *Spectral Masking* terhadap kualitas transkripsi.
5) Evaluasi WER (*Word Error Rate*)
   Kinerja transkripsi yang dihasilkan oleh model dievaluasi menggunakan metrik WER, yang merupakan ukuran standar untuk mengukur persentase kesalahan dalam transkripsi teks. WER dihitung dengan membandingkan teks transkripsi yang dihasilkan oleh model dengan transkripsi referensi yang telah disiapkan sebelumnya. Perhitungan WER mencakup kesalahan dalam substitusi kata, penghapusan, dan penambahan, sehingga memberikan gambaran yang komprehensif tentang akurasi transkripsi. Rumus untuk menghitung WER dituliskan melalui Persamaan 1 **(Loubser dkk., 2024).**

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

dimana, S adalah jumlah *substitutions* (kata yang salah dihasilkan oleh sistem), D adalah jumlah *deletions* (kata yang dihilangkan oleh sistem), I adalah jumlah *insertions* (kata tambahan yang tidak ada dalam transkrip asli), dan N adalah total jumlah kata dalam transkrip referensi (kata yang benar-benar ada dalam teks asli).

6) Pelaporan dan Analisis
   Hasil evaluasi WER dari kedua skenario dianalisis secara mendalam untuk menentukan efektivitas teknik *Spectral Masking* dalam meningkatkan kualitas transkripsi teks. Analisis ini akan mencakup perbandingan kinerja model dalam kedua skenario, serta diskusi mengenai implikasi hasil penelitian terhadap pengembangan teknologi transkripsi teks, khususnya dalam konteks peningkatan aksesibilitas bagi komunitas tuna rungu.

3. HASIL DAN PEMBAHASAN

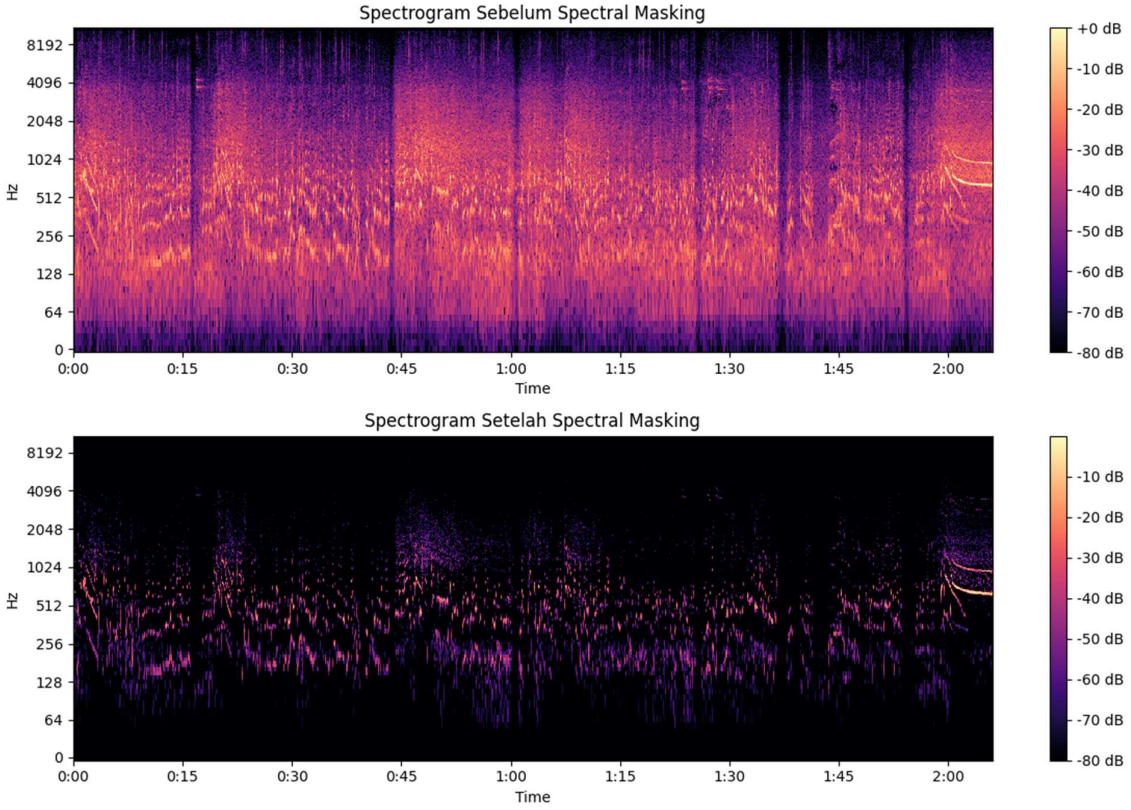
Dalam bab ini, akan dibahas hasil dari penerapan teknologi transkripsi video menggunakan WAV2Vec 2.0 yang dikombinasikan dengan teknik spectral masking. Analisis meliputi evaluasi terhadap *Word Error Rate* (WER) dan parameter kesalahan transkripsi seperti *substitution, insertion*, dan *deletion* pada 3 jenis video, yaitu video podcast, video dengan *noise*, serta video dengan background music. Sebagai bahan analisis lebih lanjut, perbandingan transkripsi sebelum dan sesudah penerapan spectral masking akan dievaluasi untuk menilai efektivitas metode *spectral masking* sebagai teknik *noise reduction* dalam meningkatkan akurasi transkripsi audio.

3.1 Analisis Pengaruh *Noise Reduction* Menggunakan *Spectral Masking*

Subbab ini menyajikan hasil analisis terhadap efek dari penerapan teknik *noise reduction* menggunakan *spectral masking* pada salah satu data uji video. Video yang dianalisis merupakan rekaman pertandingan sepakbola dengan karakteristik audio yang terdiri dari suara komentator dan *noise* latar belakang yang signifikan, terutama sorak-sorai penonton. Evaluasi terhadap perubahan pada spektrum audio sebelum dan sesudah penerapan *spectral masking*, dengan fokus pada peningkatan kejelasan suara komentator, dilakukan dengan menggunakan *spectrogram*. *Spectrogram* merupakan salah satu *tool* untuk mengidentifikasi perubahan dalam distribusi energi audio **(Sadeghi dkk., 2020)**. *Spectrogram* memberikan representasi visual dari energi pada berbagai rentang frekuensi, memungkinkan identifikasi komponen *noise* dan suara utama **(Yuliani dkk., 2021)**. Gambar 2 menunjukkan hasil *spectrogram* dari video uji VN1.mp4 yang merupakan jenis *live video* dengan *background noise* dalam bentuk suara penonton. Detail analisis dari spectogram diuraikan sebagai berikut.

1) Sebelum penerapan *spectral masking*
   a. Dominan Warna Terang (10-50 dB): mengindikasikan adanya energi yang signifikan di sebagian besar rentang frekuensi, yang konsisten dengan lingkungan siaran langsung sepakbola yang bising. Warna terang ini mewakili kombinasi dari riuh sorak penonton, suara komentator, dan *noise* lainnya.
   b. Garis Vertikal: menunjukkan sinyal yang konstan pada frekuensi tertentu, yang disebabkan oleh noise konstan seperti dengungan listrik, atau mungkin juga merupakan artefak dari proses perekaman atau pengolahan audio.

2) Setelah penerapan *spectral masking*
   a. Dominan Gelap: menunjukkan bahwa *spectral masking* telah berhasil mengurangi energi *noise* di sebagian besar rentang frekuensi.
   b. Pola Rentang Warna (40-70 dB): Pola ini kemungkinan besar mewakili suara komentator, yang sekarang lebih menonjol karena noise latar belakang telah dikurangi. Rentang warna 40-70 dB menunjukkan bahwa suara komentator memiliki intensitas yang bervariasi, yang wajar dalam percakapan manusia.



Gambar 2. Hasil Spectogram Sebelum dan Sesudah Penerapan *Spectral Masking*

Sebelum penerapan *spectral masking, spectrogram* menunjukkan dominasi warna terang yang mencerminkan tingginya energi *noise* di berbagai frekuensi. Namun, setelah *spectral masking* diterapkan, terjadi pengurangan yang signifikan pada *noise*, yang tercermin dalam dominasi warna gelap dan peningkatan visibilitas rentang frekuensi yang berkaitan dengan suara komentator. Hasil ini menunjukkan efektivitas *spectral masking* dalam mengurangi *noise* untuk memperjelas inti audio yang dibutuhkan dalam proses transkripsi (dalam hal ini suara komentator).

3.2 Analisis Hasil Transkripsi Video

Untuk mengevaluasi kinerja sistem transkripsi otomatis yang dikembangkan dalam penelitian ini, metrik WER digunakan sebagai indikator utama, untuk mengukur tingkat akurasi sistem dalam mengenali kata-kata dari audio. Metrik ini dihitung berdasarkan persentase kesalahan yang terdiri dari tiga komponen utama: *substitutions, insertions*, dan *deletions,* yang kemudian dibandingkan dengan transkripsi referensi yang benar. Analisis WER beserta komponennya dilakukan pada berbagai jenis video seperti podcast, video dengan *noise* signifikan, dan video dengan background music. Perbandingan transkripsi sebelum dan sesudah *spectral masking* (SM) dilakukan untuk mengevaluasi efektivitas metode ini dalam mengoptimalkan kinerja model WAV2Vec 2.0. Tabel 1 menunjukkan total nilai kesalahan transkripsi antara video asli (WAV2Vec 2.0) dan video dengan *noise reduction/spectral masking* (WAV2VEC 2.0 + SM).

Tabel 1. Hasil Pengujian Kesalahan Transkripsi (dalam Satuan Jumlah Kata)

Jenis Kesalahan Transkripsi	Video Podcast		Live Video		Music Video	
	Video Asli	Video dengan NR/SM	Video Asli	Video dengan NR/SM	Video Asli	Video dengan NR/SM
Subtitution	358	74	261	139	53	29
Insertion	41	18	76	66	84	73
Deletion	39	6	101	21	12	1

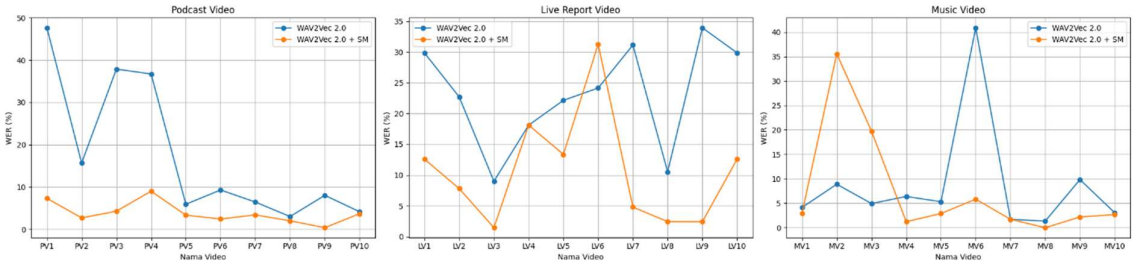
Dalam analisis ini, pengaruh *noise reduction* menggunakan *spectral masking* (NR/SM) terhadap akurasi transkripsi pada tiga jenis video berbeda—video podcast, video dengan background noise, dan video dengan background music—telah dievaluasi secara mendalam.

1) Pada video podcast dengan 2648 kata, tanpa NR/SM, jumlah kesalahan substitusi mencapai angka yang sangat tinggi, yaitu 358, yang mengindikasikan kesulitan sistem dalam mengenali kata-kata dengan akurat. Jumlah kesalahan *insertion* dan *deletion* juga cukup signifikan. Namun, setelah penerapan NR/SM, terjadi penurunan yang sangat signifikan pada semua jenis kesalahan, terutama substitusi, yang turun dari 358 menjadi 74. Hal ini menunjukkan bahwa NR/SM sangat efektif dalam meningkatkan akurasi transkripsi pada video podcast.
2) Pada video dengan background *noise* yang terdiri dari 1934 kata, tanpa NR/SM, jumlah kesalahan substitusi masih tinggi, mencapai 261, meskipun lebih rendah dibandingkan dengan video podcast tanpa NR/SM. Namun, jumlah kesalahan *insertion* dan *deletion* lebih tinggi, menunjukkan tantangan tambahan dalam mengenali kata-kata di tengah noise latar belakang. Dengan penerapan NR/SM, jumlah kesalahan substitusi berhasil dikurangi dari 261 menjadi 139, meskipun tidak seefektif pada video podcast. Jumlah kesalahan *insertion* dan *deletion* juga menurun, tetapi tetap cukup tinggi.
3) Pada video dengan background music yang memiliki 2034 kata, tanpa NR/SM, jumlah kesalahan substitusi relatif rendah, yaitu 53, dibandingkan dengan dua jenis video lainnya. Namun, jumlah kesalahan *insertion* sangat tinggi, mencapai 84, yang menunjukkan bahwa musik latar belakang menyebabkan sistem memasukkan kata-kata tambahan. Setelah NR/SM diterapkan, semua jenis kesalahan mengalami penurunan, terutama substitusi dan *insertion*. Namun, jumlah *insertion* masih cukup tinggi, yaitu 73, jika dibandingkan dengan video podcast yang menggunakan NR/SM.

Secara keseluruhan, NR/SM secara signifikan meningkatkan akurasi transkripsi pada semua jenis video, terutama dalam mengurangi kesalahan substitusi. Namun, efektivitas NR/SM bervariasi tergantung pada jenis *noise* yang ada. NR/SM paling efektif pada video podcast, diikuti oleh video dengan background *noise*, dan paling kurang efektif pada video dengan background music. Video dengan background music menghadirkan tantangan unik, karena menyebabkan sistem melakukan kesalahan *insertion* yang lebih tinggi, yakni memasukkan kata-kata tambahan yang tidak ada dalam audio asli.

3.2.1 Analisis Video Podcast

Sesuai dengan skenario pengujian yang telah diuraikan sebelumnya, pengujian dilakukan pada sepuluh video podcast yang menampilkan berbagai pembicara untuk menyediakan variasi input audio yang akan diolah oleh sistem. Video podcast dipilih sebagai contoh video dengan tingkat *noise* minimum, memungkinkan evaluasi yang lebih fokus pada kemampuan model dalam menangani variasi intonasi dan kejelasan suara. Durasi video yang digunakan dalam proses pelatihan dan pengujian model adalah 2 menit, dengan jumlah kata yang melebihi 2500 kata per video. Hasil komparasi pengujian transkripsi video disajikan dalam Gambar 3a.



(a) **(b)** **(c)**

Gambar 3. Hasil Pengujian WER antara Video Asli (WAV2Vec 2.0) dan Video dengan Noise Reduction (WAV2Vec 2.0 + Spectral Masking).

Penerapan spectral masking secara signifikan mengurangi WER pada semua video podcast yang diuji. Penurunan terbesar terlihat pada podcast-1 dengan WER turun dari 47.64% menjadi 7.30%, dan pada podcast-9 dari 8.01% menjadi 0.35%. Secara keseluruhan, WER rata-rata untuk semua video berkurang dari 17.46% pada model asli (WAV2VEC 2.0) menjadi 3.83% setelah penerapan *spectral masking*. Perbedaan rata-rata ini mengindikasikan bahwa metode *spectral masking* sangat efektif dalam meningkatkan akurasi transkripsi, terutama dalam kondisi audio yang berpotensi mengandung *noise*. Pengurangan signifikan ini menunjukkan bahwa *spectral masking* mampu menekan tingkat kesalahan dalam transkripsi, sehingga memberikan hasil yang lebih akurat dan dapat diandalkan pada berbagai jenis video podcast.

3.2.2 Analisis *Live Video*

Skenario pengujian kedua dilakukan pada sepuluh video *live report* yang menampilkan berbagai kegiatan langsung untuk menyediakan variasi input audio yang akan diolah oleh sistem. Video *live report* dipilih sebagai contoh video dengan tingkat *noise* yang bervariasi, memungkinkan evaluasi yang lebih fokus pada kemampuan model dalam menangani variasi intonasi, kejelasan suara, dan tantangan tambahan berupa *noise* latar belakang yang sering terjadi dalam situasi langsung. Durasi video yang digunakan dalam proses pelatihan dan pengujian model adalah 2 menit, dengan jumlah kata antara 1000-2000 kata per video. Hasil komparasi pengujian transkripsi *live* disajikan dalam Gambar 3b.

Penggunaan *spectral masking* pada *live video* (video dengan background *noise*) secara umum menunjukkan penurunan signifikan dalam WER pada sebagian besar video yang diuji. Penurunan terbesar terlihat pada video VN9, di mana WER turun drastis dari 33.94% menjadi 2.42%, dan pada video VN7 dari 31.14% menjadi 4.79%. Namun, pada video VN6, terjadi peningkatan WER dari 24.15% menjadi 31.29% setelah penerapan *spectral masking*, menunjukkan bahwa metode ini mungkin kurang efektif terhadap jenis *noise* tertentu. Secara keseluruhan, nilai rata-rata WER untuk video asli adalah 23.14%, sementara setelah penerapan *spectral masking* menurun menjadi 10.68%.. Hasil ini mengindikasikan bahwa *spectral masking* berhasil meningkatkan akurasi transkripsi pada mayoritas video dengan background *noise*, mengurangi tingkat kesalahan dan meningkatkan keandalan hasil transkripsi.

3.2.3 Analisis *Music Video*

Skenario pengujian ketiga dilakukan pada sepuluh video musik yang menampilkan berbagai genre dan variasi audio untuk menyediakan input yang beragam bagi sistem. Video musik dipilih sebagai contoh dengan tingkat *noise* dan latar belakang musik yang bervariasi, memungkinkan evaluasi yang lebih mendalam terhadap kemampuan model dalam menangani kompleksitas suara, kejelasan vokal, dan tantangan tambahan berupa musik latar yang sering mengaburkan kata-kata yang dinyanyikan. Durasi video yang digunakan dalam proses

pelatihan dan pengujian model adalah 2 menit, dengan jumlah kata antara 1000-2000 kata per video. Hasil komparasi pengujian transkripsi video berlatarbelakang music disajikan dalam Gambar 3c.

Penerapan *spectral masking* pada video dengan background music menghasilkan variasi dalam WER. Pada sebagian besar video, *spectral masking* berhasil mengurangi WER secara signifikan. Penurunan terbesar tercatat pada video BM6, di mana WER turun dari 40.88% menjadi 5.84%, serta pada video BM8 dari 1.35% menjadi 0.00%. Namun, dalam beberapa kasus, seperti pada video BM2 dan BM3, terjadi peningkatan WER setelah penerapan *spectral masking*. WER pada video BM2 meningkat dari 8.89% menjadi 35.56%, dan pada video BM3 dari 4.94% menjadi 19.75%. Peningkatan WER ini sebagian besar disebabkan oleh peningkatan kesalahan *insertion*, di mana sistem transkripsi menambahkan kata-kata yang tidak ada dalam audio asli. Kesalahan *insertion* ini kemungkinan besar disebabkan oleh *spectral masking* yang tidak sepenuhnya berhasil menghilangkan background music, sehingga sistem kesulitan membedakan antara suara latar dan ucapan sebenarnya, yang menyebabkan peningkatan jumlah kata yang salah dikenali atau disisipkan.

4. KESIMPULAN

Penelitian ini didorong oleh kebutuhan untuk meningkatkan akurasi transkripsi audio dalam kondisi dengan gangguan suara yang signifikan, yang sering menghambat kualitas transkripsi, terutama bagi penyandang tuna rungu. Penerapan spectral masking terbukti efektif dalam menurunkan WER pada sebagian besar video yang diuji, termasuk podcast, video dengan background *noise*, dan video dengan background music. Teknik ini berhasil meningkatkan akurasi transkripsi, meskipun terdapat keterbatasan dalam menangani *noise* kompleks seperti background music, yang terkadang meningkatkan WER akibat kesalahan *insertion*. *Spectral masking* menunjukkan potensi besar dalam optimasi teknologi WAV2Vec 2.0, terutama dalam mengurangi *noise* latar belakang, sehingga meningkatkan aksesibilitas dan pengalaman pengguna bagi penyandang tuna rungu. Inovasi ini penting untuk mendukung inklusi digital dan membuat teknologi pengenalan suara lebih adaptif terhadap berbagai kondisi audio. Kelanjutan penelitian ini mencakup adaptasi model pengenalan suara agar lebih responsif terhadap berbagai jenis noise latar serta pengujian metode ini pada berbagai bahasa dan dialek untuk memastikan aplikabilitas yang luas.

UCAPAN TERIMA KASIH

Penelitian ini mendapat dukungan secara finansial dari Direktorat Riset, Teknologi dan Pengabdian kepada Masyarakat Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi melalui hibah Penelitian Dosen Pemula tahun pelaksanaan 2024. Ucapan terimakasih juga ditujukan kepada Institut Teknologi dan Bisnis Asia Malang atas dukungan fasilitas selama pelaksanaan penelitian.

DAFTAR RUJUKAN

Andra, M. B., & Usagawa, T. (2020). Automatic Transcription and Captioning System for Bahasa Indonesia in Multi-Speaker Environment. *2020 5th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, (pp. 51–56).

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Retrieved from http://arxiv.org/abs/2006.11477

Chen, L.-W., & Rudnicky, A. (2023). Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 1–5).

Cryssiover, A., & Zahra, A. (2024). Speech recognition model design for Sundanese language using WAV2VEC 2.0. *International Journal of Speech Technology*, *27*(1), 171–177.

Direktorat Analisis dan Pengembangan Statistik. (2022). *Analisis Tematik Kependudukan Indonesia (Fertilitas Remaja, Kematian Maternal, Kematian Bayi, dan Penyandang Disabilitas)*. Badan Pusat Statistik.

Ferdiansyah, D., & Sri Kusuma Aditya, C. (2024). Implementasi Automatic Speech Recognition Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper. *Jurnal Teknik Elektro Dan Komputer TRIAC*, *11*(1), 2615–7764.

Getman, Y., Al-Ghezi, R., Grosz, T., & Kurimo, M. (2023). Multi-task wav2vec2 Serving as a Pronunciation Training System for Children. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, (pp. 36–40).

Gondi, S. (2022). Wav2Vec2.0 on the Edge: Performance Evaluation. Retrieved from http://arxiv.org/abs/2202.05993

Jain, R., Barcovschi, A., Yiwere, M. Y., Bigioi, D., Corcoran, P., & Cucu, H. (2023). A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition. *IEEE Access*, *11*, 46938–46948.

Javanmardi, F., Kadiri, S. R., & Alku, P. (2024). Exploring the Impact of Fine-Tuning the Wav2vec2 Model in Database-Independent Detection of Dysarthric Speech. *IEEE Journal of Biomedical and Health Informatics*, *28*(8), 4951–4962.

Kang, T., Han, S., Choi, S., Seo, J., Chung, S., Lee, S., Oh, S., & Kwak, I.-Y. (2024). Experimental Study: Enhancing Voice Spoofing Detection Models with wav2vec 2.0. Retrieved from http://arxiv.org/abs/2402.17127

Kozhirbayev, Z. (2023). Kazakh Speech Recognition: Wav2vec2.0 vs. Whisper. *Journal of Advances in Information Technology*, *14*(6), 1382–1389.

Loubser, A., De Villiers, P., & De Freitas, A. (2024). End-to-end automated speech recognition using a character based small scale transformer architecture. *Expert Systems with Applications*, *252*, 124119

Mozilla. (2017). *Common Voice Dataset*. Retrieved from https://commonvoice.mozilla.org/en/datasets

Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. Retrieved from http://arxiv.org/abs/1703.09452

Ragano, A., Benetos, E., & Hines, A. (2022). Learning Music Representations with wav2vec 2.0. Retrieved from http://arxiv.org/abs/2210.15310

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., … Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit. Retrieved from http://arxiv.org/abs/2106.04624

Sadeghi, M., Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R. (2020). Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, (pp. 1788–1800).

Smolik, T., Krupicka, R., & Klempir, O. (2024). Assessing Speech Intelligibility and Severity Level in Parkinson's Disease Using Wav2Vec 2.0. *2024 47th International Conference on Telecommunications and Signal Processing (TSP)*, (pp. 231–234).

Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., & Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. Retrieved from http://arxiv.org/abs/2202.12233

Tirta, L., Santoso, J., & Setyati, E. (2022). Pengenalan Lirik Lagu Otomatis Pada Video Lagu Indonesia Menggunakan Hidden Markov Model Yang Dilengkapi Music Removal. *Journal of Information System,Graphics, Hospitality and Technology*, *4*(2), 86–94.

Tokyo Institute of Technology Multilingual Speech Corpus (TITML). (2008). *TITML-IDN Dataset*. Retrieved from https://research.nii.ac.jp/src/en/TITML-IDN.html

Undang - Undang Republik Indonesia Nomor 8 Tahun 2016 Tentang Penyandang Disabilitas, 1 (2016).

Wang, D., & Chen, J. (2018). Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), (pp. 1702–1726).

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages. Retrieved from http://arxiv.org/abs/2012.12121

Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech Enhancement Using Deep Learning Methods: A Review. *Jurnal Elektronika Dan Telekomunikasi*, *21*(1), 19.