Pengembangan Sistem Koreksi Bacaan Doa Harian Anak Menggunakan Fine-Tuning Model Whisper

Umaymatun Az-Zauraa'
Program Studi Informatika
Universitas Islam Indonesia
Daerah Istimewa Yogyakarta, Indonesia
22523283@students.uii.ac.id

Hari Setiaji, S.Kom., M.Eng.
Program Studi Informatika
Universitas Islam Indonesia
Daerah Istimewa Yogyakarta
hari.setiaji@uii.ac.id

Abstract*—Pendidikan agama Islam pada usia dini, khususnya dalam pembiasaan doa sehari-hari, menghadapi tantangan dalam hal pengawasan ketepatan pelafalan. Metode konvensional sering kali terkendala waktu pendamping, sementara aplikasi digital yang ada umumnya hanya menyediakan fitur pemutaran audio tanpa koreksi otomatis. Penelitian ini bertujuan untuk mengembangkan sistem koreksi bacaan doa anak dengan memanfaatkan teknologi *Automatic Speech Recognition* (ASR) melalui metode *fine-tuning* pada model Whisper OpenAI (tarteel-ai/whisper-base-ar-quran). Model dilatih menggunakan *dataset hybrid* yang dikurasi dari rekaman langsung di Taman Pendidikan Al-Qur'an (TPA) dan sumber daring. Evaluasi dilakukan menggunakan metrik *Word Error Rate* (WER) dan *Character Error Rate* (CER), serta diimplementasikan dalam prototipe berbasis Streamlit yang mengintegrasikan algoritma *Levenshtein Distance* untuk memberikan umpan balik akurasi bacaan. Hasil penelitian menunjukkan bahwa proses *fine-tuning* mampu meningkatkan kinerja model secara signifikan, dengan penurunan WER dari 0.991 (*baseline*) menjadi 0.271 dan CER dari 0.448 (*baseline*) menjadi 0.059. Pengujian prototipe terhadap responden anak menunjukkan bahwa sistem mampu mengenali variasi pelafalan anak dengan tingkat akurasi tinggi dan memberikan umpan balik visual yang relevan. Penelitian ini membuktikan bahwa adaptasi model ASR pada domain spesifik suara anak dapat menjadi solusi efektif dalam mendukung pembelajaran doa mandiri yang interaktif.

Keywords— Automatic Speech Recognition, Whisper, Fine-tuning, Doa Harian Anak, Levenshtein Distance.

I. PENDAHULUAN

Pendidikan agama Islam pada usia dini memegang peranan krusial dalam pembentukan karakter dan perilaku positif, salah satunya melalui pembiasaan praktik ibadah dan doa sehari-hari yang dapat memperkuat nilai moral pada anak [1]. Usia dini sering kali disebut sebagai masa keemasan (*golden age*) di mana anak memiliki potensi belajar sangat besar, sehingga perhatian pendidikan pada fase ini sangat menentukan tumbuh kembang mereka [2]. Namun, dalam praktiknya, anak-anak sering kali mengalami kesulitan dalam melafalkan doa dengan *makhraj* dan tajwid yang benar. Anak-anak kerap mengalami kesulitan spesifik pada perubahan fonem dari tebal ke tipis, atau tertukarnya pelafalan huruf dengan *makhraj* yang berdekatan [2]. Metode pembelajaran konvensional seperti *talaqqi* yang bergantung pada supervisi guru dan orang tua sering kali kurang efektif karena tidak didukung oleh media visual yang dapat membantu anak memosisikan mulut saat melafalkan huruf yang sulit [3].

Di sisi lain, tren digitalisasi Islam terus meningkat sebagai respons terhadap gaya belajar generasi masa kini yang menuntut metode interaktif dan fleksibel [4]. Meskipun aplikasi edukasi Islam terus berkembang, mayoritas *platform* tersebut hanya terbatas pada penyajian fitur pemutaran audio saja. Sebuah studi oleh Khairani et. al mengungkap bahwa belum ada aplikasi populer di pasaran yang menyematkan fitur koreksi bacaan otomatis untuk domain doa anak [5]. Padahal, umpan balik korektif ini sangat penting untuk meningkatkan keterlibatan dan pemahaman anak [4], sekaligus untuk memastikan proses belajar berjalan dengan benar sejak awal, terutama untuk doa berbahasa Arab yang memiliki kaidah pelafalan yang ketat.

Perkembangan teknologi *Automatic Speech Recognition* (ASR), khususnya model Whisper dari OpenAI, menawarkan solusi potensial melalui kemampuannya yang *robust* terhadap *noise* dan variasi bahasa [6]. Dengan kemampuan multibahasa dan toleransi terhadap variasi fonetik, Whisper memiliki potensi besar dalam mengenali bacaan doa berbahasa Arab yang diucapkan oleh anak-anak. Model ini juga dapat dikolaborasikan dengan *speech analysis* untuk mengidentifikasi letak kesalahan pelafalan dan memberi umpan balik secara otomatis. Penelitian sebelumnya membuktikan bahwa Whisper memiliki kinerja unggul dalam mengenali Bahasa Arab dengan *Word Error Rate* (WER) di bawah 10% untuk domain formal, serta mampu meningkatkan akurasi transkripsi secara signifikan pada pembelajaran bahasa Arab [5].

Kendati demikian, studi *benchmarking* oleh Talafha, et al. menunjukkan bahwa Whisper sangat berpotensi untuk digunakan dalam pendidikan keagamaan anak, tetapi memerlukan proses *fine-tuning* agar mampu mengenali karakteristik vokal anak dengan lebih akurat [7]. Oleh karena itu, penelitian ini bertujuan mengembangkan sistem koreksi bacaan doa dengan melakukan *fine-tuning* pada model Whisper (menggunakan basis tarteel-ai/whisper-base-ar-quran) menggunakan *dataset hybrid* yang dikurasi dari lingkungan TPA dan sumber daring untuk menjembatani kesenjangan tersebut. Sebagai implementasi praktis, penelitian juga menghadirkan prototipe aplikasi berbasis Streamlit yang mengintegrasikan algoritma *Levenshtein Distance/Similarity* untuk memberikan umpan balik terkait letak kesalahan bacaan anak. Secara spesifik, penelitian ini dirancang untuk menjawab pertanyaan penelitian berikut: (1) Seberapa efektif metode *fine-tuning* pada model Whisper dalam meningkatkan akurasi pengenalan bacaan doa anak dibandingkan model *baseline*? dan (2) Bagaimana implementasi sistem koreksi otomatis dapat memberikan umpan balik visual yang informatif bagi pembelajaran anak?

II. STUDI LITERATUR

A. *Perbandingan Model Speech Recognition dalam Pengembangan Aplikasi Berbahasa Arab*

Dalam proses pengembangan sistem koreksi bacaan untuk aplikasi pembelajaran doa anak berbasis *Automatic Speech Recognition* (ASR), pemilihan model *speech recognition* menjadi aspek penting yang harus dikaji untuk

menentukan keberhasilan implementasinya. Meskipun berbagai pendekatan seperti Whisper, Google Speech API, CNN/Sphinx, dan Wav2Vec 2.0 telah diterapkan dalam penelitian sebelumnya, sebagai model transkripsi audio berbahasa Arab, tetapi sifatnya masih *general*, yakni audio bacaan Quran dengan suara orang dewasa. Sehingga, belum ada model yang secara khusus dioptimalkan untuk domain doa anak-anak.

Tinjauan ini dilakukan untuk membandingkan model-model *speech recognition* yang telah disebutkan, dengan mempertimbangkan beberapa konteks khusus dalam pembelajaran anak. Hal ini dikarenakan, efektivitas masing-masing model akan sangat dipengaruhi oleh konteks penggunaannya, baik dari segi jenis *dataset* pelatihan, fleksibilitas dalam pelatihan ulang, kemampuan beroperasi secara *offline*, respons *real-time*, serta kinerjanya dalam menangani kompleksitas fonetik Bahasa Arab. Sehingga, analisis komparatif yang disajikan dalam TABEL I menjadi landasan untuk mengidentifikasi model paling adaptif bagi kebutuhan sistem koreksi bacaan doa anak yang akan dikembangkan.

TABEL I. PERBANDINGAN MODEL ASR BERBAHASA ARAB DALAM PENELITIAN TERDAHULU

Model ASR	***Dataset***	**Hasil**	**Kelebihan**	**Kekurangan**
Whisper (OpenAI)	Kosakata Bahasa Arab [5]	Akurasi = 95.52%	Mendukung *fine-tuning* [5] [8], *offline* [8], multibahasa, mendukung bahasa arab [5] [7], cocok untuk anak dengan penyesuaian [5]	Performa menurun pada dialek spesifik tanpa *fine-tuning* [8] [7]
Google Speech API	Surah pendek [9]	Akurasi = 92-100%	Mudah diintegrasikan [9], mendukung bahasa arab [10]	Tidak mendukung *fine-tuning*, *online* [9] [10], tidak mampu memverifikasi kaidah tajwid Al-Quran [10]
CNN/ Sphinx	Hadits dan surat Al-Fatihah	Akurasi = 39-89%	Ringan untuk segmentasi fonem [11], cocok untuk sistem kecil	Tidak akurat untuk transkripsi panjang [11]
Wav2 Vec 2.0	Bahasa Arab	WER = 7-10%	Mendukung *fine-tuning* [12], bisa *offline* [12], Akurasi tinggi [12] [13], *pretraining* kuat	Komputasi berat [13], tidak tersedia versi ringan untuk *edge devices*

Berdasarkan tabel di atas, model Whisper menjadi pilihan utama yang memenuhi kriteria, yakni mendukung bahasa Arab, dapat beroperasi secara *offline*, fleksibilitasnya untuk dilatih ulang sesuai dengan domain anak, serta performanya yang cukup baik untuk *dataset* kosakata berbahasa Arab. Model ini juga telah digunakan dalam beberapa penelitian terdahulu, untuk eksperimen pada anak dengan hasil yang cukup baik [5] [7]. Sedangkan model lain seperti Google Speech API tidak mendukung fleksibilitas pelatihan ulang, CNN dan Sphinx tidak memberi dukungan pada transkripksi panjang, akurasinya pun terbilang tidak konsisten, dan Wav2Vec 2.0 terlalu berat untuk implementasi pada perangkat yang akan dirancang.

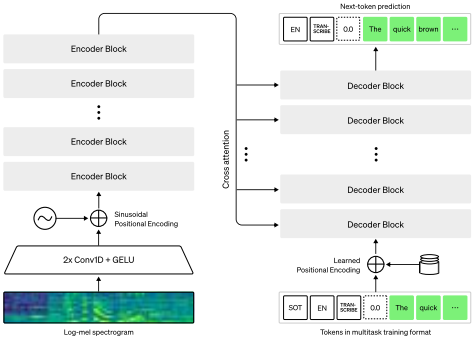
B. Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) merupakan cabang kecerdasan buatan yang memungkinkan sistem untuk mengenali dan mengubah sinyal suara atau ucapan manusia menjadi representasi teks secara otomatis melalui pencocokan pola sinyal dan representasi numerik [14]. Sistem ASR modern umumnya terdiri dari beberapa komponen penting dan krusial: modul pra-pemrosesan untuk membersihkan *noise*, ekstraksi fitur untuk mengambil karakteristik sinyal, serta model klasifikasi dan model bahasa untuk memprediksi teks keluaran sesuai konteks [15]. ASR bekerja dengan mengenali pola suara yang masuk, mengekstraksi fitur, dan mencocokkannya dengan model bahasa untuk menghasilkan representasi teks [16]. Dalam pengembangannya, pendekatan ASR modern berbasis *deep learning* kini lebih dominan dibandingkan metode konvensional seperti *Hidden Markov Model* (HMM) karena keunggulannya dalam menangani variasi suara dan kompleksitas bahasa.

Salah satu kemajuan mutakhir dalam teknologi ASR saat ini adalah model pengenalan suara Whisper yang dikembangkan oleh OpenAI. Model ini menunjukkan performa luar biasa dalam mengenali ucapan di berbagai kondisi, mulai dari aksen yang beragam hingga variasi *noise* atau kebisingan pada latar belakang, berkat kemampuannya beroperasi dalam pengaturan *zero-shot* [6]. Kemampuan ini menjadikan Whisper unggul dibandingkan model tradisional yang sangat bergantung pada pelatihan berbasis domain spesifik, karena model ini dapat beradaptasi tanpa memerlukan pelatihan ulang ekstensif pada *dataset* baru [6].

Meskipun teknologi ASR telah berkembang pesat, penerapannya pada domain spesifik seperti doa anak dalam bahasa Arab memiliki tantangan tersendiri. Dari sisi linguistik, bahasa Arab memiliki keragaman pada panjang pendeknya harakat, kompleksitas struktur bahasa, serta fonem yang unik. Selain itu, suara anak-anak ikut memberikan tantangan tambahan karena karakteristik vokal yang dimiliki belum stabil, variasi kecepatan bicara, dan kemungkinan *noise* pada latar suara. Hal ini menuntut pengembangan ASR yang tidak hanya akurat secara komputasi, tetapi juga adaptif terhadap keterbatasan sumber daya data pada bahasa tertentu dan karakteristik suara

C. Whisper OpenAI



Gambar 1. Arsitektur Whisper [17]

Whisper merupakan model *Automatic Speech Recogiton* (ASR) *end-to-end* yang dibangun diatas arsitektur *transformer encoder-decoder* [6]. Model ini dilatih menggunakan pendekatan *weak supervision* pada *dataset* masif berdurasi lebih dari 680.000 yang mencakup data

audio multibahasa, sehingga mampu mendukung pengenalan ucapan dari berbagai bahasa dengan cukup akurat [6] [7]. Whisper memiliki beberapa keunggulan utama, diantaranya mampu berjalan secara *offline*, memungkinkan pelatihan ulang atau *fine-tuning* untuk kebutuhan spesifik, serta memiliki struktur *encoder-decoder transformer* yang memudahkan integrasi dalam *pipeline* ASR.

Seperti yang disajikan dalam Gambar 1, arsitektur yang dimiliki oleh Whisper menggunakan pendekatan *end-to-end* yang diimplementasikan sebagai *transformer encoder-decoder* [17]. Mekanisme kerja Whisper diawali dengan proses segmentasi audio, di mana Whisper memproses audio yang masuk sebagai *input* dengan memecahnya menjadi bagian berdurasi 30 detik, yang kemudian dikonversi menjadi visual *log-Mel spectogram*. Representasi fitur ini kemudian diproses oleh *encoder* untuk memahami konten akustik, lalu diterjemahkan menjadi teks oleh *decoder.*

D. Adaptasi Domain Spesifik pada ASR

Model ASR yang dilatih pada *dataset* umum sering kali mengalami penurunan performa/kinerja ketika diuji pada domain yang berbeda, kondisi ini dikenal sebagai *domain missmatch* [18] [19] [20]. Dalam konteks pengenalan ucapan anak, ketidaksesuaian ini terjadi karena model yang *general* biasanya dilatih menggunakan suara orang dewasa dengan karakteristik suaranya yang stabil. Sedangkan suara anak-anak memiliki frekuensi yang beragam dan cenderung lebih tinggi dengan variasi kejelasan artikulasi dan pelafalan, yang menyulitkan model standar untuk melakukan penerjemahan secara akurat [21].

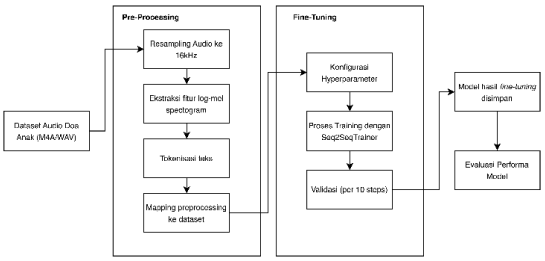
Untuk mengatasi kendala tersebut, teknik adaptasi domain melalui proses pelatihan ulang model atau *fine-tuning model* menjadi solusi yang efektif [8] [22]. *Fine-tuning* memungkinkan model *pre-trained* seperti Whisper, untuk menyesuaikan parameter bobotnya terhadap *dataset* yang spesifik [8]. Studi menunjukkan bahwa penerapan *transfer learning* dari model yang telah mempelajari bahasa terkait seperti Bahasa Arab ke data anak-anak akan meningkatkan akurasi pengenalan kata dibandingkan melakukan pelatihan dari awal [5].

E. Levenshtein Distance

Levenshtein distance atau yang sering disebut sebagai *edit distance*, merupakan algoritma matematis yang digunakan untuk mengukur perbedaan antara 2 sekuens teks [23]. Algoritma ini menghitung jumlah minimum operasi pengeditan, seperti penyisipan, penghapusan, serta penggantian yang terjadi saat perubahan satu *string* ke *string* lainnya terjadi. Dalam evaluasi model ASR, algoritma ini menjadi dasar dalam penilaian kinerja atau performa model dengan perhitungan metrik standar *Word Error Rate* (WER) dan *Character Error Rate* (CER) [19]. Selain itu, dalam membangun sistem koreksi bacaan, *Levenshtein Similarity Ratio* antara teks transkripsi yang dihasilkan oleh model dan teks referensi doa dimanfaatkan untuk memberikan skor akurasi bacaan pada pengguna. Skor yang ditampilkan berupa persentase beserta kategori kualitas bacaan, serta detail letak kesalahan bacaan agar pengguna dapat memperbaiki bacaan doanya dengan lebih mudah.

III. METODOLOGI PENELITIAN

Penelitian ini berfungsi untuk meningkatkan model ASR Whisper untuk mengenali bacaan doa anak melalui proses *fine-tuning* dengan *dataset* spesifik yang dikurasi melalui beberapa sumber. Langkah awal sistem yakni proses *pre-processing* data audio dengan mengkonversi laju sampel dalam *dataset* ke frekuensi 16kHZ untuk mencapai kinerja yang maksimal sesuai dengan kebutuhan model Whisper. Kemudian, pemrosesan data berlanjut ke dua tahapan penting yakni tokenisasi dan ekstraksi fitur ke bentuk *log-mel spectogram*. Selanjutnya, data melalui proses *mapping* dan proses *training* untuk *fine-tuning* model dapat dimulai. Detail alur proses *fine-tuning model* dapat dilihat pada Gambar 2.



Gambar 2. Alur Proses *Fine-Tuning* Model ASR Whisper

A. Dataset

Dataset yang digunakan dalam penelitian ini merupakan *dataset* audio bacaan doa harian anak yang dikumpulkan secara *hybrid* dari dua sumber utama. Sumber pertama yakni rekaman audio yang diambil secara langsung dari tiga Taman Pendidikan Al-Qur'an (TPA) yang ada di Kabupaten Sleman. Kemudian, sumber kedua yakni data tambahan yang diperoleh dari sumber daring/internet yang relevan, yang menyediakan bacaan doa anak dalam bahasa Arab. Sumber daring yang dimaksud yakni pencarian pada *platform* YouTube dan *search engine* Google. Penggabungan dua sumber ini dilakukan untuk memperluas variasi audio suara, memperkaya gaya pelafalan, serta meningkatkan *robustness* model terhadap *noise,* perbedaan karakteristik, maupun artikulasi.

Pada bagian pengumpulan data secara langsung, setiap anak diminta untuk melafalkan doa-doa pendek pilihan yang umum diajarkan pada tingkat pembelajaran dasar-dasar keagamaan umat muslim, seperti doa sebelum tidur, doa bangun tidur, dan doa sebelum makan. Untuk meningkatkan jumlah sampel dan memberi variasi pengucapan, setiap anak diminta menghafal satu doa sebanyak satu hingga tiga kali. Rekaman diambil menggunakan media ponsel yang terhubung dengan *microphone wireless* untuk mengurangi *noise,* dengan format hasil audio yakni M4A. Proses transkripsi dilakukan secara manual dengan penyesuaian harakat, konsistensi penulisan teks Arab, serta pengecekan ulang untuk kesalahan anotasi,

Sementara itu, data tambahan yang diperoleh dari sumber daring/internet, merupakan data rekaman doa anak-anak dengan konten sehari-hari yang dikurasi, lalu dipotong sesuai kebutuhan pelatihan. Data ini kemudian di konversi ke format standar yakni WAV, lalu ditranskripsikan secara manual sesuai dengan format *dataset* rekaman langsung agar konsisten.

Secara keseluruhan, karakteristik *dataset* audio yang digunakan dalam proses pelatihan, validasi, dan pengujian model ASR untuk kebutuhan sistem koreksi bacaan doa anak, adalah sebagai berikut:

- *Dataset* diambil dari dua sumber utama, yakni kombinasi rekaman dari 3 TPA di wilayah Kabupaten Sleman (TPA Ar-Rahmah Sawitsari, TPA Al-Baitul

Makmur Wonorejo, dan TPA An-Nahrawi Plosokuning) dengan total 32 anak (17 laki-laki dan 15 perempuan) dengan rentang usia 6-12 tahun, serta data daring terkurasi.

- *Dataset* terdiri dari 3 jenis doa, yakni doa sebelum tidur, doa bangun tidur, dan doa sebelum makan, yang masing-masing terdiri dari 100-120 data audio dengan format M4A/WAV.
- *Dataset* memiliki *sampling rate* 16 kHZ, mono.
- Durasi rata-rata audio 7.99 detik
- Total durasi *dataset* 42 menit 36.94 detik
- Untuk *dataset* TPA yang diambil secara langsung, setiap anak menghafal sebanyak 1-3 kali untuk setiap doa.

Untuk memperjelas distribusi *dataset*, TABEL II merangkum jumlah audio dalam setiap jenis doa.

TABEL II. RANGKUMAN DISTRIBUSI DATASET

Doa	Jumlah Audio	Total Durasi	Rata-Rata Durasi	Jumlah Anak/Sumber
Doa sebelum makan	100	7.58 detik	12 menit 38.24 detik	32 anak + 38 data daring
Doa sebelum tidur	120	5.90 detik	11 menit 48.2 detik	31 anak + 55 data daring
Doa bangun tidur	100	10.91 detik	18 menit 10.51 detik	100 data daring

Dataset yang telah dikumpulkan tersebut, kemudian dibagi menjadi tiga kelompok data, yakni data latih, data validasi, dan data uji. Pembagian dilakukan dengan pendekatan *stratified sampling* berdasarkan jenis doa, dengan detail pembagian, yakni data latih 80%, data validasi 10%, dan data uji 10%. Pendekatan ini dipilih untuk memastikan setiap doa terdistribusi secara merata di setiap kelompok data, sehingga model tidak bias pada doa tertentu.

B. Pra-Pemrosesan Dataset Audio

Sebelum masuk ke tahap pelatihan, data audio melalui beberapa tahapan *pre-processing*. Tahapan ini dilakukan untuk memastikan seluruh data audio berada dalam format yang sama dan cocok dengan *pipeline* pelatihan model ASR. Berkas audio dibaca menggunakan modul *torchaudio*, kemudian dinormalisasi melalui beberapa tahapan:

1) Resampling audio ke 16.000 Hz: audio yang sudah dibaca, lalu dikonversi ke 16 kHz agar sesuai dengan spesifikasi model ASR Whisper dari OpenAI.

2) Ekstraksi fitur log-mel spectogram: dengan memanfaatkan modul *WhisperFeatureExtractor*, audio yang sudah berada pada frekuensi 16 kHz kemudian dikonversi ke representasi fitur standar Whisper, yakni *log-mel*. Hasil ekstraksi ini kemudian disimpan sebagai *input_features* yang menjadi masukan untuk *encoder* Whisper.

3) Tokenisasi teks: teks transkripsi doa yang telah dinormalisasi diproses menggunakan *tokenizer* Whisper. Token ID disimpan dalam kolom *labels* sebagai luaran model saat pelatihan.

4) Mapping preprocessing ke dataset: ketiga tahapan tersebut diterapkan ke *dataset train, validation,* dan *test* melalui fungsi *prepare_dataset()* dalam *dataset.map()*. Format fitur inilah yang siap digunakan untuk pelatihan serta evaluasi.

C. Arsitektur Model

Model ASR yang digunakan sebagai dasar pengembangan sistem koreksi pada penelitian ini yakni model ASR tarteel-ai/whisper-base-ar-quran yang tersedia secara *open-source* pada *platform* HuggingFace. Model ini merupakan bentuk *fine-tuned model* dari model aslinya yakni openai/whisper-base. Varian Whisper ini dipilih karena telah dilatih khusus menggunakan *dataset* berbahasa Arab khususnya Al-Qur'an.

D. Pelatihan Model

Proses pelatihan model dilakukan dengan pendekatan *fine-tuning* menggunakan model Whisper dari tarteel-ai/whisper-base-ar-quran pada lingkungan komputasi Google Colab dengan memanfaatkan GPU NVIDIA L4 (Ada Lovelace). Proses *fine-tuning* lebih berfokus pada adaptasi model terhadap domain bacaan doa anak dengan variasi durasi dan ragam pelafalan. GPU L4 dipilih karena memiliki kapasitas memori dan kemampuan komputasi yang lebih optimal dibanding GPU T4 dan CPU yang disediakan oleh Colab. Pelatihan dilakukan menggunakan *library* HuggingFace Transformers, yang menyediakan kelas *Seq2SeqTrainer* sebagai *training loop* utama. *Dataset* yang telah melalui tahap pra-pemrosesan dimuat ke dalam *trainer* sebagai *train_dataset* dan *eval_dataset*.

Pelatihan model dilakukan dengan menerapkan konfigurasi *hyperparameter* yang disesuaikan dengan ukuran *dataset*, kebutuhan stabilitas pelatihan, dan keterbatasan memori GPU L4. Detail *hyperparameter* yang digunakan tercantum dalam TABEL III.

TABEL III. KONFIGURASI HYPERPARAMETER

Hyperparameter	Nilai
Batch Size per Device	4
Gradient Accumulation Steps	2
Learning Rate	1e-4
Weight Decay	0.01
Warmup Steps	50
Max Steps	100
Evaluation Strategy	Steps
Evaluation Steps	10
Save Strategy	Steps
Save Steps	100
Logging Strategy	Steps
Logging Steps	10
FP16	True
Predict with Generate	True

E. Pengujian dan Evaluasi Model

Pada tahap ini, model yang telah selesai dilatih melalui proses *fine-tuning*, diuji menggunakan *dataset* pengujian (data uji) untuk mengukur kemampuan model dalam mengenali bacaan doa yang tidak pernah dilihat sebelumnya, saat pelatihan. *Dataset* pengujian diperoleh dari hasil pemisahan *dataset* validasi menjadi dua bagian secara *stratified* agar distribusi tiap jenis doa dalam kelompok data tetap seimbang. Data dibagi dengan proporsi 50% untuk data validasi dan 50% untuk data uji dari folder valid.

Selanjutnya, evaluasi model dilakukan dengan menggunakan dua metrik utama, yakni *Word Error Rate* (WER) dan *Character Error Rate* (CER). Dalam penelitian ini, WER digunakan untuk mengukur tingkat kesalahan model pada level kata, sedangkan CER digunakan untuk mengukur tingkat kesalahan model pada level karakter. Hal ini dikarenakan, CER akan lebih sensitif terhadap susunan atau fonem bahasa Arab yang memiliki harakat, huruf, serta

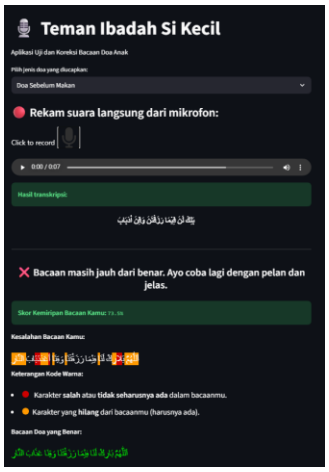
variasi kecil lainnya dalam doa. Secara matematis, WER dan CER dapat dihitung dengan persamaan (1) dan (2) di bawah ini, di mana S adalah substitusi, D adalah *deletion* atau penghapusan, I adalah *insertion* atau penyisipan, dan N adalah jumlah kata atau karakter pada label:

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

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

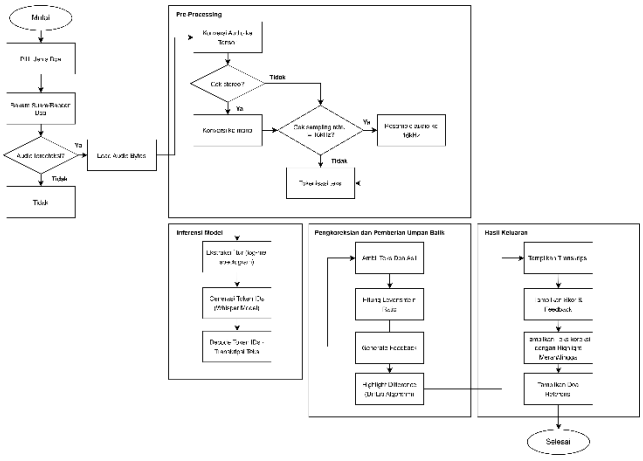
Hasil dari kedua metrik ini menjadi indikator utama yang menentukan kualitas model dalam mengenali bacaan doa anak. Selain itu, kedua metrik ini juga digunakan untuk membandingkan performa antar percobaan, serta menjadi referensi kualitas model sebelum diintegrasikan ke aplikasi/sistem koreksi bacaan.

F. Implementasi Prototipe Sistem



Gambar 3. Antarmuka Prototipe Sistem Koreksi Bacaan Doa Anak

Untuk memberikan gambaran alur penggunaan sistem koreksi bacaan yang dibangun, serta menguji kemampuan model hasil *fine-tuning* dalam mengenali dan mengoreksi bacaan doa anak, implementasi prototipe sistem dilakukan dengan memanfaatkan *framework* Streamlit sebagai antarmuka interaktif. Tampilan antarmuka disajikan dalam Gambar 3. Sistem ini memungkinkan pengguna merekam suara secara langsung melalui mikrofon, kemudian model Whisper memproses audio tersebut untuk mendapatkan transkripsi.



Gambar 4. Alur Sistem Koreksi Bacaan Anak

Sebagaimana yang dapat dilihat dalam Gambar 4, proses pengenalan suara melalui beberapa tahapan setelah pengguna selesai merekam suara menggunakan mikrofon. Tahapan ini meliputi pemuatan audio dari *browser*, normalisasi format audio melalui proses *resampling* ke 16 kHZ, ekstraksi fitur *log-mel spectogram*, inferensi menggunakan model, hasil transkripsi keluar, dan sistem akan melakukan perhitungan skor kemiripan bacaan menggunakan *Levenshtein Similarity Ratio*. Sistem juga dilengkapi dengan fitur penyorotan perbedaan karakter antara bacaan pengguna dan teks doa asli menggunakan *character-level diff highlighting*, sehingga kesalahan bacaan doa anak dapat terlihat secara visual.

G. Algoritma Evaluasi Bacaan dan Similarity Score

Proses evaluasi bacaan yang diimplementasikan dalam prototipe, dilakukan dengan membandingkan hasil transkripsi model Whisper terhadap referensi asli teks doa menggunakan pendekatan *string similarity*. Metode yang digunakan yakni *Levenshtein Distance*. Pendekatan ini sesuai dengan kasus bacaan doa pada anak-anak karena kesalahan yang umumnya muncul pada pelafalan doa oleh anak yakni kekeliruan huruf atau huruf tertukar, hilangnya huruf pada bacaan, serta panjang pendeknya bacaan yang menyebabkan karakter tertentu hilang.

Untuk menilai kemiripan bacaan, prototipe memanfaatkan algoritma bawaan dalam modul *Levenshtein* yakni *levenshtein ratio*, yang secara matematis dapat didefinisikan dalam persamaan (3):

$$Similarity = 1 - \frac{LD(pred,ref)}{\max(len(pred),len(ref))} \quad (3)$$

dengan pred adalah hasil transkripsi yang dibuat oleh model ASR dan ref adalah referensi teks doa asli dengan penulisan yang benar. Nilai asli *similarity* berada pada rentang 0-1, di mana nilai mendekati 1 menunjukkan bahwa kedua teks tersebut memiliki kesamaan yang hampir identik. Dalam tampilan antarmuka pada prototipe nilai ini, dimunculkan sebagai persentase kemiripan bacaan dengan label 'Skor Kemiripan Bacaan Kamu:'. Skor yang ditampilkan akan berada pada rentang 0-100%, Selain itu, untuk memberikan umpan balik yang mudah dipahami oleh anak-anak dan orang tua maupun guru TPA, nilai *similarity* kemudian di kategorisasikan ke dalam tiga tingkat kualitas bacaan sebagaimana ditunjukkan pada TABEL IV.

TABEL IV. KATEGORI KUALITAS BACAAN DOA

Rentang *Similarity*	**Kategori**	**Bentuk Umpan Balik**	**Makna**
≥ 0.90	Benar	✅ Bacaanmu sudah benar!	Bacaan sangat mirip, kesalahan minor
0.75–0.89	Perlu perbaikan	⚠️ Bacaanmu cukup bagus, coba perbaiki sedikit lagi.	Ada kesalahan kecil di beberapa kata yang dapat diperbaiki
< 0.75	Salah/ kurang tepat	❌ Bacaan masih jauh dari benar. Ayo coba lagi dengan pelan dan jelas.	Banyak kesalahan, perlu mencoba lagi atau latihan lebih lanjut

Selain memberikan kategori penilaian, prototipe juga menampilkan penyorotan pada bagian bacaan yang dinilai kurang tepat untuk membantu pengguna dalam mengenali letak kesalahan bacaan. Proses ini dilakukan menggunakan pendekatan *character-level diff* dari *difflib.ndiff*, di mana karakter yang tidak sesuai ditandai dengan warna merah, dan

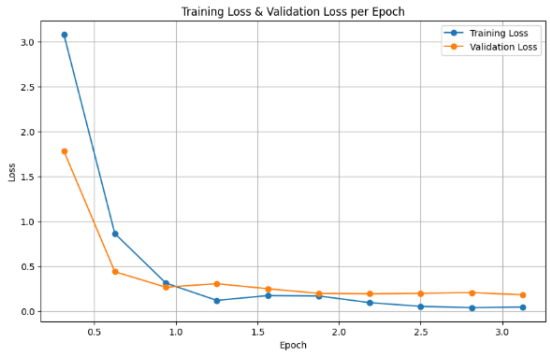
karakter tambahan yang dihasilkan dari transkripsi model (yang seharusnya tidak ada) diberi warna jingga. Pendekatan ini membuat proses evaluasi lebih informatif dan membantu anak untuk belajar secara visual dengan melihat langsung letak kesalahan pada bacaan mereka.

H. Prosedur Pengujian Prototipe

Proses pengujian prototipe sistem koreksi bacaan doa harian dilakukan secara langsung kepada responden anak-anak. Pengujian melibatkan 8 orang partisipan anak dengan usia 6-12 tahun dari TPA An-Nahrawi Plosokuning. Setiap anak diminta untuk memilih satu doa yang ingin dibaca atau dihafalkan, kemudian suara direkam melalui mikrofon pada antarmuka prototipe. Seluruh hasil pengujian dicatat dalam tabel rekapitulasi yang memuat identitas anak (kode anak dan usia), doa yang diuji, hasil transkripsi, skor kemiripan, dan jenis kesalahan yang muncul. Pendekatan ini memastikan bahwa proses penilaian berlangsung konsisten dan objektif dengan menggambarkan performa prototipe dalam konteks penggunaan oleh anak-anak.

IV. HASIL DAN PEMBAHASAN

A. Hasil Pelatihan



Gambar 5. Grafik *Training Loss* dan *Validation Loss* per *Epoch*

Proses pelatihan model Whisper dari tarteel-ai/whisper-base-ar-quran dilakukan menggunakan GPU L4, yang mana model merupakan model *pre-trained* yang sudah dilatih pada *dataset* bahasa Arab Quran. Dalam proses pelatihan ini, model dilatih kembali dengan domain khusus yakni *dataset* bacaan doa anak yang telah dikumpulkan agar dapat mengenali karakteristik dan pola bacaan doa pada anak-anak. Gambar grafik *training loss* dan *validation loss* pada Gambar 5 menunjukkan bahwa pola pembelajaran model berjalan secara konsisten dan stabil selama proses pelatihan berlangsung tanpa *overfitting*.

Pada rentang *epoch* 0.3 sampai 0.6, terjadi penurunan *loss* yang sangat tajam, yakni dari sekitar 3.1 menjadi 0.86 pada *training loss* dan dari 1.8 menjadi 0.44 pada *validation loss*. Penurunan drastis ini menggambarkan adanya proses *rapid adaption*, di mana model secara cepat menyesuaikan diri terhadap pola dan karakteristik doa anak, seperti variasi pelafalan, intonasi, serta struktur fonetik yang ada. Selanjutnya, memasuki *epoch* 1 hingga 2, nilai *training loss* terus menurun ke rentang 0.1-0.3, sementara *validation loss* berada pada posisi yang stabil, yakni di rentang 0.2-0.3. Perbedaan kecil ini menunjukkan bahwa model tidak mengalami *overfitting* dan generalisasinya masih cukup baik terhadap data validasi.

Kemudian, pada *epoch* 2 hingga 3, nilai *loss* keduanya berada pada posisi yang cenderung stabil, yakni menyentuh angka 0.04 pada *training* dan ~0.2 pada *validation*. Hal ini menunjukkan bahwa model telah mencapai titik optimal berdasarkan *dataset* yang tersedia. Sehingga, penambahan *epoch* lebih banyak diprediksi tidak akan menghasilkan peningkatan performa yang signifikan. Melalui pelatihan ini, model Whisper terbukti sangat responsif terhadap domain khusus yang sempit seperti bacaan doa anak dan *dataset* kecil tetap efektif untuk proses *fine-tuning*.

B. Evaluasi Kinerja Model

Setelah proses pelatihan selesai, proses evaluasi kinerja model dilakukan secara objektif dengan memanfaatkan 32 data uji yang telah disiapkan dan belum dilibatkan sama sekali dalam proses pelatihan. Hal ini dilakukan untuk menguji kemampuan generalisasi model dalam mengenali variasi suara baru di luar data latih. Untuk mengukur efektivitas metode *fine-tuning*, evaluasi dilakukan dalam dua tahap: (1) Evaluasi model hasil *fine-tuning*, dan (2) Analisis komparatif terhadap model *baseline* (whisper-base-ar-quran versi asli yang tidak di *fine-tuning*).

1) Kinerja Model Setelah Fine-Tuning

TABEL V. HASIL EVALUASI KINERJA MODEL FINE-TUNED

Metrik	Nilai
Word Error Rate (WER)	0.271
Character Error Rate (CER)	0.059

Berdasarkan TABEL V, model menghasilkan WER sebesar 27.1%. Dalam konteks ASR untuk anak-anak, angka ini tergolong kompetitif dan dapat diterima. Namun, metrik CER yang rendah yakni 5.9% mengindikasikan bahwa kesalahan model secara garis besar bersifat minor, yakni pada tingkat karakter, yang mana kesalahan yang terjadi bukanlah kesalahan halusinasi kalimat oleh model, melainkan kesalahan huruf atau harakat dalam satu kata.

2) Analisis Komparatif Kinerja Model Baseline vs Fine-tuned

Untuk memvalidasi urgensi *fine-tuning*, evaluasi kinerja model dilakukan dengan membandingkan performa model whisper-base-ar-quran sebelum dan sesudah proses *fine-tuning* menggunakan *dataset* bacaan doa anak. Perbandingan merujuk pada dua metrik utama evaluasi, yakni WER dan CER. Hasil lengkap perbandingan dapat dilihat pada TABEL VI.

TABEL VI. PERBANDINGAN METRIK EVALUASI MODEL *BASELINE* VS *FINE-TUNED*

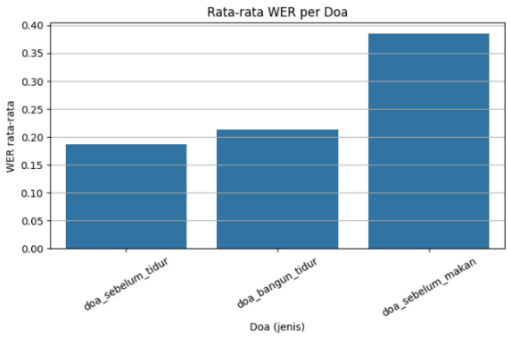
Metrik	Baseline	Fine-tuned	Peningkatan
WER	0.991	0.271	+72% lebih baik
CER	0.448	0.059	+38.9% lebih baik

Hasil eksperimen tersebut menunjukkan perbedaan performa yang sangat signifikan. Model *baseline* yang belum dilatih menggunakan *dataset* doa anak, memiliki performa yang sangat rendah dengan WER yang mendekati 1.0. Yang secara praktis menunjukkan bahwa model gagal dalam mengenali doa anak, ditandai oleh ketidaksesuaian hampir seluruh kata yang ditranskripsikan oleh model ASR dengan target atau *ground truth* dari transkrip asli audio. Hal ini dapat terjadi karena model awal tidak dibangun secara

spesifik untuk domain doa anak yang sangat berbeda dengan bacaan Al-Qur'an yang menjadi fokus pelatihan model aslinya.

Sebaliknya, setelah dilakukan proses *fine-tuning*, peningkatan performa terjadi secara signifikan pada setiap metrik yang ada. WER turun dari 0.991 menjadi 0.271, yang menandakan adanya peningkatan akurasi kata sebesar 72%. Penurunan CER dari 0.448 menjadi 0.059 juga menunjukkan bahwa kesalahan yang tersisa sebagian besar berupa perbedaan kecil pada huruf atau harakat, bukan kesalahan struktur kalimat secara keseluruhan. Temuan ini menunjukkan bahwa pendekatan *fine-tuning* merupakan langkah krusial dan efektif dalam mengadaptasi model ASR untuk kebutuhan domain spesifik seperti bacaan doa anak dan model yang telah di *fine-tune* sudah layak digunakan sebagai sistem koreksi bacaan doa anak dalam prototipe aplikasi.

3) Analisis Kinerja Berdasarkan Jenis Doa



Gambar 6. Rata-rata WER per jenis doa

Analisis lebih dalam dilakukan untuk melihat korelasi antara kompleksitas fonetik doa dengan tingkat kesalahan model. Evaluasi WER per doa yang ditunjukkan pada Gambar 6 menunjukkan pola korelasi yang jelas antara kesulitan pelafalan bagi anak dengan tingkat kesalahan model. Doa sebelum tidur dengan WER ~0.19, memiliki WER terendah. Hal ini dikarenakan doa ini memiliki struktur kalimat pendek dan minim huruf dengan makhraj dari tenggorokan, sehingga model lebih mudah melakukan prediksi yang akurat. Pada kasus doa bangun tidur, WER yang dihasilkan yakni ~0.21. Peningkatan WER ini disebabkan oleh kekeliruan model dalam mentranskripsikan karakter *Zai* (ز) atau *Dzal* (ذ) pada kata '*alladzii* (الَّذِيْ)' (model mentranskripsikan karakter *zai* ke karakter *dzal*). WER tertinggi terlihat pada jenis doa sebelum makan. WER yang dihasilkan yakni ~0.38, di mana model kesulitan untuk mentranskripsikan huruf-huruf tertentu dalam rangkaian kata tertentu seperti *Qaf* (ق) atau *Kaf* (ك) dalam rangkaian kata '*razaqtanaa* (رَزَقْتَنَا)' dan '*waqinaa* (وَقِنَا)', *Kha*' (خ) atau *Kaf* (ك) pada rangkaian kata '*baariklanaa* (بَارِكْ لَنَا)', *'Ain* (ع) atau *Alif* (أ) serta *Zai* (ز) atau *Dzal* (ذ) pada rangkaian kata ''*adzaaban-naar* (عَذَابَ النَّار)'. Kesalahan ini terjadi karena adanya ketidakstabilan artikulasi anak serta kesulitan mereka dalam melafalkan huruf-huruf sulit ini sehingga model menerjemahkannya sebagai *noise* dan berujung pada peningkatan kesalahan kata.

C. Hasil Pengujian Prototipe

Pengujian prototipe dilakukan untuk menilai performa sistem dalam mengenali dan mengevaluasi bacaan doa anak secara langsung. Pengujian dilaksanakan di TPA An-Nahrawi Plosokuning dengan melibatkan delapan orang anak. Setiap anak diminta untuk membaca satu doa yang dipilih di depan mikrofon, kemudian rekaman suara diolah oleh prototipe berbasis Streamlit yang telah diintegrasikan dengan model Whisper yang telah di *fine-tuning*.

Melalui pengujian ini, diperoleh gambaran bagaimana sistem koreksi bacaan doa anak ini dapat bekerja dalam kondisi nyata, termasuk kemampuan model dalam mengenali variasi suara dan pelafalan doa oleh anak, kualitas bacaan anak, serta karakter kesalahan yang biasa terjadi pada pelafalan doa oleh anak. Hasil pengujian dirangkum dalam TABEL VII, dengan identitas anak disamarkan demi menjaga privasi.

TABEL VII. TABEL REKAP PENGUJIAN PROTOTIPE PER DOA

No	Anak	Hasil Transkripsi	Status/ *Similarity*	Jenis Kesalahan
Doa Sebelum Makan				
1	A	اللَّهُمَّ بَرِكْ لَنَا فِيْمَا رَزَقْتَنَا وَقِنَا أَذَابَ النَّار	Benar/ 97.8%	1 karakter hilang, substitusi huruf (أ→ع)
2	B	اللَّهُمَّ بَرِكْ لَنَ فِيْمَا رَزَقْتَنَا وَقِنَا أَذَابَ النَّار	Benar/ 98.6%	2 karakter hilang
3	C	اللَّهُمَّ بَارِكْ لَنَا فِيْمَا رَزَقْتَنَا وَقِنَا عَذَابَ النَّار	Benar/ 100%	-
4	D	اللَّهُمَّ بَرِكْ لَنَا فِيْمَا رَزَقْتَنْ وَقِنَا عَذَابَ النَّار	Benar/ 98.6%	2 karakter hilang
Doa Sebelum Tidur				
5	E	بِسْمِكَ اللَّهُمَّ أَحْيَا وَ بِسْمِكَ أَمُوْتُ	Benar/ 100%	-
6	F	بِسْمِكَ اللَّهُمَّ أَحْيَ وَ بِسْمِكَ أَمُتْ وَ أَمُتْ	Perlu perbaikan/ 88.0%	2 karakter hilang, penyisipan kata lain
7	G	بِسْمِكَ اللَّهُمَّ أَحْيَ وَ بِسْمِكَ أَمُوْتُ	Benar/ 98.9%	1 karakter hilang
8	H	بِسْمِكَ اللَّهُمَّ أَحْيَ وَ بِسْمِكَ أَمُوْتُ وَ أَمُتْ	Perlu perbaikan/ 88.2%	1 karakter hilang, penyisipan kata lain
Doa Bangun Tidur				
9	C	الْحَمْدُ للهِ الَّذِيْ أَحْيَانَا بَعْدَ مَا أَمَاتَنَا وَإِلَيْهِ النُّشُوْرُ	Benar/ 96.3%	1 karakter hilang, penyisipan spasi
10	E	الْحَمْدُ للهِ الَّذِيْ أَحْيَانَا بَعْدَ مَا أَمَاتَنَا وَإِلَيْهِ النُّشُوْرُ	Benar/ 96.9%	penyisipan spasi

Hasil pengujian prototipe menunjukkan bahwa sistem mampu menjalankan fungsi perekaman, transkripsi, dan evaluasi bacaan doa anak secara efektif.

D. Pembahasan

Hasil pengujian model menjawab pertanyaan penelitian pertama, di mana metode *fine-tuning* terbukti efektif meningkatkan akurasi secara signifikan ditandai dengan penurunan WER sebesar 72%. Hal ini juga mengonfirmasi bahwa adaptasi domain sangat krusial untuk mengatasi ketidakstabilan artikulasi anak yang tidak dapat ditangani oleh model *baseline*. Terkait pertanyaan penelitian kedua, integrasi algoritma *Levenshtein* pada prototipe berhasil menerjemahkan selisih transkripsi menjadi umpan balik visual dengan penyorotan letak kesalahan, yang memungkinkan koreksi mandiri oleh anak tanpa memerlukan pendampingan intensif.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian dan pengujian sistem, dapat disimpulkan bahwa penerapan metode *fine-tuning* pada model Whisper terbukti secara efektif dapat mengatasi masalah *domain missmatch* pada suara anak. Keberhasilan ini ditandai dengan peningkatan performa yang signifikan, yaitu penurunan WER dari 0.991 pada model *baseline* menjadi 0.271, serta pencapaian CER sebesar 0.059. Meskipun akurasi secara umum meningkat, model masih menghadapi tantangan dalam mengenali huruf dengan *makhraj* yang berdekatan, terutama *Qaf* dan *'Ain*, yang terlihat dari tingginya tingkat kesalahan pada doa dengan artikulasi yang cukup sulit. Selain itu, implementasi prototipe aplikasi berbasis Streamlit yang diintegrasikan dengan algoritma *levenshtein distance* berhasil berfungsi degan baik dalam memberikan evaluasi korektif secara *real-time*, sehingga memudahkan pengguna untuk mengidentifikasi serta memahami letak kesalahan bacaan melalui skor kemiripan dan penyorotan visual.

Penelitian ini berkontribusi melalui: (1) Penyediaan *dataset hybrid* bacaan doa anak berbahasa Arab dengan dialek lokal; (2) Pengembangan model ASR Whisper yang dioptimalkan untuk domain anak; dan (3) Penerapan *framework* evaluasi berbasis *Levenshtein Distance* untuk umpan balik pedagogis. Namun, penelitian ini memiliki keterbatasan, terutama pada ukuran *dataset* (320 audio dengan total durasi ~42 menit) yang relatif kecil untuk standar *deep learning*. Hal ini membatasi generalisasi model terhadap variasi aksen yang lebih luas di luar wilayah pengambilan sampel.

B. Saran

Demi pengembangan sistem yang lebih optimal di masa mendatang, beberapa saran yang dapat diajukan untuk penelitian selanjutnya adalah sebagai berikut:

1. Menambah jumlah dan variasi data latih, khususnya pada pelafalan huruf-huruf yang serupa maupun huruf *halqiah* untuk meningkatkan sensitivitas model.
2. Mengembangkan sistem menjadi aplikasi seluler demi aksesibilitas yang lebih baik, serta menambahkan fitur deteksi hukum tajwid spesifik.
3. Memperluas cakupan *dataset* mencakup doa-doa harian lain agar aplikasi dapat mendukung pembelajaran yang lebih komprehensif.

REFERENSI

[1] Z. Pamuji, M. Rogib, A. Basit dan M. S. Yahya, "Implementation of Religious Culture to Develop Children's Character in Early Childhood Education," *Jurnal Pendidikan Usia DIni,* vol. 18, no. 1, pp. 81-98, 2024.

[2] D. A. L. E. Al Azhim dan L. N. Kholidah, "Problematika Pelafalan Huruf Hijaiyah pada Anak Usia Dini di Rhoudhotu Tarbiyatil Qur'an (RTQ) Al-Ghozali Tlogomas Malang," *JoLLA: Journal of Language, Literature, and Arts,* vol. 1, no. 1, pp. 62-75, 2021.

[3] A. M. Haryadi dan L. Marlina, "Phonetic Analysis of Makhārij al-Ḥurūf Pronunciation Errors in the Recitation of Surah Al-Fātiḥah by Grade VIII Students at SMPIT Nurul Yaqien," *Tatsqifiy: Jurnal Pendidikan Bahasa Arab,* vol. 6, no. 2, pp. 139-146, 2025.

[4] M. Ibadurrahman, A. Supriyadi, M. N. Solahudin, W. Ridwan, L. Nuraini, A. Solihat dan Andriyana, "Mobile Learning Applications for Islamic Studies: A Systematic Review of Design Principles and Learning Outcomes," *Journal of International Multidisciplinary Research,* vol. 3, no. 8, pp. 12-19, 2025.

[5] D. Khairani, T. Rosyadi, Arini, I. L. Rahmatullah dan F. F. Antoro, "Enhancing Speech-to-Text and Translation Capabilities for Developing Arabic Learning Games: Integration of Whisper OpenAI Model and Google API Translate," *JURNAL TEKNIK INFORMATIKA (JTI),* vol. 17, no. 2, pp. 203-212, 2024.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey dan I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022.

[7] B. Talafha, A. Waheed dan M. Abdul-Mageed, "N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition," *INTERSPEECH,* pp. 5092-5096, August 2023.

[8] V. Timmel, C. Paonessa, M. Vogel, D. Perruchoud dan R. Kakooe, "Fine-tuning Whisper on Low-Resource Languages for Real-World Applications," *CC BY 4.0,* pp. 1-7, December 2024.

[9] A. Akbar, A. Y. Husodo dan A. Zubaidi, "Implementasi Google Speech API Pada Aplikasi Koreksi Hafalan Al-Qur'an Berbasis Android," *Jurnal Teknologi Informas, Komputer, dan Aplikasinya (JTIKA),* vol. 1, no. 1, pp. 1-8, Maret 2019.

[10] M. Assisi, A. Septiarini, A. H. Kridalaksana dan M. Wati, "Rancang Bangun Aplikasi Hafalan Al-Quran dengan Google Speech API Berbasis Android," *Jurnal Rekayasa Teknologi Informasi (JURTI),* vol. 6, no. 1, pp. 26-35, 2022.

[11] R. Hadiyansah dan R. Andamira, "Convolutional Neural Network (CNN) for Detecting Al-Qur'an Reciting and Memorizing," *Khazanah Journal of Religion and Technology,* vol. 1, no. 2, pp. 44-48, 2023.

[12] O. Mohamed dan S. A. Aly, "Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset," *arXiv:2110.04425,* 2021.

[13] A. Baevski, H. Zhou, A. Mohamed dan M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," pp. 1-12, June 2020.

[14] S. M. Isa, "Speech Recognition," Bina Nusantara University, 8 May 2019. [Online]. Available: https://mti.binus.ac.id/2019/05/08/speech-recognition/.

[15] M. Malik, M. K. Malik, K. Mehmood dan I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications,* no. 80, pp. 9411-9457, 2021.

[16] A. Graves, A.-r. Mohamed dan G. Hinton, "Speech Recognition With Deep Recurrent Neural Networks," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 6645-6649, 2013.

[17] Open AI, "Introducing Whisper," Open AI, 21 September 2022. [Online]. Available: https://openai.com/id-ID/index/whisper/.

[18] S. Khurana, N. Moritz, T. Hori dan J. L. Roux, "Unsupervised Domain Adaption for Speech Recognition via Uncertainty Driven Self-Training," *arXiv:2011.13439v2,* 2021.

[19] M. A. Mensah, I. Wiafe, A. Ekpezu, J. K. Appati, D. Abdulai, A. N. Wiafe-Akenten, F. E. Yeboah dan G. Odame, "Benchmarking Akan ASR Models Across Domain-Specific Datasets: A Comparative Evaluation of Performance, Scalability, and Adaptability," *arXiv:2507.02407,* 2025.

[20] T. Raissi, N. Rossenbach dan R. Schluter, "Analysis of Domain Shift across ASR Architectures via TTS-Enabled Separation of Target Domain and Acoustic Conditions," *arXiv:2508.09868,* 2025.

[21] S. P. Dubagunta, S. H. Kabil dan M. Magimai.-Doss, "Improving Children Speech Recognition Through Feature Learning from RAW Speech Signal," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5736-5740, 2019.

[22] P. G. Shivakumar dan P. Georgiou, "Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations," *arXiv:1805.03322,* 2018.

[23] M. U. Sadiq dan M. M. Yousaf, "Distributed Algorithm for Parallel Edit Distance Computation," *Computing and Informatics,* vol. 39, no. 4, pp. 757-779, 2021.