Implementasi *Automatic Speech Recognition* Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper

Danny Ferdiansyah, Christian Sri Kusuma Aditya
Program Studi Informatika, Universitas Muhammadiyah Malang
Jl. Raya Tlogomas No.246, Babatan, Tegalgondo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144
E-mail: dannyferdiansyah389@webmail.umm.ac.id, christianskaditya@umm.ac.id

***Abstract*— Implementasi *Automatic Speech Recognition* untuk memprediksi bacaan sering digunakan dalam kehidupan sehari-hari. Salah satu tujuan dilakukannya penelitian ini adalah untuk mengurangi angka buta mengaji Al-Qur'an pada umat Islam dengan mengimplementasikan ASR sebagai prediksi huruf hijaiyah dan bacaan dengan text ayat-ayat suci Al-Qur'an sebagai target. Data diambil dari *platform YouTube* dengan suara-suara *murottal* dari Syeikh Mahmoud Al-Hussary. Ada banyak metode *deep learning* ASR yang dapat digunakan untuk memprediksi kata (*transcribing*), contohnya adalah *Wav2vec 2.0* dan *OpenAI-Whisper*. Hasil dari metode *Wav2vec 2.0* menunjukkan nilai *Character Error Rate* (CER) dalam memprediksi ayat suci Al-Qur'an dari rentang 0.226 (23%) ~ 0.677 (68%). Hasil dari metode *OpenAI-Whisper* menunjukkan performa yang lebih bagus daripada *Wav2vec 2.0* dengan nilai *Character Error Rate* (CER) dari rentang 0.064 (6%) ~ 0.172 (17%). Hasil dari kedua metode yang telah diusulkan mengimplikasikan bahwa nilai *error* yang rendah menjadi metode yang terbaik dengan kesalahan yang minimal.**

Kata Kunci—Automatic Speech Recognition, Wav2vec 2.0, OpenAI-Whisper, Al-Qur'an

I. PENDAHULUAN

Al-Qur'an adalah firman Allah SWT yang diturunkan kepada Nabi Muhammad SAW untuk menentang pihak-pihak yang menentangnya dan menjadi pedoman hidup bagi umat Islam [1]. Bacaan Al-Qur'an adalah bagian paling penting dalam penanaman nilai agama dan moral agar jiwa umat Islam tumbuh diatas fitrah [2]. Pendidikan nilai agama dan moral menjadi pondasi dan harus ditanamkan kepada anak usia dini agar tetap tertanam di dalam benak pikiran dan jiwa anak. Pembelajaran Baca Tulis Al-Qur'an (BTAQ) merupakan pelajaran sebagai proses pembelajaran untuk mengenal dan mempelajari bacaan yang terkandung di dalam ayat-ayat Al-Qur'an. Dalam membaca Al-Qur'an, umat Islam dituntut untuk membaca secara tartil dan lafadz yang ada di Al-Qur'an harus benar-benar terbaca sebagaimana Allah berfirman dalam QS. Al-Muzamil Ayat 4: "*Dan Bacalah Al-Qur'an dengan tartil*" [3].

Automatic Speech Recognition (ASR) adalah sebuah teknologi yang menjadi populer dan sering digunakan manusia untuk penerapan *Voice To Text*, *Translate*, *Character Error* dan lain-lain [4]. Sistem ASR biasanya terdiri dari beberapa modul seperti pemisahan ucapan untuk menangani ucapan yang tumpang tindih [5], identifikasi pembicaraan dan untuk mentranskripsikan setiap ucapan [6]. Baru-baru ini *speech community* memiliki metode terbaru yang awalnya dari metode pemodelan *hybrid* menjadi pemodelan *End-to-End* (E2E) yang secara langsung melakukan *transcribing* ucapan menjadi *output text* [7]. Karena relevansi bahasa lisan yang sangat penting dalam kehidupan manusia, sangat penting bahwa sistem *Automatic Speech Recognition* (ASR) mampu menangani variabilitas dalam cara orang berbicara (misalnya, karena perbedaan pembicara, demografi, gaya berbicara yang berbeda, dan pengguna yang memiliki kemampuan berbeda) [8]. Sistem ASR menjanjikan untuk memberikan interpretasi yang objektif dari ucapan manusia.

Tantangan internal saat ini adalah meningkatnya angka buta mengaji Al-Qur'an, hal ini disebabkan melemahnya sistem agama pada jalur pendidikan formal, kurang perhatiannya orang tua dalam membimbing anaknya pada pengajaran Al-Qur'an. Salah satu pembelajaran yang sering digunakan saat ini yaitu memakai model *deep learning*. Pada pembuatan model *deep learning*, dibutuhkan suatu dataset yang dapat mewakili bahasa tertentu dikarenakan setiap bahasa memiliki karakteristik dan ciri khas masing-masing [9].

Deep learning memiliki banyak metode seperti *Deep Speech*, *Hidden Markov Model* dan *Connectionist Temporal Classification Model* [10]. Pada paper yang berjudul "*ETLT 2021 : Shared Task on Automatic Speech Recognition for Non-Native Children's Speech*" [4], hasil dari paper ini menunjukkan bahwa performa terbaik saat melakukan *processing* ditunjukkan dengan hasil *Word Error Rate* (WER) yang baik (untuk bahasa Inggris) atau bahkan sama (untuk bahasa Jerman) karena dalam melakukan *transcribing* dalam bahasa Jerman memiliki *best system WER* sebesar 23.50%.

Pada paper yang berjudul "*Quran Recitation Recognition using End-to-End Deep Learning*" [11], bertujuan untuk mengusulkan model *Deep Learning* dengan *End-to-End* yang baru untuk mengenali bacaan Al-Qur'an dengan metode *Convolutional Neural Network* (CNN) dengan *Connectionist Temporal Classification* (CTC) sebagai tahap *modelling*. Hasil dari paper ini menunjukkan bahwa model yang diusulkan mencapai kinerja yang baik dalam hal *Word Error Rate* (WER) dan *Character Error Rate* (CER) pada ASR.

Pada paper yang berjudul "*Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of The Person*" [12], Metode yang digunakan dalam paper ini adalah model *Whisper* yang di integrasi dengan deteksi batas heuristik. Hasil dari paper ini menunjukkan bahwa model atau metode

OpenAI's Whisper mencapai *Word Error Rate* (WER) dan *Character Error Rate* (CER) yang tidak terlalu besar yaitu 14.50% dan 8.13%.

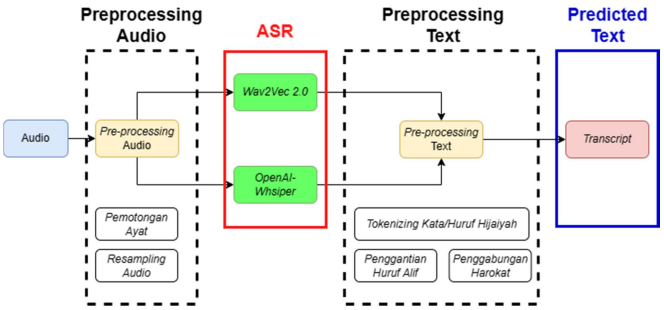
Pada paper yang berjudul "*Can Whisper Perform Speech-Based In-Context Learning?*" [13], metode yang digunakan dalam penelitian ini adalah dengan melakukan integrasi antara *Speech-based in Context Learning* (SICL) dengan model *Whisper*. Hasil dari paper ini menunjukkan bahwa model yang diusulkan mencapai *Word Error Rate* (WER) yang cukup baik yaitu sebesar 36.40%.

Wav2Vec adalah sebuah kerangka kerja untuk pembelajaran mandiri representasi dari *raw audio* [14]. Metode *Wav2Vec* menggunakan konsep "*self-supervised learning*" yang memungkinkan model untuk memahami data tanpa memerlukan label (data tanpa transkripsi teks). Baru-baru ini, beberapa model untuk *Automatic Speech Recognition* (ASR) yang menggunakan *pre-trained* yang *self-supervised* telah dirilis termasuk *Wav2Vec* [15] dan *VQ-wav2vec* [16]. Beberapa penelitian terbaru telah berhasil menerapkan representasi dari model-model ini sebagai fitur untuk *Automatic Speech Recognition* (ASR) [17], [18] dan [19]. *OpenAI* adalah sebuah perusahaan di bidang *Artificial Intelligence* yang telah mendominasi dan membantu manusia dalam mengerjakan *task* dengan menggunakan platform GPT-3 ataupun GPT-4. Selain itu, *OpenAI* membuat sebuah terobosan baru di bidang *speech recognition* yaitu *Whisper* [20]. *OpenAI-Whisper* sangat berguna dalam bidang *speech recognition* dan telah dikembangkan oleh *developer* melalui *Hugging Face* yang dapat diakses dan digunakan oleh semua orang.

Sejalan dengan kekurangan penelitian-penelitian sebelumnya, peneliti mengeksplorasi dataset dengan mencampurkan data reciter yang berbeda dengan bacaan 25 ayat suci Al-Qur'an, penggunaan dari model *wav2vec 2.0* [14], sebagai pengekstrak fitur untuk *Automatic Speech Recognition* (ASR) dan diterapkan dengan menggunakan tambahan library dari *wav2vec* yaitu Wav2Vec2ForCTC dan Wav2Vec2Processor untuk *modelling*. Selain *wav2vec 2.0*, *OpenAI-Whisper* juga digunakan sebagai perbandingan antar 2 metode dengan menggunakan *modeling* yang tersedia di *Hugging Face* yaitu "tarteel-ai/whisper-tiny-ar-quran", sehingga hasil *Automatic Speech Recognition* (ASR) bacaan Al-Qur'an mempunyai nilai *character error rate* (CER) yang optimal.

II. BAHAN DAN METODE

Tujuan dari penelitian ini adalah untuk implementasi *Automatic Speech Recognition* (ASR) dari *Wav2Vec 2.0* dan *OpenAI-Whisper* untuk menentukan metode mana yang lebih cocok dalam *speech recognition* dengan data audio bacaan Al-Qur'an. Rancangan penelitian yang diusulkan ditunjukkan pada gambar 1 dibawah ini.



Gambar 1. Rancangan Penelitian dan *Pre-processing*

Tahapan yang pertama dilakukan adalah pengumpulan data audio dan dilakukan *preprocessing audio* menggunakan masing-masing model dari *Wav2vec 2.0* dan *OpenAI-Whisper*. Selanjutnya dilakukan *ASR modelling* dengan dua metode yang diusulkan dan dilanjutkan oleh *preprocessing text*. Hasil dari *preprocessing text* adalah sebuah *predicted text*.

A. Dataset

Peneliti mengumpulkan dataset audio yang didapatkan dari *YouTube* dengan *reciter* atau suara dari *Qari'* terkenal. Nama *reciter* yang akan dijadikan sumber suara adalah Syeikh Mahmoud Al-Hussary. Tujuan dari pengambilan *reciter* ini adalah untuk pengujian seberapa pengaruhnya makhorijul huruf, bacaan dan cara mengaji pada pengujian *Automatic Speech Recognition* (ASR). Ayat-ayat Al-Qur'an yang akan diambil sebagai pengujian adalah sebanyak 25 ayat. Dataset yang dikumpulkan adalah berupa audio berekstensi wav, berikut adalah sampel bacaan-bacaan ayat suci Al-Qur'an yang akan digunakan sebagai pengujian *Automatic Speech Recognition*.

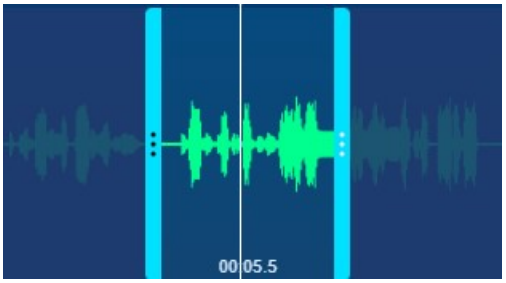
TABLE I. AYAT SUCI AL-QUR'AN SEBAGAI PENGUJIAN

No	Teks Target	Surah
1	يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا لَا تُقَدِّمُوْا بَيْنَ يَدَيِ اللّٰهِ وَرَسُوْلِهٖ وَاتَّقُوا اللّٰهَ اِنَّ اللّٰهَ سَمِيْعٌ عَلِيْمٌ	Al-Hujurat Ayat 1
2	لَمْ يَكُنِ الَّذِيْنَ كَفَرُوْا مِنْ اَهْلِ الْكِتٰبِ وَالْمُشْرِكِيْنَ مُنْفَكِّيْنَ حَتّٰى تَأْتِيَهُمُ الْبَيِّنَةُ	Al-Bayyinah Ayat 1
3	لَكُمْ دِيْنُكُمْ وَلِيَ دِيْنِ	Al-Kafirun Ayat 6
4	اِذَا زُلْزِلَتِ الْاَرْضُ زِلْزَالَهَا	Az-Zalzalah Ayat 1
5	سَبَّحَ لِلّٰهِ مَا فِي السَّمٰوٰتِ وَمَا فِي الْاَرْضِ وَهُوَ الْعَزِيْزُ الْحَكِيْمُ	Al-Hasyr Ayat 1

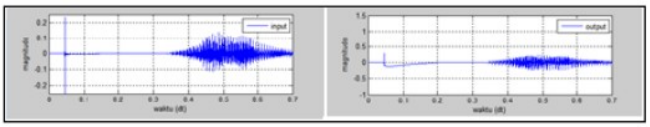
B. Preprocessing Audio

Tahap *preprocessing audio* ditunjukkan pada gambar 1. Tahap ini dilakukan untuk menghilangkan gangguan (*noise*) dalam data agar sistem *Automatic Speech Recognition* (ASR) dapat mengenali huruf hijaiyah yang dibacakan oleh *reciter*. Selain itu, ada tahapan pemotongan ayat agar sesuai dengan *target text*. Sinyal suara harus dibersihkan dan dilakukan *resampling* untuk memastikan bahwa semua file audio memiliki frekuensi yang sesuai dengan ekstensi. Pada metode *Wav2Vec 2.0*, tahap *preprocessing audio* dilakukan

dengan menggunakan library Wav2Vec2Processor. Sedangkan untuk metode *OpenAI-Whisper*, tahap *preprocessing audio* dilakukan dengan menggunakan *transformer* dari Hugging Face. Selain itu, adapun teknik *resampling* pada audio yang bertujuan untuk menggabungkan audio yang telah di *preprocessing* dan mengubah *high frequency* menjadi *low frequency*. *Pre-processing* ini bertujuan untuk memastikan bahwa semua audio memiliki kualitas yang baik sebelum dilakukan pelatihan model. Contoh tahapan dari *pre-processing audio* ditunjukkan pada gambar 2 dan gambar 3 [21] dibawah ini.



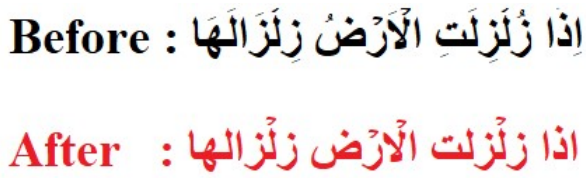
Gambar 2. Tahapan Pemotongan Ayat (*Pre-processing Audio*)



Gambar 3. Tahapan *Resampling* (*Pre-processing Audio*)

C. *Pre-processing Text*

Tahap *pre-processing text* ditunjukkan pada gambar 1. Tahap *pre-processing text* dilakukan untuk mengenali perbedaan antara bacaan waqaf, tashdid, fathatain dan lain-lain. *Pre-processing text* merupakan sebuah proses untuk mendapatkan informasi mengenai teks asli dengan hasil [22]. Teknik yang pertama dilakukan adalah melakukan *tokenizing*. *Tokenizing* dilakukan untuk memisahkan kata-kata dan huruf hijaiyah. Tahapan penggantian huruf alif pun juga dilakukan untuk mengenali huruf alif yang sebenarnya tanpa adanya huruf atau harakat lain seperti hamzah yang berada pada satu alif. Selain penggantian huruf, penggabungan harokat pada *preprocessing* juga dilakukan sebagai tahap akhir dari *preprocessing* agar kalimat atau kata yang sudah tergabung bisa dibaca kembali sebagai *predicted text*. Contoh hasil dari *preprocessing text* ditunjukkan pada gambar 4.



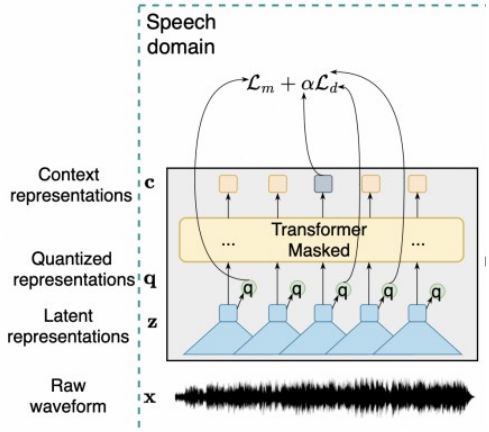
Gambar 4. Contoh Hasil *Pre-processing Text*

D. *Model Wav2Vec 2.0*

Wav2Vec 2.0 telah dilatih dengan menggunakan metode *self-supervised learning* (SSL) pada data audio dengan skala besar seperti yang ada pada gambar 5 [23]. Sebelum representasi ke dalam satuan teks, beberapa urutan *frame* yang berurutan disamarkan secara acak, kemudian digantikan oleh *vectorizer* dari Wav2Vec 2.0. Model ini mengacu pada optimalisasi kombinasi dari *contrastive loss* dan *diservity loss* $L_m + \sigma L_d$ (σ mengacu pada *hyper-parameter* atau penyeimbangan). Formula [24] dari model ini ditunjukkan pada persamaan 1.

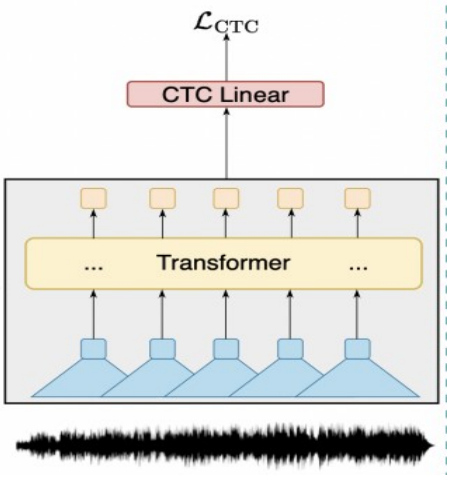
$$L_m = -log \frac{exp(sim(c_t, q_t)/k)}{\sum_{\tilde{q} \sim Q_t} exp(sim(c_t, \tilde{q}_t)/k)} \quad (1)$$

Dimana c_t adalah *t*-th *frame* dari representasi sebuah teks, Q_t merepresentasikan semua kemungkinan representasi teks, *value* dari *k* adalah 0,1 dan *sim* mengacu pada *cosine similarity*. *Diservity loss* L_d dirancang ntuk mendorong penggunaan semua input. Untuk mengetahui secara lengkap kegunaan *diservity loss*, dapat dilihat pada referensi [24].



Gambar 5. *Pre-training* Data Audio

Setelah dilakukan *pre-training*, *fine tuning* juga dilakukan dengan memberikan *labeled text* agar bisa dilakukan prediksi melalui *CTC Linear*. Pada gambar 6 [23], ditunjukkan bahwa kuantisasi data dihilangkan dan diganti menjadi pemakaian *transformer*. Model ini dilatih dengan mengoptimalisasi *connection temporal classification* (CTC) *loss* L_{CTC} [23]. Misal, *ground-truth transcription* adalah ω^* sebagai urutan dari tokenisasi karakter, maka formula [23] *CTC Loss* ditunjukkan pada persamaan 2.



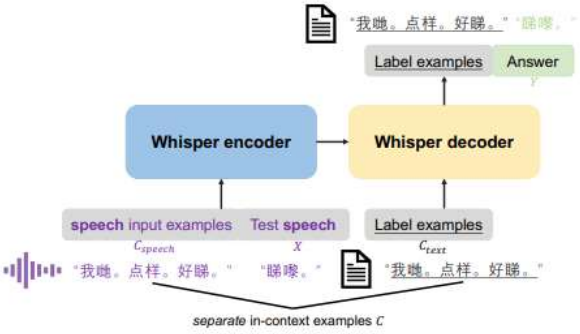
Gambar 6. *Fine-tuning* Data Audio

$$L_{CTC} = -log \sum_{\pi \in \beta^{-1}(\omega^*)} \prod_{t=1}^{T} p(\pi_t \mid f_t) \quad (2)$$

Dimana f_t bergantung pada c_t di dalam stage, T adalah jumlah *frame*, β adalah sebuah fungsi untuk *mapping* atau mengurutkan urutan $\pi_{1:T}$ to ${\omega^*}_{1:N}$ (dimana N adalah jumlah dari token karakter) dengan menghilangkan duplikasi karakter dan kata kosong menggunakan *inverse function* $\beta^{-1}(\omega^*)$ mengacu pada semua elemen CTC yang disusun dari ω^*. Probabilitas $p(\pi_t \mid f_t)$ dikomputasi melalui *CTC linear* yang diikuti oleh *softmax operation*. Disamping *supervised CTC loss*, *pseudo-labeling* dibuat ketika melakukan *fine-tuning*. Untuk mengetahui lebih lanjut tentang formula, dapat dilihat pada referensi [24] dan [25].

E. Model OpenAI-Whisper

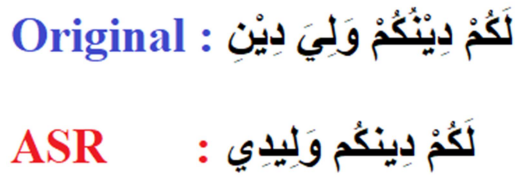
Rangkaian model OpenAI-Whisper menggunakan *encoder-decoder* arsitektur mulai dari 39 juta parameter (kecil) hingga 1,55 miliar parameter (besar) dengan fitur multi bahasa sebanyak 98 bahasa [26]. Format data dirancang untuk memungkinkan Whisper untuk melakukan banyak tugas dengan fleksibilitas dan ketangguhan. Pada OpenAI-Whisper, disediakan *encoder* dan *decoder* untuk transkripsi teks menggunakan *transformers*. Gambar 7 [20] menunjukkan *encoder* bekerja ketika mendapatkan input dari audio yang kemudian diproses oleh *decoder*. *Decoder* melakukan komparasi antara audio yang telah diterima dengan teks original (*label example*). Hasil komparasi kemudian dijadikan *text transcript* dari hasil yang terbaik sebagai *output*.



Gambar 7. Contoh ASR Bahasa China Menggunakan OpenAI-Whisper

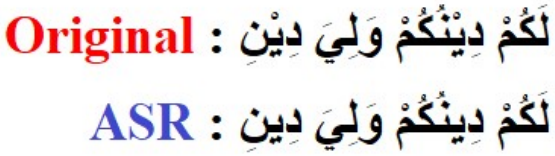
III. HASIL DAN PEMBAHASAN

Pada bab hasil dan penelitian, dijelaskan hasil dari penelitian dan pembahasan, fokus kepada *performance* dengan dataset 25 ayat suci Al-Qur'an menggunakan metode Wav2vec 2.0 dan OpenAI-Whisper. Data audio yang dipakai adalah data dari *reciter* Syeikh Mahmoud Al-Hussary. Data audio ini diproses dengan *pre-processing audio*, *pre-processing text* dan *modelling*, kemudian akan diprediksi *text transcript* melalui teks target yaitu *original text* dari Al-Qur'an. Dengan menggunakan ASR, ada kemungkinan bahwa ASR melakukan *error learning* atau *learning loss* dalam memprediksi huruf hijaiyah maupun bacaan Al-Qur'an seperti pada gambar 8.



Gambar 8. ASR *transcribing* Wav2vec 2.0

Pada gambar 8 saat memprediksi kata menggunakan metode Wav2vec 2.0, ASR bisa saja salah dalam memprediksi huruf hijaiyah karena adanya sebuah dialek dan aksen yang bersifat samar. Seperti contoh di QS. Al-Kafirun Ayat 6 huruf hijaiyah "ن" terdeteksi oleh ASR, ada kemungkinan huruf "ن" tidak terbaca karena *mad thabi'i* yang dibaca panjang 2 harakat saat membaca "وَلِيدِي" yang lebih dominan ke huruf "ي" daripada huruf "ن". Tetapi, ketika menggunakan metode OpenAI-Whisper (gambar 9), ASR mampu memprediksi huruf hijaiyah dengan tepat karena menggunakan model yang sudah tersedia di *Hugging Face* khusus untuk Al-Qur'an yaitu "tarteel-ai/whisper-tiny-ar-quran". Meskipun tepat, bacaan yang berharakat sukun dihilangkan karena sudah dianggap bacaan mati.



Gambar 9. ASR *transcribing* OpenAI-Whisper

A. Evaluasi Metode Wav2Vec 2.0

Pada pengujian ASR menggunakan metode Wav2vec 2.0, dilakukan evaluasi dengan dengan memberikan *predicted text* untuk membandingkan hasil antara *original text* dengan *predicted text* dari bacaan Syeikh Mahmoud Al-Hussary seperti pada Table 2.

TABLE II. HASIL PREDICTED TEXT METODE WAV2VEC 2.0

Target Text	Predicted Text
يَاأَيُّهَا الَّذِيْنَ اٰمَنُوْا لَا تُقَدِّمُوْا بَيْنَ يَدَيِ اللّٰهِ وَرَسُوْلِهٖ وَ اتَّقُوا اللّٰهَ اِنَّ اللّٰهَ سَمِيْعٌ عَلِيْمٌ	يَا أَيُهَا اللَذِينَ أَعمَنٌ لاتٌ قَدِّمُوبَينَ يَا دَيلَّ وَرَسُولِه وَتَّكُومإِنَ المَّهَا سَامِيِ اللنَّالم
لَمْ يَكُنِ الَّذِيْنَ كَفَرُوْا مِنْ اَهْلِ الْكِتٰبِ وَالْمُشْرِكِيْنَ مُنْفَكِّيْنَ حَتّٰى تَأْتِيَهُمُ الْبَيِّنَةُ	لَمْ يَكُنِ الَّذِينَ كَفَرو مِنْ أَهْلِلْكِتَابِ وَالْمُشْرِكِينَ مُنْفَكِّينَ حَتَّى تَأْتِيَ هُمُلِبَين
لَكُمْ دِيْنُكُمْ وَلِيَ دِيْنِ	لا تتدينكوم واليأدين
اِذَا زُلْزِلَتِ الْاَرْضُ زِلْزَالَهَا	إِذَا لُزلَةِ الأَرضُ زِيلْزَالَهَا
سَبَّحَ لِلّٰهِ مَا فِى السَّمٰوٰتِ وَمَا فِى الْاَرْضِ وَهُوَ الْعَزِيْزُ الْحَكِيْمُ	سبح لله ما في السماوات وما في الأرض وهو العزيز الحكيم

B. Evaluasi Metode OpenAI-Whisper

Pada pengujian ASR menggunakan metode OpenAI-Whisper, dilakukan evaluasi dengan dengan memberikan *predicted text* untuk membandingkan hasil antara *original text* dengan *predicted text* dari bacaan Syeikh Mahmoud Al-Hussary seperti pada Table 3.

TABLE III. HASIL PREDICTED TEXT METODE OPENAI-WHISPER

Target Text	Predicted Text
يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا لَا تُقَدِّمُوْا بَيْنَ يَدَيِ اللّٰهِ وَرَسُوْلِهٖ وَ اتَّقُوا اللّٰهَ اِنَّ اللّٰهَ سَمِيْعٌ عَلِيْمٌ	يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تُقَدِّمُوا بَيْنَ يَدَيِ اللَّهِ وَرَسُولِهِ وَاتَّقُوا اللَّهَ إِنَّ اللَّهَ سَمِيعٌ عَلِيمٌ
لَمْ يَكُنِ الَّذِيْنَ كَفَرُوْا مِنْ اَهْلِ الْكِتٰبِ وَالْمُشْرِكِيْنَ مُنْفَكِّيْنَ حَتّٰى تَاْتِيَهُمُ الْبَيِّنَةُ	لَمْ يَكُنِ الَّذِينَ كَفَرُوا مِنْ أَهْلِ الْكِتَابِ وَالْمُشْرِكِينَ مُنْفَكِّينَ حَتَّى تَأْتِيَهُمُ الْبَيِّنَةُ
لَكُمْ دِيْنُكُمْ وَلِيَ دِيْنِ	لَكُمْ دِينُكُمْ وَلِيَ دِينِ
اِذَا زُلْزِلَتِ الْاَرْضُ زِلْزَالَهَا	إِذَا زُلْزِلَتِ الْأَرْضُ زِلْزَالَهَا
سَبَّحَ لِلّٰهِ مَا فِى السَّمٰوٰتِ وَمَا فِى الْاَرْضِ وَهُوَ الْعَزِيْزُ الْحَكِيْمُ	سَبَّحَ لِلَّهِ مَا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ وَهُوَ الْعَزِيزُ الْحَكِيمُ

C. *Perbandingan Metode Wav2vec 2.0 dan OpenAI-Whisper*

Dalam section ini akan dijelaskan tentang perbandingan performa antara metode Wav2vec 2.0 dengan OpenAI-Whisper. Pada Table 4, masing-masing performa pada kedua metode diambil dari nilai *character error rate* (CER). Secara garis besar OpenAI-Whisper memiliki performa yang bagus daripada Wav2vec 2.0. Karena OpenAI-Whisper secara garis besar pemodelan dan pemrosesan data sudah menggunakan teknologi *Artificial Intelligence* yang mampu mengenali seluruh bahasa. Nilai CER OpenAI-Whisper yang didapatkan sangat baik dan konsisten dalam memprediksi sebuah huruf hijaiyah dengan rentang 0.064 (6%) ~ 0.172 (17%) sebagai nilai terburuknya. Sedangkan, dari metode Wav2vec 2.0 masih menggunakan *pre-trained model* yang sangat lama dalam melakukan komputasi pada huruf-huruf hijaiyah, sehingga nilai dari CER pada wav2vec 2.0 menjadi kurang baik.

TABLE IV. PERBANDINGAN CHARACTER ERROR RATE SETIAP METODE

Surah & Ayat	CER (%) Wav2vec 2.0	CER (%) OpenAI-Whisper
Al-Hujurat Ayat 1	46,76	16,54
Al-Bayyinah Ayat 1	24,36	15,12
Al-Kafirun Ayat 6	58,06	6,45
Az-Zalzalah Ayat 1	28,20	15,38
Al-Hasyr Ayat 1	44,82	17,24

Selain itu, nilai CER dari metode OpenAI-Whisper secara signifikan memiliki nilai yang tetap, dikarenakan model dari Whisper memiliki keakuratan dalam memprediksi bacaan Al-Qur'an. Secara algoritma, metode OpenAI-Whisper jauh lebih unggul dalam memprediksi kata atau bacaan Al-Qur'an karena adanya model yang sempurna daripada Wav2vec 2.0 yang harus memakai komponen CTC dalam memprediksi bacaan Al-Qur'an.

Perbandingan yang dilakukan selanjutnya adalah melakukan visualisasi atau ringkasan pada kedua metode yang telah di implementasikan dengan menggunakan *mean* sebagai rata-rata nilai CER. Dari hasil visualisasi yang telah didapatkan dari dua metode yang berbeda. OpenAI-Whisper memiliki CER yang lebih kecil daripada Wav2vec 2.0. Semakin kecil nilai CER maka semakin akurat juga model dalam melakukan *transcripting text*. Perbandingan performa metode ditunjukkan dalam gambar 10 dibawah ini.

Gambar 10. Hasil Perbandingan *Character Error Rate* Dua Metode

Pada metode Wav2vec 2.0, hasil dari CER lebih tinggi daripada OpenAI-Whisper. Hasil CER yang tinggi (30% ~ 35%) pada metode Wav2vec 2.0 lebih tinggi karena adanya model yang tidak kompleks dalam memprediksi bacaan Al-Qur'an, model tidak berhasil atau tidak sempurna dalam mempelajari huruf dan panjang pendek bacaan *reciter*. Model yang tidak kompleks pada Wav2vec 2.0 menjadikan nilai CER yang terlalu besar karena model tidak dapat memprediksi huruf hijaiyah secara menyeluruh. Secara hirarki, proses dari Wav2vec 2.0 menggunakan *convolutional neural network* (CNN) yang berdasar pada jaringan saraf tiruan. Sedangkan, OpenAI-Whisper secara proses berdasarkan dari *transformer* yang dapat mengolah data secara sempurna karena adanya proses *tokenizing* di dalam *transformer*.

Proses dari Wav2vec 2.0 secara mendasar adalah melakukan prediksi kata dengan tahapan *preprocessing*, melakukan pelatihan model dan didapatkan hasil. Tetapi, proses yang ada di OpenAI-Whisper lebih signifikan dengan adanya tahapan *transformer* sebelum melakukan *training model*. Dengan analisa proses yang telah dijelaskan, OpenAI-Whisper memiliki tahapan *transformer* sebelum *modelling*, sehingga CER yang didapatkan juga lebih kecil, lebih kecil nilai CER maka dapat dinyatakan model telah berhasil melakukan prediksi yang sempurna dalam memprediksi bacaan Al-Qur'an.

IV. KESIMPULAN

Pendidikan Al-Qur'an adalah hal yang sangat penting bagi umat muslim untuk bekal di akhirat kelak. Melakukan evaluasi bacaan mengaji atau bacaan Al-Qur'an adalah hal yang harus dilakukan karena untuk mempertebal keimanan manusia, memelihara ketakwaan kepada Allah SWT atau menjaga diri agar tetap berada di jalan yang tepat dan lurus. Faktor-faktor yang mempengaruhi angka buta mengaji adalah kurangnya minat, sibuk dengan dunia dan tidak ada niat. Penelitian sebelumnya memberikan model yang membutuhkan komputasi cukup lama dan tidak adanya sebuah teks transkrip dalam memprediksi bacaan Al-Qur'an. Tujuan dari penelitian ini adalah untuk mengimplementasi ASR pada bacaan mengaji dengan menggunakan metode Wav2vec 2.0 dan OpenAI-Whisper.

Hasil yang didapatkan dalam paper ini adalah metode Wav2vec 2.0 memiliki kekurangan yang signifikan dalam memprediksi kata ataupun huruf hijaiyah dengan CER paling buruk 0.580 atau 58%. OpenAI-Whisper memberikan solusi dalam penelitian ini dengan *best CER* sebesar 0.064 atau 6%. Dari visualisasi pada Gambar 10, OpenAI-Whisper juga memberikan CER yang paling kecil dan paling akurat daripada Wav2vec 2.0. Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih beragam terkait reciter bacaan Al-Qur'an yaitu dengan memberikan kategori usia, kemampuan mengaji dan fitur yang lebih luas untuk mengoptimalkan penelitian dan mengeksplorasi berbagai metode *deep learning* ataupun *machine learning*.

DAFTAR PUSTAKA

[1] A. J. Muhammad Yasir, *Studi Al-Quran*, vol. 53, no. 9. 2016.

[2] I. Sri Maharani, "Pembelajaran Baca Tulis Al- Qur ' an Anak Usia Dini," vol. 4, no. 2, pp. 1288–1298, 2020.

[3] D. I. Fitriani and F. Hayati, "Penerapan Metode Tahsin untuk Meningkatkan Kemampuan Membaca Al-Qur'an Siswa Sekolah Menengah Atas," *J. Pendidik. Islam Indones.*, vol. 5, no. 1, pp. 15–31, 2020, doi: 10.35316/jpii.v4i1.227.

[4] R. Gretter *et al.*, "ETLT 2021: Shared task on automatic speech recognition for non-native children's speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 3, pp. 1923–1927, 2021, doi: 10.21437/Interspeech.2021-1237.

[5] S. Chen *et al.*, "Continuous speech separation with conformer," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021–June, pp. 5749–5753, 2021, doi: 10.1109/ICASSP39728.2021.9413423.

[6] N. Kanda *et al.*, "Streaming Speaker-Attributed ASR with Token-Level Speaker Embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022–September, pp. 521–525, 2022, doi: 10.21437/Interspeech.2022-253.

[7] R. De Mori, "Recent advances in automatic speech recognition," *Signal Processing*, vol. 1, no. 2, pp. 95–123, 1979, doi: 10.1016/0165-1684(79)90013-6.

[8] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," 2021, [Online]. Available: http://arxiv.org/abs/2103.15122

[9] T. Novela, Martin; Basaruddin, "Dataset Suara Dan Teks Berbahasa Indonesia Pada Rekaman," vol. 11, no. 2, pp. 61–66, 2021.

[10] O. Iosifova, I. Iosifov, V. Sokolov, O. Romanovskyi, and I. Sukaylo, "Analysis of automatic speech recognition methods," *CEUR Workshop Proc.*, vol. 2923, pp. 252–257, 2021.

[11] A. Al Harere and K. Al Jallad, "Quran Recitation Recognition using End-to-End Deep Learning," pp. 1–22, 2023, [Online]. Available: https://arxiv.org/abs/2305.07034v1

[12] L. R. S. Gris, R. Marcacini, A. C. Junior, E. Casanova, A. Soares, and S. M. Aluísio, "Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of life histories of the Museum of the Person," 2023, [Online]. Available: http://arxiv.org/abs/2305.14580

[13] S. Wang, C.-H. H. Yang, J. Wu, and C. Zhang, "Can Whisper perform speech-based in-context learning," 2023, [Online]. Available: https://arxiv.org/abs/2309.07081v1

[14] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 551–555, 2021, doi: 10.21437/Interspeech.2021-703.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "WAV2vec: Unsupervised pre-training for speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019–September, pp. 3465–3469, 2019, doi: 10.21437/Interspeech.2019-1873.

[16] A. Baevski, S. Schneider, and M. Auli, "Vq-Wav2Vec: Self-Supervised Learning of Discrete Speech Representations," *8th Int. Conf. Learn. Represent. ICLR 2020*, pp. 1–12, 2020.

[17] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020–October, pp. 3755–3759, 2020, doi: 10.21437/Interspeech.2020-1212.

[18] M. MacAry, M. Tahon, Y. Esteve, and A. Rousseau, "On the Use of Self-Supervised Pre-Trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition," *2021 IEEE Spok. Lang. Technol. Work. SLT 2021 - Proc.*, pp. 373–380, 2021, doi: 10.1109/SLT48900.2021.9383456.

[19] J. Boigne, B. Liyanage, and T. Östrem, "Recognizing More Emotions with Less Data Using Self-supervised Transfer Learning," 2020, [Online]. Available: http://arxiv.org/abs/2011.05585

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022, [Online]. Available: http://arxiv.org/abs/2212.04356

[21] H. Heriyanto, H. Jayadianti, and J. Juwairiah, "The Implementation Of Mfcc Feature Extraction And Selection of Cepstral Coefficient for Qur'an Recitation in TPA (Qur'an Learning Center) Nurul Huda Plus Purbayan," *RSF Conf. Ser. Eng. Technol.*, vol. 1, no. 1, pp. 453–478, 2021, doi: 10.31098/cset.v1i1.417.

[22] A. Khumaidi and R. L. Pradana, "Identifikasi Penyebab Cacat Pada Hasil Pengelasan Dengan Image Processing Menggunakan Metode Yolo," *J. Tek. Elektro dan Komput. TRIAC*, vol. 9, no. 3, pp. 107–112, 2022, [Online]. Available: https://journal.trunojoyo.ac.id/triac/article/view/15997

[23] L. Ou, X. Gu, and Y. Wang, "Transfer Learning of wav2vec 2.0 for Automatic Lyric Transcription," 2022, [Online]. Available: http://arxiv.org/abs/2207.09747

[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 2020–December, pp. 1–19, 2020.

[25] Q. Xu *et al.*, "Self-training and pre-training are complementary for speech recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021–June, pp. 3030–3034, 2021, doi: 10.1109/ICASSP39728.2021.9414641.

[26] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017–December, no. Nips, pp. 5999–6009, 2017.