JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 10
No. 1
April 2024

Pengolahan Korpus Dataset Audio Bacaan Al-Qur'an Menggunakan Metode Wav2Vec 2.0

Aminudin[#1], Ilyas Nuryasin[#2], Saiful Amien[*3], Galih Wasis Wicaksono[#4], Didih Rizki Chandranegara[#5], I'anatut Thoifah[*6], Wahyu Rizky[#7], Danny Ferdiansyah[#8], Kiara Azzahra[#9], Fildzah Lathifah[#10], Khairunnisa Aulyah[#11]

[#]*Program Studi Informatika Universitas Muhammadiyah Malang*
Jl. Raya Tlogomas No 234 Malang

[1]aminudin2008@umm.ac.id, [2]ilyas@umm.ac.id, [4]galih.w.w@umm.ac.id,
[5]didihrizki@umm.ac.id,[7]yuurzky@webmail.ac.id, [8]dannyferdiansyah389@webmail.ac.id,
[9]kiarazzahraaa@webmail.ac.id, [10]fildzahlthfh@webmail.ac.id, [11]nisaaulyah12@webmail.ac.id,

[*]*Fakultas Agama Islam Universitas Muhammadiyah Malang*
Jl. Raya Tlogomas No 246 Malang

[3]amien75@umm.ac.id, [6]thoifah @umm.ac.id

***Abstrak*— Pengembangan sistem otomasi pengenalan ucapan (*Automatic Speech Recognition*/ASR) di dalam membaca Al-Qur'an dibutuhkan korpus data audio bacaan Al-Qur'an dan beranotasi dengan transkripsi tekstual agar dapat diproses oleh algoritma machine learning. Pemrosesan Korpus dataset ini dibangun mengingat belum adanya dataset beserta pemrosesanya menggunakan metode tertentu untuk keperluan riset di dalam pengembangan ASR. Paper ini menyajikan kumpulan *corpus dataset* dan pengolahannya menggunakan metode *Wav2Vec 2.0* dengan total 24 ribuan dataset hasil dari rekaman dari 170 santri dengan jenjang umur 4 sampai dengan 16 tahun. Pemrosesan korpus dataset dibuat mengikuti standar metode Wav2Vec 2.0 agar dapat digunakan sebagai data latih pada pemrosesan machine learning. Wav2Vec merupakan model yang dapat mempelajari representasi vektor dari masukan sinyal suara dengan proses pembelajaran *self-supervised learning*. Wav2Vec juga mampu menangani perbedaan aksen dan karakteristik pembaca Al-Qur'an yang bervariasi dan lebih akurat karena menggunakan deep learning. Dari hasil pengujian menggunakan parameter *Precision* didapatkan hasil accuracy sebesar 65.52%, *precision* dengan nilai 0.83 *Recall* dengan nilai 0.66 dan *F1-Score* dengan nilai 0.73 serta *Word Error Rate* (WER) dengan nilai 0.5. Diharapkan dengan adanya pemrosesan korpus dataset ini dapat membantu pengembangan dan riset terkait automasi sistem bacaan Al-Qur'an dengan teknik deep learning dan meningkatkan minat generasi milenial untuk belajar Al-Qur'an dengan memanfaatkan teknologi terkini.**

***Kata kunci*— Dataset, Al-Qur'an, Text-to-Speech, Audio Speech Recognition, Wav2Vec**

I. PENDAHULUAN

Al-Qur'an adalah firman *Allah SWT* yang diturunkan kepada *Nabi Muhammad SAW* untuk 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. Pendidikan nilai agama dan moral menjadi pondasi dan harus ditanamkan kepada anak usia dini agar tetap tertanam di dalam benak pikiran dan jiwa anak [2]. 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 Q.Ss Al-Muzamil Ayat 4: "Dan Bacalah Al-Qur'an dengan tartil"[3].

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 [3]. Dengan adanya tantangan tersebut perlu kiranya dihadirkan sebuah teknologi pembelajaran cara baca Al-Qur'an sebagai *complementary learning* untuk dapat mendukung ketersediaan tata cara baca Al-Qur'an yang dapat digunakan secara mandiri oleh pebelajar. Salah satu teknologi pembelajaran yang dapat digunakan untuk mendeteksi pelafalan cara baca Al-Qur'an adalah dengan menggunakan teknologi ASR. Khusus untuk teknologi tersebut agar dapat digunakan secara akurat maka harus didukung dengan dataset yang mumpuni. Paper ini membahas pengolahan dataset yang dikumpulkan dan direkam anak-anak di usia 4-7 tahun (Tingkat 1), 8-11 (Tingkat 2) dan 12-16 (Tingkat 3) yang mempelajari *Al-Qur'an* di tingkat TPQ. Sebanyak 13 TPQ di Malang Raya yang turut untuk membantu pelaksanaan kegiatan ini, 10 santri dari masing-masing TPQ ikut turut serta di dalam membantu penelitian ini.

Korpus dataset Al-Qur'an yang bersifat secara publik masih sangat terbatas jumlahnya dikarenakan banyak kriteria yang terkandung di dalam dataset tersebut

Submitted 05-10-2023; Revised 08-01-2024; Accepted 03-03-2024

diantaranya (1) ketidakseragaman orang yang membaca Al-Qur'an karena setiap pembaca memiliki gaya, tempo, intonasi, aksen yang berbeda. Keterbatasan ini dapat menyulitkan untuk mengenali model variasi suara (2) Dialektika bahasa Arab yang sangat bervariasi di mana memiliki masing-masig huruf memiliki kemiripan dan jika ada sedikit saja kesalahan maka bisa mengubah makna dan artinya (3) Kesulitan di dalam melakuan anotasi data yakni pendalaman tentang konteks makna, linguistik, fonetik untuk menghasilkan dataset yang bermutu tinggi.

Salah satu model pembelajaran machine learning dalam pemrosesan korpus dataset suara yang 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 [4]**.** *Deep learning* memiliki banyak metode seperti *Deep Speech*, *Hidden Markov Model* dan *Connectionist Temporal Classification Model* sering digunakan oleh peneliti lain untuk digunakan sebagai *Automatic Speech Recognition* (ASR) [5].

Tetapi pada kenyataannya, HMM dan Deep Speech memiliki kelebihan dan kekurangannya masing-masing. *Hidden Markov Model* (HMM) memiliki *natural framework* untuk membangun model-model ucapan yang temporal dan tersembunyi sebagai rangkaian *spectral vectors* yang menampung jarak dari frekuensi audio akan tetapi memiliki kelemahan limitasi dalam pemodelan dan akurasinya tidak terlalu akurat dalam generalisasi data yang tidak terlihat[6]. *DeepSpeech* mempunyai kelebihan untuk membedakan setiap pengucapan aksen dalam *Speech Recognition*, tetapi memiliki kelemahan dalam pengembangan model. *DeepSpeech* mulai ditinggalkan dan tidak ada pengembangan model dari *developer*. *Connectionist Temporal Classification Model* (CTC) mempunyai kelebihan menyelaraskan antara fitur ucapan dan transkripsi karakter dalam *speech* [7], tetapi memiliki konvergensi model yang lambat, kurang akurat dalam memprediksi kata benda dan membutuhkan external *Language Model* (LM). *Wav2Vec 2.0* mampu mengatasi beberapa kekurangan metode HMM, CTC dan *DeepSpeech*, yaitu model pembelajaran *End-to-End* (tanpa perlu memodelkan *hidden state* seperti HMM)[8] fleksibilitas dalam data, konteks dan model (tidak memiliki limitasi model), kecepatan pelatihan dan pengenalan karakter yang fleksibel (tanpa membutuhkan external *Language Model* seperti CTC).

Pada paper yang dibahas sebelumnya, *dataset* berupa suara dan teks dalam Bahasa Indonesia menggunakan model Tacotron2 menghasilkan MOS sebesar 4.01 [4]. Di paper ini, *dataset* disiapkan dalam audio Bahasa Arab dari *record* yang telah dilakukan oleh santri dan menggunakan metode *Wav2Vec 2.0* untuk pemodelan *dataset*. *Wav2Vec* adalah sebuah kerangka kerja untuk pembelajaran mandiri representasi dari *raw audio* [9]. Baru-baru ini, beberapa model untuk *Automatic Speech Recognition* (ASR) yang menggunakan *pre-trained* yang *self-supervised* telah dirilis termasuk *Wav2Vec* dan *VQ-wav2vec* [10]. Beberapa penelitian terbaru [11], [12], [13] telah berhasil menerapkan representasi dari model-model ini sebagai fitur untuk ASR. Sejalan dengan penelitian-penelitian tersebut, dalam paper ini dieksplorasi penggunaan dari model *wav2vec* 2.0 [9], sebuah versi perbaikan dari model *wav2vec* asli, sebagai pengekstrak fitur untuk ASR dan diterapkan dengan menggunakan tambahan *library Wav2Vec2ForCTC*, *B2Vec2Processor*, sehingga hasil *Automatic Speech Recognition* (ASR) bacaan Al-Qur'an mempunyai nilai *word error rate* (WER) yang optimal.

II. METODE PENELITIAN

Metode Wav2Vec merupakan model *Automatic Speech Recognition* (ASR) yang mampu mengidentifikasi sinyal suara dan menerjemahkannya menjadi teks dengan tingkat akurasi yang tinggi walaupun data pelatihan yang berlabel terbatas ketersediaannya. Metode Wav2Vec menggunakan konsep "*self-supervised learning*" yang memungkinkan model untuk memahami data tanpa memerlukan label (data tanpa transkripsi teks). Dengan demikian, model ini dapat bekerja efektif tanpa bergantung pada jumlah data yang memiliki label.

Di dalam paper ini menggunakan metode Wav2Vec, dengan mengumpulkan data audio bacaan Al-Qur'an dalam bentuk rekaman mengaji yang melibatkan anak-anak berusia 4 hingga 16 tahun. Pengumpulan dataset ini dilakukan secara langsung dengan mengunjungi 13 TPQ di wilayah Malang Raya dan merekam setiap kegiatan mengaji dari dengan total 170 santri kemudian target hasil perekaman 200 jam.

A. Pengumpulan Dataset

Dataset audio Al-Qur'an yang diperlukan untuk melatih model Wav2Vec2. Dataset ini merupakan komponen kunci dalam pengembangan transkripsi otomatis bacaan audio Al-Qur'an. Audio dipilih sebagai bacaan Al-Qur'an dari berbagai sumber yang tersedia offline. Dataset dipastikan bahwa audio yang dipilih memiliki kualitas audio yang baik dan representatif

B. Preprocessing

Tahap ini merupakan langkah pertama yang sangat penting. Pada tahap ini, sinyal suara harus dibersihkan dari gangguan (noise) dan dilakukan dilakukan proses resampling untuk memastikan bahwa semua file audio memiliki frekuensi sampel yang seragam menggunakan software Adobe Audition. Selanjutnya, proses normalisasi audio juga perlu dilakukan dengan membagi data dalam sinyal suara oleh nilai amplitudo maksimum guna mencapai konsistensi dalam tingkat volume audio. Tindakan ini bertujuan untuk memastikan bahwa semua audio memiliki tingkat volume yang seragam, yang pada gilirannya membantu menghindari distorsi suara dan meningkatkan akurasi transkripsi.

C. Pelatihan Model

Dataset akan dibagi menjadi tiga bagian utama untuk keperluan pelatihan dan pengujian model Wav2Vec2, yaitu data *train*, data *validation*, dan data *test*. Dari ketiga dataset ini akan dilakukan ke tahap selanjutnya yaitu :

- *Pretraining* (Pra - Pelatihan)

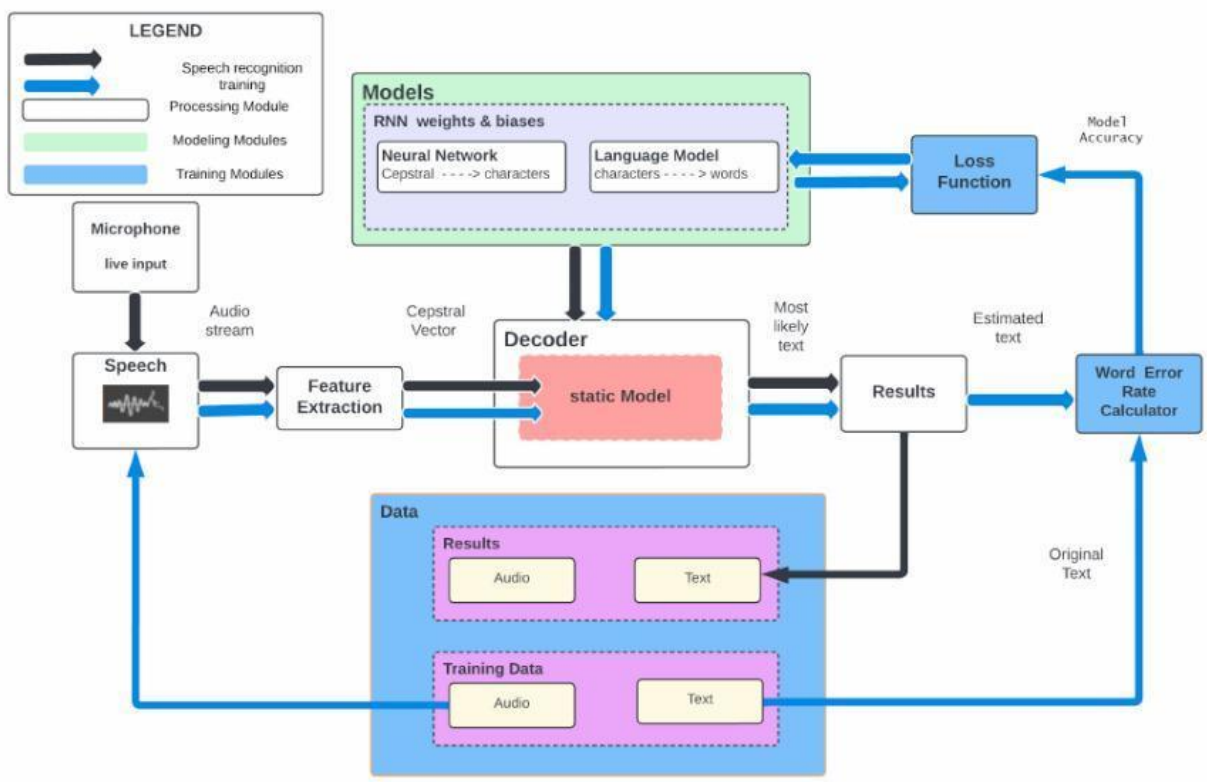
Pada tahap ini, model akan diberikan data *train* untuk mempelajari struktur audio dasar tanpa memerlukan transkripsi teks dalam beberapa iterasi yang membutuhkan pengoptimalan parameter model untuk meminimalkan kesalahan dalam menerjemahkan audio menjadi teks. Sedangkan, untuk data *validation* akan digunakan untuk mengukur sejauh mana model Wav2Vec2 berhasil dalam memahami representasi audio.

- WER(*Word of Error*)

Word Error Rate (WER) adalah metrik yang umumnya digunakan untuk mengukur kinerja sistem pengenalan suara atau sistem transkripsi otomatis, terutama dalam kasus pengenalan ucapan atau teks oral. WER mengukur seberapa baik sistem dapat mengenali dan mentranskripsikan ucapan menjadi teks. WER dihitung sebagai perbandingan antara jumlah kata yang disalin dengan benar, diubah, atau dihapus oleh sistem, dibandingkan dengan total jumlah kata dalam transkripsi referensi. Secara formal, WER dapat dihitung menggunakan formula berikut :

$$WER = \frac{Insertion + Deletion + Substitution}{Total\ Kata\ dalam\ Referensi} x100$$

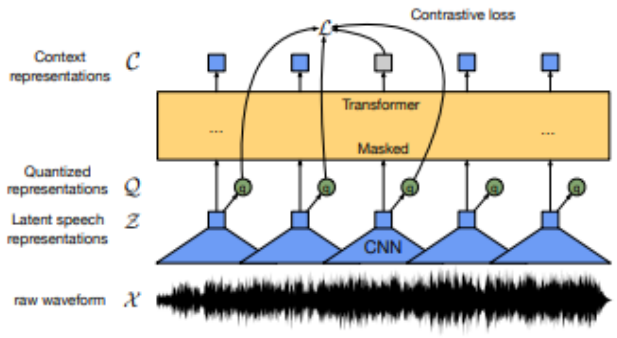
WER menyediakan gambaran yang holistik tentang kesalahan pengenalan suara dan transkripsi, mencakup perubahan kata, penghapusan kata, dan penambahan kata. Semakin rendah nilai WER, semakin baik kinerja sistem. Nilai WER didefinisikan ketika mendekati 1 maka nilai WER nya semakin bagus.



Gambar 1. Flowchart langkah-langkah penelitian

D. Modelling Dataset

Dataset yang telah diproses sebelumnya kemudian digunakan untuk proses pelatihan dengan model Wav2Vec2 yang menggunakan jaringan Convolutional Neural Network. Tahap pelatihan menggunakan 70% data latih dan 15% data uji. Model dilatih untuk mencerna spektogram ucapan dan menghasilkan transkripsi teks bahasa Arab. Model Wav2Vec2 terdiri dari encoder fitur konvolusi multi-layer $f: X \rightarrow Z$ yang mengambil input raw audio X dan mengeluarkan representasi ucapan laten $z_1 \quad z_T$ dengan T untuk setiap langkah waktu. Setiap layer kemudian dibawa ke Transformer $g: Z \rightarrow C$ untuk membangun representasi $c_1 \quad c_T$ yang menangkap informasi dari seluruh urutan. Keluaran dari encoder fitur di diskritisasi ke q_T dengan modul kuantisasi $Z \rightarrow Q$ untuk merepresentasikan target (Gambar 2) dalam tujuan self-supervised.



Gambar. 2 Ilustrasi framework Wav2Vec2 yang tergabung untuk mempelajari representasi ucapan yang disesuaikan dengan konteks

Untuk data latih model self-supervised, model Wav2Vec2 dipisahkan dengan output melalui fitur encoder z ke finite set dalam representasi ucapan lewat kuantisasi produk. Kuantisasi produk ini digunakan untuk memilih representasi berdasarkan berbagai codebooks. Contoh diberikan G codebooks, atau grup, dengan V dari entri e $\in R^{V \times d/G}$, diambil satu entri dari setiap codebook dan menggabungkan vektor yang dihasilkan e_1 e_G dan menerapkan transformasi linier $R^d \rightarrow R^f$ untuk mendapatkan q $\in R^f$. Fitur luaran encoder z dipetakan ke logit 1 $\in R^{G \times V}$ dan probabilitas untuk memilih entri codebook ke-v untuk grup g adalah:

$$P_{g,v} = \frac{exp(l_{g,v}+n_v)/T}{\Sigma_{k=1}^{V} exp(l_{g,k}+n_k)/T} \quad (1)$$

dimana T adalah non-negative temperature, $n = -log(-log(u))$ dan u adalah sampel uniform dari $U(0,1)$. Selama proses forward pass, codeword i dipilih berdasarkan $i = argmax_j P_{g,j}$ dan untuk backward pass, gradien yang sebenarnya dari softmax Gumbel digunakan x

E. Skenario Uji

Di dalam *training* dan *testing*, akan dievaluasi model menggunakan *validation* untuk mencari parameter yang optimal, seperti halnya *network architecture*, *learning rate*, dan banyak iterasi di dalam training. Setelah mengidentifikasi parameter terbaik dalam set validasi, dilanjutkan untuk melatih model menggunakan seluruh dataset pelatihan, termasuk set validasi. Selanjutnya, model dievaluasi menggunakan dua himpunan independen, yaitu himpunan uji dan himpunan inferensi. Dalam scenario pengujian akan dihitung *Word Error Rate* (WER) dalam mengevaluasi kinerja rekaman bacaan Al-Qur'an model pengenalan suara. WER digunakan untuk mengukur tingkat kesalahan dalam mengenali teks yang diucapkan dengan benar. Semakin rendah nilai WER, semakin baik kinerja model dalam pengenalan suara dan interpretasi teks bacaan Al-Qur'an.

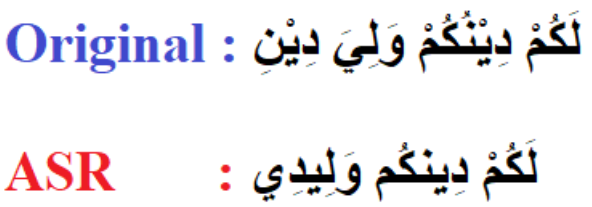
Kunci untuk pelatihan adalah membuat modifikasi kecil pada parameter untuk melihat kesesuaian tiap percobaan.. Setiap kali menggunakan set pelatihan dan set validasi untuk melatih model. Kemudian model akan diuji dengan set pengujian, dan akan menghasilkan WER. Perlu dicatat, bahwa kesalahan dalam pengucapan huruf, makhraj dan aksen dari bacaan dapat berpengaruh pada WER. WER di dapat dari model berdampak pada teks yang diprediksi.

Sementara itu, labelling *teks target* dan *teks prediksi* juga merupakan salah satu *matrik* evaluasi yang diperhitungkan. *Matrix* ini mengukur sejauh mana model memberikan hasil yang benar. Meskipun labelling *teks target* dan *teks prediksi* ini penting, para peneliti menyadari bahwa *matrix* ini dapat menjadi kurang informatif dalam konteks ketidakseimbangan kelas data, seperti dalam tugas pengenalan suara rekaman pembacaan Al-Qur'an.

III. Hasil dan Analisa

Dataset dikumpulkan dengan melibatkan 170 santri yang secara sukarela merekam bacaan dengan bacaan membaca teks yang sudah ditentukan sebelumnya seperti buku 6 jilid di buku bacaan Tilawati, Qiro'ati dan Ummi. Dataset tersebut kemudian dilakukan proses pembersihan audio agar tidak adanya *noise* dalam data. Setelah tahap pembersihan, data audio diseragamkan sinyal dan tingkat volume nya dalam membantu menghindari distorsi suara dan meningkatkan akurasi. Setelah berhasil membersihkan data audio, menyamakan sinyal dan tingkat volume audio telah dilakukan dengan tujuan untuk menciptakan konsistensi dalam dataset, menghindari potensi distorsi suara, dan meningkatkan akurasi dalam proses pengenalan suara. Data audio yang sudah bersih dan seragam dalam hal sinyal dan volume siap untuk diproses ke tahap pelatihan model.

Sebelum melakukan pelatihan, dataset dibagi menjadi 3 (tiga) bagian utama, yaitu data pelatihan (*train*), data validasi, dan data pengujian (*test*) dengan rasio masing-masing 60:20:20. Data pengujian dengan jumlah 24 ribu total durasi 200 jam perekaman. Dari data tersebut dilakukan uji sebanyak lima kali pengujian dengan membagi 24 ribu berdasarkan rasio yang telah ditentukan. Selanjutnya, model Wav2Vec2 menjalani fase training menggunakan dataset pelatihan menggunakan model trained yang sudah ada dalam bahasa Arab. Fase ini digunakan untuk membantu model memahami hubungan antara urutan audio dan teks. Hasil dari proses ini adalah proses pelatihan model yang telah diadaptasi secara khusus untuk mengenali dan mentranskripsi bacaan Al-Qur'an dalam bahasa Arab. Diharapkan model yang telah dibangun ini siap digunakan untuk penelitian lebih lanjut atau aplikasi yang memerlukan pengenalan suara dalam konteks bahasa Arab dengan tingkat akurasi yang optimal. Pada Gambar 3, dapat dilihat bahwa ASR membuat beberapa kesalahan dalam memprediksi kata.



Gambar. 3 Contoh pemrosesan data Original Text dan hasil ASR.

Saat memprediksi kata, ASR bisa saja salah dalam memprediksi huruf karena adanya sebuah dialek dan aksen yang bersifat samar. Seperti contoh, Di QS. Al-Kafirun Ayat 6 huruf hijaiyah "ن" tidak terdeteksi oleh ASR, ada kemungkinan huruf "ن" tidak terbaca karena *mad thobi'i* yang dibaca panjang 2 harakat saat membaca "وَلِيدِي" yang lebih dominan ke huruf "ي" daripada huruf "ن".

TABEL I
HASIL PREDIKSI WAV2VEC2 DENGAN CONTOH TARGET

No.	Teks Target	Teks Prediksi
1	قُلْ أَعُوذُ بِرَبِّ ٱلنَّاس	قُلْ أَعُوذُ بِرَبِّنَا
2	لَمْ يَكُنِ الَّذِيْنَ كَفَرُوْا مِنْ اَهْلِ الْكِتٰبِ وَالْمُشْرِكِيْنَ مُنْفَكِّيْنَ حَتّٰى تَأْتِيَهُمُ الْبَيِّنَة ُ	لَمْ يَكُنْ الَّذِينَ كَفَرُوا مِنْ أَهْلِ الْكِتَابِ وَالْمُشْرِكِينَ مُ فَكِينَ حَتَّى تَأْتِيَهُمُ الْبَيِنَ
3	لَكُمْ دِيْنُكُمْ وَلِيَ دِيْنِ	لَكُمْ دِينكُم وَلِيدِي
4	اِذَا زُلْزِلَتِ الْاَرْضُ زِلْزَالَهَا	إِذَا زُنْزِلَتِ الأَرضِ لْزَالَهَا
5	وَمِنْ اٰيٰتِهٖٓ اَنْ خَلَقَ لَكُمْ مِّنْ اَنْفُسِكُمْ اَزْوَاجًا لِّتَسْكُنُوْٓا اِلَيْهَا وَجَعَلَ بَيْنَكُمْ مَّوَدَّةً وَّرَحْمَةً ۗ اِنَّ فِيْ ذٰلِكَ لَاٰيٰتٍ لِّقَوْمٍ يَّتَفَكَّرُوْنَ	وَمِنْ آيَاتِ أَنْ خَلَقَ لَكُم مِنْ أَنْفُسِكُمْ أَزْوَاجًا لِتَسْكُنُو إِلَيْهَا وَجَعَلَ بَيْنَكُمَّ وَدَّةً وَرَحْمَ إِنَّ فِي ذَلِكَ لَآيَاتٍ لِقَومِيَّةَ فَكَّرُو

Tabel 1 menunjukkan contoh perbandingan antara teks target (yang seharusnya dibaca) dan teks prediksi yang dihasilkan oleh Automatic Speech Recognition (ASR). Pada contoh pertama, ASR seharusnya mengenali kata "ٱلنَّاس" (an-nas) sebagai teks target, tetapi menghasilkan "رَبِّنَا" (rabbina), yang berarti "Tuhan kami". Kesalahan ini bisa disebabkan oleh dialek atau aksen yang berbeda dalam pengucapan "ٱلنَّاس". Pada contoh kedua, ASR tidak mengenali kata "مُنْفَكِّيْن" (munfakkin) dengan benar. Kesalahan ini mungkin disebabkan oleh pengucapan yang tidak jelas atau dialek yang berbeda. Pada contoh ketiga, ASR tidak mengenali kata "وَلِيَ دِيْنِ" (waliya deeni) dengan benar. Kesalahan ini bisa disebabkan oleh pengucapan yang tidak jelas atau perbedaan dialek. Pada contoh keempat, ASR mengalami kesalahan dalam mendeteksi kata "زُلْزِلَت" (zulzilat) dengan benar. Kesalahan ini bisa disebabkan oleh ketidakjelasan dalam pengucapan. Pada contoh kelima, ASR mengalami kesalahan dalam mendeteksi kata "إِيَّة" (iyyata) dengan benar. Kesalahan ini bisa disebabkan oleh ketidakjelasan dalam pengucapan atau variasi dalam dialek.

Dalam penggunaan ASR, berbagai jenis kesalahan dapat terjadi saat mengenali dan mentranskripsikan ucapan menjadi teks. Kesalahan-kesalahan ini adalah contoh bagaimana ASR dapat salah memprediksi kata-kata dalam teks karena variasi dalam bahasa dan pengucapan. Kesalahan ASR bisa disebabkan karena substitusi kata, penghilangan kata, perubahan kata, ketidakjelasan dalam pengucapan, kesalahan dalam pengenalan tanda baca atau spasi, kesalahan dalam pengenalan nomor atau angka, serta penggantian aksen atau dialek. Bahasa seringkali dipengaruhi oleh dialek regional, aksen, atau variasi individu. ASR harus bisa mengatasi berbagai variasi ini untuk dapat mengenali ucapan yang akurat sehingga untuk meningkatkan akurasi ASR, penting memiliki pemahaman tentang konteks dan pengetahuan bahasa yang lebih luas untuk memperbaiki hasil yang dihasilkan oleh ASR, terutama saat bekerja dengan teks yang bersifat samar atau mengandung variasi dialek.

TABEL II
HASIL PREDIKSI WAV2VEC2 DENGAN ACCURACY, PRECISION, RECALL, F1-SCORE DAN WORD ERROR RATE

Pengujian	Akurasi	Presisi	Recall	F1	WER
1	65.52%	0.83	0.66	0.73	0.5
2	27.97%	0.30	0.28	0.29	0.75
3	41.94%	0.57	0.42	0.48	0.75
4	46.15%	0.49	0.46	0.47	1.0
5	10.94%	0.12	0.11	0.12	0.9

Tabel 2 menampilkan hasil prediksi yang dilakukan dari penerapan model Wav2Vec2 pada kumpulan data audio yang berisi bacaan Al-Qur'an. Evaluasi hasil ini dilakukan dengan menggunakan berbagai matrik evaluasi standar, termasuk Akurasi, *Precision*, *Recall*, *F1-Score*, serta *Word Error Rate* (WER). Akurasi mencerminkan sejauh mana model berhasil mengklasifikasikan audio secara tepat dan dalam analisis ini, dapat diamati bahwa tingkat akurasi yang dihasilkan tidak rata
dengan performa tertinggi pada pengujian pertama dan dan performa yang terendah pada pengujian yang terakhir. Hal ini disebabkan dataset yang dihasilkan ketika recording sangat bervariasi atau bisa disebabkan hasil datasetnya ada noisy sehingga mengakibatkan akurasinya menjadi naik turun. *Precision* mengukur jumlah prediksi positif yang

tepat dibandingkan dengan total prediksi positif, sementara *Recall* mengukur jumlah prediksi positif yang tepat dibandingkan dengan total data yang sebenarnya positif. *F1-Score*, sebagai rata-rata dari *precision* dan *recall*, memberikan gambaran keseluruhan tentang performa model.

Hasil prediksi ini juga telah berhasil mencapai tingkat *Word Error Rate* (WER) 0.5%. Tingkat *Word Error Rate* (WER) sebesar 0.5% adalah indikator bahwa model ini telah mengalami pelatihan yang baik dan mampu mengekstraksi fitur audio dari ucapan dengan tingkat akurasi yang tinggi. Ini menunjukkan kemajuan yang signifikan dalam pengenalan dan transkripsi audio dalam konteks bacaan Al-Qur'an. Namun demikian, tidak dapat diabaikan bahwa masih terdapat error dalam prediksi model. Salah satu penyebabnya adalah keterbatasan *pre-trained* model yang digunakan untuk bacaan dan penulisan Al-Qur'an. *Pre-trained* model ini mungkin belum memahami variasi yang ada dalam bacaan Al-Qur'an dengan cukup mendalam, sehingga menghasilkan beberapa kekeliruan dalam transkripsi. Hal ini menunjukkan bahwa terdapat ruang untuk pengembangan lebih lanjut dalam pemahaman model terhadap pembahasan khusus seperti bacaan Al-Qur'an.

IV. KESIMPULAN

Berdasarkan hasil pengujian menggunakan Metode Wav2Vec dengan konsep "*self-supervised learning*" yang memungkinkan model untuk memahami data tanpa memerlukan label (data tanpa transkripsi teks). Dengan demikian, model ini dapat bekerja efektif tanpa bergantung pada jumlah data yang memiliki label. Hasil dari proses pelatihan adalah model yang telah diadaptasi secara khusus yang dapat mengenali dan mentranskripsi ucapan bahasa Arab dengan tingkat akurasi yang lebih tinggi daripada model yang belum melalui fine-tuning. Hasil prediksi yang telah dilakukan menggunakan Wav2Vec terhadap dataset audio bacaan Al-Qur'an berhasil mencapai tingkat *Word Error Rate* (WER) 1%. Hal ini menjelaskan bahwa model sudah terlatih dengan cukup baik sehingga mampu mengekstraksi fitur audio ucapan dengan akurat. Meskipun demikian, masih terdapat sedikit kekeliruan terhadap model prediksi dikarenakan *pre-trained* model untuk bacaan atau penulisan bacaan Al-Qur'an masih belum masif sehingga belum bisa mencapai akurasi yang baik. Untuk penelitian lebih lanjut, diperlukan dataset audio dengan aksen, makhroj dan pelafalan yang jelas untuk mengurangi tingkat *Word Error Rate* (WER). Selain itu, perlu adanya pengembangan model baru yang mampu memprediksi sekaligus aksen dan makhraj agar tingkat WER tidak terlalu tinggi.

UCAPAN TERIMA KASIH / ACKNOWLEDGMENT

Kami mengucapkan rasa terima kasih kepada semua pihak yang telah memberikan kontribusi dan dukungan penuh dalam menyelesaikan penelitian dan penulisan paper ini dengan sepenuhnya. Kami juga ingin mengucapkan rasa terima kasih kepada Direktorat Jenderal Pendidikan Tinggi (Ditjen Dikti) yang melalui program *Matching Fund* telah memberikan dukungan keuangan yang sangat berarti dalam pelaksanaan penelitian ini sehingga memungkinkan penelitian kami berjalan lancar dan sukses.

REFERENSI

[1] Yasir, Muhammad, Jamaruddin, and Ade, *Studi Al-Qur'an*. 2002.
[2] S. Maharani, "Pembelajaran Baca Tulis Al-Qur'an Anak Usia Dini," *Jurnal Pendidikan Tambusai*, 2020.
[3] D. I. Fitriani, "Penerapan Metode Tahsin untuk Meningkatkan Kemampuan Membaca Al-Qur'an Siswa Sekolah Menengah Atas," *Jurnal Pendidikan Islam Indonesia*, vol. 5, no. 1, 2020, doi: 10.35316/jpii.v4i1.227.
[4] M. Novela and T. Basaruddin, "DATASET SUARA DAN TEKS BERBAHASA INDONESIA PADA REKAMAN PODCAST DAN TALK SHOW," *Agustus*, vol. 11, no. 2, pp. 61–66.
[5] O. Iosifova, I. Iosifov, V. Sokolov, O. Romanovskyi, I. Sukaylo Ender Turing OÜ, and P. str, "Analysis of Automatic Speech Recognition Methods," 2021.
[6] A. M. Deshmukh, "Comparison of Hidden Markov Model and Recurrent Neural Network in Automatic Speech Recognition," *European Journal of Engineering Research and Science*, vol. 5, no. 8, pp. 958–965, Aug. 2020, doi: 10.24018/ejers.2020.5.8.2077.
[7] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 1408–1412. doi: 10.21437/Interspeech.2019-1938.
[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.05862
[9] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.03502
[10] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.05453
[11] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning 'BERT-like' Self Supervised Models to Improve Multimodal Speech Emotion Recognition," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.06682
[12] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.09212
[13] J. Boigne, B. Liyanage, and T. Östrem, "Recognizing More Emotions with Less Data Using Self-supervised Transfer Learning," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.05585