Action Research Literate
Vol. 9, No. 2, Februari 2025
ISSN: 2808-6988

SPEECH RECOGNITION DENGAN WHISPER DALAM BAHASA INDONESIA

Ezra William[1], Amalia Zahra[2]
Universitas Bina Nusantara, Indonesia
Email: ezra.william@binus.ac.id, amalia.zahra@binus.edu

ABSTRAK

Perkembangan teknologi kecerdasan buatan telah mendorong kemajuan dalam pengenalan suara (speech recognition), terutama dalam mendukung komunikasi digital yang lebih efisien. Salah satu model terbaru yang banyak digunakan adalah Whisper, yang dikembangkan oleh OpenAI dengan kemampuan pengenalan suara multibahasa yang diklaim memiliki akurasi tinggi. Namun, tantangan utama dalam implementasi teknologi ini di Indonesia adalah keterbatasan sumber daya data dalam bahasa lokal serta variasi aksen yang signifikan. Oleh karena itu, penelitian ini dilakukan untuk mengevaluasi kinerja model Whisper dalam mengenali dan mentranskripsi suara berbahasa Indonesia. Penelitian ini bertujuan untuk menganalisis tingkat akurasi Whisper dalam pengenalan ucapan bahasa Indonesia berdasarkan Word Error Rate (WER) serta membandingkannya dengan model XLS-R dan XLSR-53. Metode yang digunakan dalam penelitian ini adalah pendekatan komparatif dengan melakukan fine-tuning terhadap model Whisper menggunakan dataset Common Voice 13 dalam bahasa Indonesia. Evaluasi model dilakukan dengan mengukur WER pada tahap pelatihan dan pengujian. Hasil penelitian menunjukkan bahwa model Whisper memiliki performa terbaik dibandingkan model XLS-R dan XLSR-53 dalam mengenali ucapan bahasa Indonesia. Nilai WER Training yang diperoleh adalah 22.33505%, sedangkan nilai WER Testing adalah 19.774909%. Hal ini menunjukkan bahwa model Whisper lebih unggul dalam menangani variasi aksen dan kondisi akustik dibandingkan dengan model lainnya. Keunggulan ini terutama disebabkan oleh pelatihan berbasis data yang lebih besar serta kemampuan adaptasi model terhadap berbagai bahasa. Implikasi penelitian ini memberikan kontribusi dalam pengembangan teknologi speech recognition berbahasa Indonesia serta meningkatkan aksesibilitas bagi pengguna dalam berbagai sektor, seperti pendidikan, layanan publik, dan teknologi komunikasi.

Kata Kunci: Speech Recognition, Speech-to-text, Whisper, Common Voice 13.

ABSTRACT

The development of artificial intelligence technology has driven progress in speech recognition, especially in supporting more efficient digital communication. One of the latest models that is widely used is Whisper, developed by OpenAI with multilingual speech recognition capabilities that are claimed to have high accuracy. However, the main challenge in implementing this technology in Indonesia is the limited data resources in local languages and significant accent variations. Therefore, this study was conducted to evaluate the performance of the Whisper model in recognizing and transcribing Indonesian speech. This study aims to analyze the level of accuracy of Whisper in recognizing Indonesian speech based on the Word Error Rate (WER) and compare it with the XLS-R and XLSR-53 models. The method used in this study is a comparative approach by fine-tuning the Whisper model using the Common Voice 13 dataset in Indonesian. Model evaluation is carried out by measuring WER at the training and testing stages. The results of the study show that the Whisper model has the best performance compared to the XLS-R and XLSR-53 models in recognizing Indonesian speech. The WER Training value obtained was 22.33505%, while the WER Testing value was 19.774909%. This shows that the Whisper model is superior in handling accent variations and acoustic conditions compared to other models. This superiority is mainly due to larger data-based

training and the model's adaptability to various languages. The implications of this research contribute to the development of Indonesian speech recognition technology and increase accessibility for users in various sectors, such as education, public services, and communication technology.

Keywords: Speech Recognition, Speech-to-text, Whisper, Common Voice 13

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (AI) telah membawa dampak besar dalam berbagai aspek kehidupan manusia, salah satunya dalam bidang pengenalan ucapan atau speech recognition. Teknologi ini telah mengalami kemajuan pesat dalam beberapa dekade terakhir, dengan implementasi yang luas di berbagai sektor seperti asisten virtual, layanan kesehatan, pendidikan, hingga sistem keamanan. Salah satu model terbaru yang menarik perhatian dalam dunia speech recognition adalah OpenAI Whisper. Model ini dikembangkan dengan pendekatan pembelajaran skala besar dan diklaim mampu mengenali berbagai bahasa dengan akurasi tinggi (Radford et al., 2022). Namun, tantangan utama dalam pengembangan teknologi speech recognition adalah bagaimana model ini dapat bekerja secara optimal dalam berbagai kondisi lingkungan, aksen, serta bahasa yang memiliki sumber daya terbatas (Gong et al., 2023).

Selain itu, isu yang berkaitan dengan aksesibilitas dan keberpihakan dalam teknologi speech recognition juga menjadi perhatian utama. Sebagian besar model yang ada saat ini lebih banyak dikembangkan dalam bahasa Inggris, sementara bahasa dengan jumlah penutur besar seperti Bahasa Indonesia masih memiliki keterbatasan dalam data pelatihan dan akurasi hasil pengenalan suara (Arisaputra & Zahra, 2022). Oleh karena itu, penelitian yang berfokus pada pengembangan speech recognition dalam bahasa Indonesia menjadi semakin penting.

Dalam konteks Indonesia, pengenalan suara memiliki tantangan tersendiri yang mencakup variasi aksen, dialek, dan gaya berbicara yang beragam. Whisper sebagai model terbaru diklaim mampu mengatasi berbagai tantangan ini dengan pelatihan berbasis data besar yang mencakup berbagai bahasa (Pratama & Amrullah, 2024). Namun, implementasi model ini dalam konteks lokal masih perlu dievaluasi lebih lanjut, terutama dalam hal tingkat kesalahan kata atau Word Error Rate (Yakubovskyi & Morozov, 2023).

Selain itu, salah satu penerapan spesifik yang telah diuji adalah penggunaan Whisper dalam transkripsi debat politik di Indonesia (Khoiroh et al., 2024). Studi ini menunjukkan bahwa meskipun Whisper memiliki performa yang baik dalam kondisi ideal, model ini masih mengalami kesulitan dalam menangkap variasi intonasi dan gaya bicara formal yang sering digunakan dalam konteks debat. Hal ini menunjukkan bahwa meskipun teknologi ini memiliki potensi besar, masih ada tantangan yang perlu diatasi untuk meningkatkan akurasi dalam bahasa Indonesia.

Dengan semakin meningkatnya kebutuhan akan teknologi speech recognition yang dapat bekerja secara optimal dalam bahasa Indonesia, penelitian ini menjadi sangat penting. Salah satu faktor utama yang mendorong urgensi penelitian ini adalah kebutuhan akan sistem transkripsi otomatis yang dapat membantu berbagai sektor, seperti pendidikan, jurnalistik, dan administrasi publik. Teknologi ini dapat membantu meningkatkan efisiensi kerja serta memberikan aksesibilitas lebih baik bagi individu dengan keterbatasan fisik, seperti tuna rungu (Saputra, 2023).

Selain itu, dengan semakin banyaknya perangkat lunak dan aplikasi yang mengandalkan pengenalan suara, pengembangan model yang lebih akurat dan sesuai dengan

kebutuhan lokal menjadi semakin mendesak (Ferdiansyah & Aditya, 2024). Whisper, sebagai salah satu model terbaru dalam teknologi speech recognition, memiliki potensi besar untuk menjawab tantangan ini, tetapi masih memerlukan penelitian lebih lanjut untuk memastikan bahwa model ini dapat beradaptasi dengan baik terhadap konteks bahasa Indonesia.

Berbagai studi telah dilakukan terkait pengenalan ucapan dalam bahasa Indonesia. Arisaputra dan Zahra (2022) mengembangkan model speech recognition berbasis XLSR-53 dan menemukan bahwa meskipun model ini memiliki tingkat akurasi yang cukup tinggi, masih terdapat kendala dalam menangkap variasi aksen dan kejelasan artikulasi. Sementara itu, (Ardila et al., 2019) melalui proyek Common Voice berusaha mengumpulkan data suara dari berbagai bahasa, termasuk bahasa Indonesia, untuk meningkatkan akurasi model speech recognition. Namun, data yang tersedia masih terbatas dibandingkan dengan bahasa Inggris.

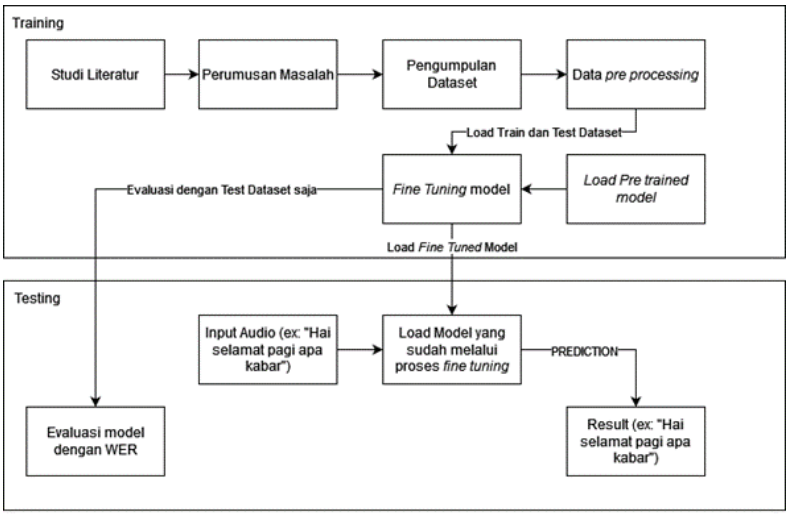
Di sisi lain, penelitian yang dilakukan oleh (Rifaldi & Cahyono, 2024) menunjukkan bahwa penggunaan Whisper dalam aplikasi notulensi dapat meningkatkan efisiensi transkripsi dalam lingkungan akademik dan profesional. Selain itu, studi yang dilakukan oleh Rekimoto (2023) menunjukkan bahwa Whisper juga dapat digunakan untuk mengubah ucapan berbisik menjadi suara normal, yang sangat berguna dalam konteks komunikasi yang lebih inklusif.

Penelitian ini berusaha memberikan pembaruan dalam studi terkait teknologi speech recognition dalam bahasa Indonesia dengan beberapa kontribusi utama. Pertama, penelitian ini akan mengevaluasi performa Whisper dalam berbagai kondisi lingkungan dan aksen di Indonesia. Kedua, penelitian ini akan membandingkan hasil transkripsi Whisper dengan model lainnya yang telah digunakan dalam bahasa Indonesia, seperti XLSR-53 (Cahyawijaya et al., 2022). Ketiga, penelitian ini akan menguji bagaimana model Whisper dapat diadaptasi untuk meningkatkan akurasi pengenalan ucapan dalam bahasa Indonesia dengan pendekatan pelatihan berbasis data lokal (Peng et al., 2023).

Penelitian ini bertujuan untuk Menganalisis tingkat akurasi Whisper dalam pengenalan ucapan bahasa Indonesia berdasarkan Word Error Rate (Yakubovskyi & Morozov, 2023). Mengevaluasi kinerja Whisper dalam berbagai kondisi lingkungan dan aksen bahasa Indonesia (Pratama & Amrullah, 2024). Membandingkan hasil transkripsi Whisper dengan model lainnya seperti XLSR-53 (Arisaputra & Zahra, 2022). Mengidentifikasi tantangan dan peluang dalam penerapan Whisper untuk kebutuhan transkripsi otomatis dalam bahasa Indonesia (Khoiroh et al., 2024). Penelitian ini diharapkan dapat memberikan kontribusi bagi pengembangan teknologi speech recognition dalam bahasa Indonesia serta meningkatkan aksesibilitas teknologi ini di berbagai sektor. Dengan demikian, hasil penelitian ini dapat menjadi dasar bagi pengembangan sistem yang lebih canggih dan sesuai dengan kebutuhan pengguna lokal.

METODE PENELITIAN

Metode penelitian yang dipakai dalam penelitian adalah komparasi model, dimana model yang akan dibandingkan adalah model Whisper, XLSR, dan XLSR-53.



Gambar 1. Tahapan Penelitian

Untuk Tahapan penelitian secara umum dilakukan oleh peneliti dari Tahap Studi Literatur, dimana penulis mempelajari Studi yang telah dilakukan di dalam topik "Speech Recognition". Kemudian, penelitian ini merumuskan masalah apa saja yang masih belum terpecahkan, dilanjutkan dengan proses pengumpulan dataset yang akan digunakan untuk proses fine tuning. Kemudian, penulis melakukan tahap Data Pre processing, dimana dataset yang sudah didapatkan diolah sehingga optimal untuk input dataset di tahap Fine Tuning. Di dalam tahap Fine Tuning, model akan melalui proses Fine Tuning dengan sumber input dataset berasal dari hasil output tahap Data Pre processing di tahap sebelumnya. Hasil output dari tahap Fine Tuning menghasilkan model yang sudah melalui tahap Fine Tuning, dimana model tersebut sudah belajar dari input dataset yang sudah disiapkan di tahap Data Pre processing. Penelitian menggunakan Dataset Common Voice 13 dalam bahasa Indonesia melakukan fine tuning di semua model Speeech Recognition di dalam penelitian ini. Untuk jumlah transkripsi dataset Common Voice dalam Bahasa Indonesia secara keseluruhan ada 3616 baris.

Dataset Common Voice 13 dalam bahasa Indonesia dibagi menjadi 3 bagian, yaitu dataset train, test, dan validation. Untuk format data audio yang ada di dalam Common Voice 13 adalah ".MP3" dengan sampling rate sebesar 48 KHz. Untuk jumlah data audio di ketiga bagian tersebut serta total waktu data audio dari masing masing bagian tersebut dijabarkan dalam tabel 2.

Dalam tahap Data Preprocessing akan dilakukan pembagian dataset dengan persentase data training sebesar 60% dan data testing sebesar 20% dan data validation sebesar 20% dari dataset Common Voice 13 dalam bahasa Indonesia. Komposisi data di dataset Common Voice 13 dalam bahasa Indonesia sudah dilakukan pembagian secara merata, dengan rincian yang dijelaskan dalam table 2, di mana penulis akan mengambil data audio di dalam bagian data training, data testing, dan data validation dengan jumlah berdasarkan tabel 2.

Untuk preprocessing dataset Common Voice 13 dalam bahasa Indonesia dilakukan dengan menghilangkan secara spesifik karakter simbol dalam bentuk teks di bagian

transcript dataset yang dihilangkan menggunakan Bahasa pemrograman Python dengan kode:“chars_to_ignore_regex='[\,\?\.\!\\;\:\"\“\%\‘\”\□\'\$\&\+\/\¡\£\¥\§\©\\xad\°\±\²\³\µ\¶\º\ ã\ä\å\ÿ\ƒ\•\€]' ”.

Di dalam tahap ini, langkah pertama adalah melakukan proses loading pre-trained model Speech recognition dari model Whisper. Setelah proses loading pre trained model Speech recognition dari model Whisper sudah dilakukan, penulis melakukan proses fine tuning di model Speech recognition dari model Whisper terhadap dataset Common Voice 13 dalam bahasa Indonesia bagian train dan bagian validation yang sudah melalui tahap data preprocessing. Pertama tama, dataset Common Voice 13 dalam bahasa Indonesia bagian train dan bagian validation yang sudah melalui tahap data preprocessing diproses dengan menggunakan library XLS-R dengan nama “Wav2Vec2Processor” sebagai processor, dan “Wav2Vec2ForCTC” sebagai model.

Untuk fungsi dari “Wav2Vec2Processor” adalah sebagai Feature Extractor dan Tokenizer dalam satu function. Untuk fungsi dari “Wav2Vec2forCTC” adalah sebagai Language Model dengan basis dasar dari model menggunakan Connectionist Temporal Classification (CTC). Setelah tahap sebelumnya sudah dilakukan, maka tahap selanjutnya adalah proses fine tuning model Model Whisper dengan dataset yang sudah diproses. Sebelum masuk ke dalam proses fine tuning model Whisper, hyperparameter tuning perlu dilakukan. Fungsi dari hyperparameter tuning adalah tempat pengaturan bagaimana fine tuning akan dilakukan di dalam sebuah model.

Setelah melakukan proses fine tuning atau hyperparameter tuning, proses evaluasi dibagi menjadi 2, yaitu evaluasi model Speech recognition dari training dan validation dataset, dan evaluasi model Speech recognition dari testing dataset. Metode evaluasi di dalam model Speech recognition dilakukan dengan metode WER (Word Error Rate), di mana jika variabel dari WER semakin kecil, semakin bagus model yang dihasilkan dari data yang diprediksi. Dalam evaluasi model Speech recognition dari training dan validation dataset, WER (Word Error Rate) yang dihasilkan didapatkan dari hasil belajar model terhadap training dan validation dataset, dimana model bisa menghasilkan output yang berbeda jika dievaluasi kembali menggunakan dataset yang berbeda, yaitu testing dataset. Inilah tujuan adanya evaluasi model Speech recognition dari testing dataset, yakni untuk membuktikan bahwa WER (Word Error Rate) yang dihasilkan dalam evaluasi model Speech recognition dari training dan validation dataset sesuai dengan kondisi evaluasi di dataset yang berbeda.

Berdasarkan Gambar 1, Tahap evaluasi model Speech recognition dari training dan validation dataset dilakukan di fase proses fine tuning model. Sedangkan evaluasi model Speech recognition dari testing dataset dilakukan setelah proses fine tuning model sudah selesai dilakukan, yang juga menandakan bahwa proses fine tuning model sudah selesai dilakukan. Berpindah ke evaluasi model Speech recognition dari testing dataset, evaluasi dimulai dari proses loading model yang sudah melalui proses fine tuning model, lalu model akan mencoba untuk mengolah input audio agar melihat seberapa jauh kemampuan model dalam mengubah input audio berupa suara menjadi kata kata.

Word Error Rate yang dihasilkan dalam proses ini lalu akan dibandingkan dengan WER yang dihasilkan dari evaluasi model Speech recognition dari training dan validation dataset untuk membuktikan bahwa model yang sudah melalui proses fine tuning tidak mengalami overfitting. Untuk metode penghitungan WER Training adalah dengan menggunakan output dari hasil proses fine tuning model, sedangkan metode penghitungan

WER Testing adalah dengan dengan mengevaluasi model dengan setiap file audio di dalam testing dataset, dimana dalam testing dataset memuat 3649 file audio. Setiap model mencoba untuk mengubah file audio di dalam testing dataset, maka WER (Word Error Rate) akan diukur, sehingga mengetahui seberapa jauh model dalam mempelajari bagaimana mengubah suara dari file audio menjadi teks dalam Bahasa Indonesia. Proses ini akan berlangsung hingga semua file audio sudah diproses. Jika semua file audio sudah diproses, maka selanjutnya adalah mendapatkan hasil WER rata rata dari WER setiap file audio. Proses ini dilakukan dengan menjumlahkan semua WER setiap file audio, lalu dibagi dengan total jumlah setiap file audio, yakni 3649 file audio.

Dengan cara ini, maka WER Testing bisa didapatkan dari proses evaluasi model Speech recognition dari testing dataset. Lalu, membandingkan performa model yang sudah melalui proses fine tuning dengan model yang berbeda, yaitu performa model Whisper dengan model XLS-R dan model XLSR-53. Tujuan penelitian melakukan prosess perbandingan ini adalah untuk menentukan manakah model terbaik di antara ketiga model ini dalam evaluasi model Speech recognition dengan menggunakan testing dataset sebagai dataset.

Salah satu cara untuk mengukur peningkatan kinerja model adalah melalui penggunaan metrik seperti Word Error Rate (WER). Metrik ini membandingkan Output Speech Recognition Model dengan referensi yang ada di Transkrip dan menghitung persentase dari kata kata yang mengalami kesalahan prediksi. Tingkat kesalahan yang lebih rendah menunjukkan kinerja yang lebih baik. Metrik untuk menilai kualitas model pengenalan ucapan adalah tingkat kesalahan kata [4]. Parameter ini dimasukkan ke dalam rumus

$$WER = (S + D + I) / N \quad (1)$$

Dengan keterangan dimana S adalah jumlah substitusi, atau variabel di mana sebuah kata diganti. D adalah jumlah penghapusan, atau variabel di mana sebuah kata dihilangkan dari transkrip. I adalah jumlah penyisipan, atau variabel di mana sebuah kata ditambahkan ke dalam kalimat. Sedangkan N adalah jumlah kata dalam referensi. Dengan metode penelitian ini, diharapkan dapat diperoleh hasil yang valid dan dapat diandalkan dalam menentukan model terbaik untuk pengenalan suara dalam bahasa Indonesia.

HASIL DAN PEMBAHASAN

Dalam pembuatan model yang penulis lakukan, penulis menggunakan bahasa pemrograman Python yang dijalankan di Google Colab dengan menggunakan GPU T4 dan Ram sebesar 15 GB. Penelitian ini memakai library yang tersedia di Google Colab tersebut seperti Transformers, sklearn, huggingface_hub, dan sebagainya. Untuk pembuatan model yang sudah dilakukan, model Speech recognition dari model Whisper menggunakan metode fine tuning di dalam library transformers dan huggingface. Untuk metode fine tuning, penulis menggunakan Tokenizer dan Feature Extractor dari library transformers. Untuk pengaturan hyperparameter tuning yang dilakukan oleh penulis dalam membuat model Whisper dengan hasil paling optimal dipaparkan dalam Tabel 1.

Tabel 1. Tabel Argument Dari Fine Tuning Training Arguments Dan Nilai Yang Diterapkan

No	*Hyperparameter*	Nilai
1	*group_by_length*	*False*
2	*per_device_train_batch_size*	16

3	*evaluation_strategy*	*Steps*
4	*gradient_checkpointing*	*True*
5	*fp16*	*True*
6	*save_steps*	1000
7	*logging_steps*	25
8	*learning_rate*	1e-5
9	*weight_decay*	0

Sumber: data diperoleh

Dalam penelitian ini, ada 2 tipe evaluasi model yaitu evaluasi model Speech recognition dari training dan validation dataset, dan evaluasi model Speech recognition dari testing dataset. Pertama tama, penulis melakukan evaluasi model Speech recognition dari training dan validation dataset. Untuk hasil evaluasi model Speech recognition dari training dan validation dataset yang sudah dilakukan oleh penulis berdasarkan hyperparameter yang sudah dijabarkan oleh penulis dipaparkan dalam Tabel 2.

Tabel 2. Hasil Model Speech Recognition Yang Diterapkan Di Model Whisper

Speech Recognition Model	*Dataset*	*Steps*	*Training Loss (%)*	*Validation Loss (%)*	*Word Error Rate (%)*
Whisper	*Common Voice 13*	1000	0.043800	0.361127	22.585492
		2000	0.005700	0.411267	22.362370
		3000	0.001500	0.436557	22.312281
		4000	0.001100	0.444734	**22.335048**

Sumber: data diperoleh

Dari hasil evaluasi model Speech recognition dari training dan validation dataset, model Speech recognition dari model Whisper berhasil mencapai nilai WER Training sebesar 22.33505%. Hasil WER Training ini adalah WER terbaik yang berhasil didokumentasikan oleh penulis dalam proses evaluasi model Speech recognition dari training dan validation dataset. Untuk Model berikut yang menjadi model pembanding adalah model XLS-R. Untuk pengaturan hyperparameter tuning yang dilakukan oleh penulis dalam membuat model XLS-R dengan hasil paling optimal dipaparkan dalam Tabel 3.

Tabel 3. Tabel Argument Dari Fine Tuning Training Arguments Dan Nilai Yang Diterapkan Di Model XLS-R

No	*Hyperparameter*	Nilai
1	*group_by_length*	*True*
2	*per_device_train_batch_size*	16
3	*evaluation_strategy*	*Epoch*
4	*num_train_epochs*	12
5	*gradient_checkpointing*	*True*
6	*fp16*	*True*
7	*save_steps*	500
8	*logging_steps*	100
9	*learning_rate*	1e-4

10	*weight_decay*	0.01
11	*save_total_limit*	5

Sumber: data diperoleh

Untuk pengaturan di dalam tabel 5 tidak jauh berbeda dengan Tabel 3, akan tetapi, model Whisper dan model XLS-R secara struktur berbeda, maka beberapa penyesuaian perlu dilakukan. Salah satu penyesuaian yang dilakukan adalah dengan mengubah learning rate menjadi 1e-4 di hyperparameter tuning model XLS-R. Untuk penyesuaian lainnya terletak pada evaluation strategy, dimana metode yang dipakai di Whisper adalah steps, sedangkan evaluation strategy yang dipakai di model XLS-R adalah Epoch. Untuk variabel num_train_epochs dalam hyperparameter tuning model XLS-R juga disesuaikan menjadi 12. Untuk performa model XLS-R saat proses evaluasi model Speech recognition dari training dan validation dataset terpapar di Tabel 4.

Tabel 4. Hasil Model Speech Recognition Yang Diterapkan Di Model XLS-R

Speech Recognition Model	*Dataset*	*Epoch*	*Training Loss (%)*	*Validation Loss (%)*	*Word Error Rate (%)*
XLS-R	*Common Voice 13*	1	4.535700	2.968087	1.000000
		2	2.932400	2.861846	1.000000
		3	2.870800	2.460367	1.000000
		4	2.460500	0.838463	0.717077
		5	0.571200	0.691731	0.691341
		6	0.423500	0.621622	0.612080
		7	0.350800	0.594245	0.576550
		8	0.295300	0.604900	0.570993
		9	0.254300	0.631946	0.560470
		10	0.198500	0.632732	0.554366
		11	0.181900	0.622929	0.541247
		12	0.174500	0.626638	**0.541019**

Sumber: data diperoleh

Dari hasil evaluasi model Speech recognition dari training dan validation dataset, model Speech recognition dari model XLS-R berhasil mencapai nilai WER Training sebesar 54.1019%, dimana hasil ini adalah hasil terbaik dari beberapa kali percobaan yang sudah dilakukan oleh penulis. Untuk Model selanjutnya yang menjadi model pembanding adalah model XLSR-53. Untuk pengaturan hyperparameter tuning yang dilakukan oleh penulis dalam membuat model XLSR-53 dengan hasil paling optimal dipaparkan dalam Tabel 5.

Tabel 5. Tabel Argument Dari Fine Tuning Training Arguments Dan Nilai Yang Diterapkan Di Model XLSR-53

No	*Hyperparameter*	Nilai
1	*group_by_length*	*True*
2	*per_device_train_batch_size*	16
3	*evaluation_strategy*	*Epoch*
4	*num_train_epochs*	12
5	*gradient_checkpointing*	*True*
6	*fp16*	*True*
7	*save_steps*	500
8	*logging_steps*	100
9	*learning_rate*	1e-4
10	*weight_decay*	0.01
11	*save_total_limit*	5

Sumber: data diperoleh

Untuk pengaturan di dalam tabel 5 tidak jauh berbeda dengan Tabel 5, dimana model XLS-R dan model XLSR-53 tidak jauh berbeda. Untuk penyesuaian yang dilakukan adalah variabel num_train_epochs dalam hyperparameter tuning model XLS-R yang disesuaikan menjadi 10. Untuk performa model XLS-R saat proses evaluasi model Speech recognition dari training dan validation dataset terpapar di Tabel 6.

Tabel 6. Hasil Model Speech Recognition Yang Diterapkan Di Model XLSR-53

Speech Recognition Model	*Dataset*	*Epoch*	*Training Loss (%)*	*Validation Loss (%)*	*Word Error Rate (%)*
XLSR-53	*Common Voice 13*	1	5.447900	2.974142	1.000000
		2	2.954300	2.929660	1.000000
		3	2.930600	2.911203	1.000000
		4	2.921600	2.907078	1.000000
		5	2.896800	2.871307	1.000000
		6	2.882200	2.844639	1.000000
		7	2.842100	2.515719	1.000000
		8	2.576300	1.577951	0.996447
		9	1.944900	0.986443	0.813192
		10	1.039800	0.856505	0.734843
		11	0.916200	0.794065	0.704323
		12	0.890900	0.789579	**0.697900**

Sumber: data diperoleh

Dari hasil evaluasi model Speech recognition dari training dan validation dataset, model Speech recognition dari model XLSR-53 berhasil mencapai nilai WER Training

sebesar 69.79%. Penulis juga melakukan evaluasi model Speech recognition dari testing dataset, dimana performa model ditentukan dari seberapa bagus model melakukan prediksi dari input Audio yang bersumber dari sub bagian Test di dataset Common Voice. Penulis juga melakukan proses evaluasi model Speech recognition dari testing dataset dengan membandingkan model Whisper, XLS-R, dan XLSR-53 untuk menentukan model yang terbaik di antara ketiga model tersebut melalui evaluasi model Speech recognition dari testing dataset. Proses dari evaluasi model Speech recognition dari testing dataset pertama tama dimulai dengan mengevaluasi model dengan setiap file audio di dalam testing dataset, dimana dalam testing dataset memuat 3649 file audio, seperti yang sudah dijabarkan di tabel 3.2. Setiap model mencoba untuk mengubah file audio di dalam testing dataset, maka WER (Word Error Rate) akan diukur, sehingga penulis mengetahui seberapa jauh model dalam mempelajari bagaimana mengubah suara dari file audio menjadi teks daalam Bahasa Indonesia.

Proses ini akan berlangsung hingga semua file audio sudah diproses. Jika semua file audio sudah diproses, maka selanjutnya adalah mendapatkan hasil WER rata rata dari WER setiap file audio. Proses ini dilakukan dengan menjumlahkan semua WER setiap file audio, lalu dibagi dengan total jumlah setiap file audio, yakni 3649 file audio. Dengan cara ini, maka WER Testing bisa didapatkan dari proses evaluasi model Speech recognition dari testing dataset. Hasil dari evaluasi dan perbandingan model yang sudah dilakukan oleh penulis berdasarkan evaluasi model Speech recognition dari training dan validation dataset dalam bentuk WER Training, dan evaluasi model Speech recognition dari testing dataset dalam bentuk WER Testing dipaparkan dalam Tabel 7.

Tabel 7. Perbandingan Model Pre Trained

No	Model	GPU	WER Training	WER Testing
1	model *Whisper*	T4	**22.33505%**	**19.774909%**
2	model XLS-R	T4	54.1019%	56.347186%
3	model XLSR-53	A100	69.79%	71.910376%

Sumber: data diperoleh

Terpaparkan di Tabel 7, hasil yang didapatkan di dalam model Whisper terbukti menghasilkan angka WER lebih baik daripada model XLS-R dan model XLSR-53 saat diterapkan proses fine tuning model. Untuk hasil terbaik nomor satu dari model yang telah melalui proses fine tuning berdasarkan evaluasi WER Testing adalah model Whisper, dengan pencapaian nilai WER Testing sebesar 19.774909%. Penulis perlu melakukan beberapa kali eksperimen di proses fine tuning model karena ada kemungkinan bahwa hasil dari performa model bisa beragam, meskipun hyperparameter dari masing masing proses fine tuning model memiliki nilai yang sama. Untuk penelitian ini, dimana eksperimen yang dilakukan menggunakan model Whisper dalam dataset Bahasa Indonesia, hasil dari penelitian ini masih mempunyai ruang untuk pengembangan.

Dengan demikian, maka dapat disimpulkan bahwa model Whisper adalah model terbaik dari segi WER Training sebesar 22.33505% dan WER Testing sebesar 19.774909%. Untuk penyebab Model Whisper lebih unggul terletak pada kelebihan model Whisper, yakni Skala Data Training yang besar, dimana model Whisper dilatih dengan 680,000 jam transkripsi audio dengan karakteristik multilingual dan beragam tujuan, salah satu dari tujuan tersebut adalah Speech Recognition. Dengan latar belakang Skala Data Training Multilingual, maka model Whisper memiliki kemampuan untuk mengenali lebih dari 1 bahasa, salah satunya adalah Bahasa Indonesia. Model Whisper juga termasuk model yang

bisa diandalkan dalam beragam tujuan, khususnya Speech Recognition jika dibandingkan dengan model lain. Model Whisper juga memiliki akurasi yang mirip dengan kemampuan manusia pada umumnya.

KESIMPULAN

Model speech recognition Whisper memiliki performa terbaik dibandingkan dengan model XLS-R dan XLSR-53 dalam pengenalan suara bahasa Indonesia. Hal ini dibuktikan dengan nilai Word Error Rate (WER) yang lebih rendah pada model Whisper, baik dalam tahap pelatihan (22.33505%) maupun pengujian (19.774909%). Keunggulan Whisper terutama disebabkan oleh skala data pelatihannya yang lebih besar, karakteristik multibahasa, dan kemampuannya dalam menangani variasi aksen serta kondisi akustik yang berbeda.Penelitian ini juga menegaskan bahwa fine-tuning model dengan dataset berbahasa Indonesia dari Common Voice 13 dapat meningkatkan akurasi transkripsi suara. Proses fine-tuning yang dilakukan menunjukkan bahwa model Whisper memiliki ketahanan lebih baik terhadap berbagai kondisi lingkungan dibandingkan dengan model lain yang diuji.Selain itu, hasil penelitian ini memberikan kontribusi dalam pengembangan teknologi speech recognition di Indonesia dengan menunjukkan bahwa model Whisper dapat menjadi solusi unggulan untuk kebutuhan transkripsi otomatis. Namun, masih terdapat ruang untuk pengembangan lebih lanjut, seperti eksplorasi model dengan dataset yang lebih besar, peningkatan akurasi dalam variasi aksen daerah, serta integrasi dengan sistem aplikasi berbasis suara di berbagai sektor. Penelitian selanjutnya disarankan untuk mengeksplorasi penggunaan Whisper dalam konteks multibahasa serta membandingkan kinerjanya dengan model speech recognition lain di skala global.

DAFTAR PUSTAKA

Ahmad Dhani Wafiy, Barlian Henryranu Prasetio, 2017, Penerapan Model Whisperpada Embedded Systemuntuk Speech To Text, Jurnal Pengemb Angan Teknologi Informasi Dan Ilmu Komputer.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common Voice: A Massively-Multilingual Speech Corpus. Https://Arxiv.Org/Abs/1912.06670. Https://Doi.Org/10.48550/Arxiv.1912.06670

Arisaputra, P., & Zahra, A. (2022). Indonesian Automatic Speech Recognition With Xlsr-53. Ingenierie Des Systemes D'information, 27(6), 973–982. Https://Doi.Org/10.18280/Isi.270614

Ferdiansyah, D., & Aditya, C. S. K. (2024). "Implementasi Automatic Speech Recognition Bacaan Al-Qur'an Menggunakan Metode Wav2vec 2.0 Dan Openai-Whisper." Triac: Jurnal Teknologi Informasi Dan Komunikasi, 14(2), 45-56. Doi: Https://Doi.Org/10.21107/Triac.V11i1.24332

Gong, Y., Khurana, S., Karlinsky, L., & Glass, J. (2023). "Whisper-At: Noise-Robust Automatic Speech Recognizers Are Also Strong General Audio Event Taggers." Arxiv Preprint Arxiv:2307.03183. Doi:10.48550/Arxiv.2307.03183

Hasan, F., Li, Y., Foulds, J., Pan, S., & Bhattacharjee, B. (2023). "Teach Me With A Whisper: Enhancing Large Language Models For Analyzing Spoken Transcripts Using Speech Embeddings." Arxiv Preprint Arxiv:2311.07014.

Peng, Y., Tian, J., Yan, B., Berrebbi, D., Chang, X., Li, X., ... & Watanabe, S. (2023). "Reproducing Whisper-Style Training Using An Open-Source Toolkit And Publicly Available Data." Arxiv Preprint Arxiv:2309.13876. Doi: 10.48550/Arxiv.2309.13876

Radford, A., Kim, J. W., Xu, T., Brockman, G., & Mcleavey, C. (2022). "Robust Speech Recognition Via Large-Scale Weak Supervision." Openai Research, 1-15. Https://Doi.Org/10.48550/Arxiv.2212.04356

Rekimoto, J. (2023). "Wesper: Zero-Shot And Realtime Whisper To Normal Voice Conversion For Whisper-Based Speech Interactions." Arxiv Preprint Arxiv:2303.01639. Doi:10.48550/Arxiv.2303.01639

Riefkyanov Surya Adia Pratama, Agit Amrullah, 2024, Analysis Of Whisper Automatic Speech Recognition Performance On Low Resource Language, Pilar Nusa Mandiri : Journal Of Computing And Information System Publishing, Doi: Https://Doi.Org/10.33480/Pilar.V20i1.4633.

Rifaldi, R., & Cahyono, A. B. (2024). "Pengembangan Fitur Speech Recognition Pada Aplikasi Notulensi Menggunakan Whisper Ai." Innovative: Journal Of Social Science Research, 4(6), 8555-8566. Doi: Https://Doi.Org/10.31004/Innovative.V4i6.17269

Roissyah Fernanda Khoiroh, Eric Julianto, Safrizal Ardana Ardiyansa, Haidar Ahmad Fajri, Implementasi Speech Recognition Whisper Pada Debat Calon Wakil Presiden Republik Indonesia, Explore. Doi:10.35200/Ex.V14i2.115.

Saputra, R. D. (2023). "Perancangan Sistem Pengenalan Ucapan Multibahasa (Arab & Indonesia) Berbasis Openai Whisper & Voiced-Unvoiced Classification." Jurnal Teknologi Terpadu, 12(1), 23-34

Yakubovskyi, R., & Morozov, Y. (2023). Speech Models Training Technologies Comparison Using Word Error Rate. Advances In Cyber-Physical Systems, 8(1), 2022. Doi:10.23939/Acps2023.01.074