INNOVATIVE: Journal Of Social Science Research
Volume 4 Nomor 6 Tahun 2024 Page 8555-8566
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Pengembangan Fitur *Speech Recognition* Pada Aplikasi Notulensi Menggunakan Whisper AI

Raihan Rifaldi[1], Andhik Budi Cahyono[2]✉
Universitas Islam Indonesia
Email: 105230101@uii.ac.id[2]✉

Abstrak

Rangkuman rapat perusahaan sangat membantu memastikan bahwa semua peserta terinformasi dengan baik dan tersampaikan. Namun proses penulisan rangkuman masih dilakukan secara manual, yang seringkali mengakibatkan kesalahan dan keterlambatan ketika melakukan penulisan. Untuk mengatasi masalah ini, dikembangkan SPERCO, sebuah aplikasi berbasis *speech recognition* menggunakan Whisper *Model* dari OpenAI. SPERCO dirancang untuk mencatat rangkuman secara otomatis dengan fitur *Speaker Diarization* yang dapat mengidentifikasi berbagai pembicara. Teknologi ini diimplementasi di Hugging face yang sangat membantu dalam menyediakan PTM (*Pre-Trained Models*). Hasil dari temuan ini dapat mengurangi *human error* dan membantu dalam pengembangan evaluasi perusahaan secara lebih mendalam.

Kata Kunci: *Speech Recognition, Whisper AI, Speaker Diarization, Notulensi*

Abstract

Company meeting summaries are very helpful in ensuring that all participants are well-informed and communicated with. However, the process of writing summaries is still done manually, which often results in errors and delays when writing. This problem can be addressed by SPERCO. SPERCO (Speech Recognition) is an application that can assist note-takers in automatically recording summaries using a voice feature developed using OpenAI's Whisper model technology to improve accuracy, time efficiency, and is enhanced with a Speaker Diarization feature that can identify multiple speakers simultaneously. This technology is implemented in Hugging Face, which is very helpful in providing PTM (Pre-Trained Models). The use of this application is expected to reduce human error and assist in the development of more in-depth company evaluations.

Keyword: *Speech Recognition, Whisper, Audio Transcript, Speaker Diarization, Minutes of Meeting*

PENDAHULUAN

Dalam beberapa dekade terakhir, teknologi pengenalan suara (*speech recognition*) telah berkembang pesat dan memberikan dampak signifikan pada berbagai bidang, termasuk dalam dunia korporasi. Salah satu penerapan yang mendapatkan perhatian khusus adalah dalam konteks pembuatan notulensi rapat. Proses pencatatan manual yang dilakukan selama rapat sering kali memakan waktu lama dan rentan terhadap kesalahan manusia. Hal ini dapat menyebabkan ketidaksesuaian antara hasil rapat dan laporan yang dihasilkan, yang pada akhirnya dapat berdampak negatif terhadap pengambilan keputusan(Tobin et al., 2024).

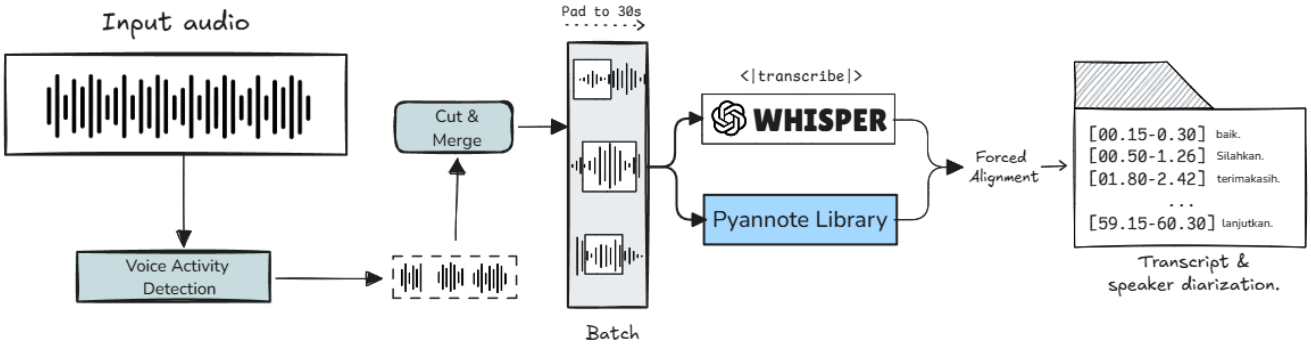
Beberapa perusahaan menghadapi tantangan dalam memastikan bahwa setiap rapat tercatat dengan akurat dan efisien. Mengingat volume rapat yang diadakan setiap hari, penting bagi perusahan untuk mengadopsi solusi yang mampu mengotomatiskan proses notulensi guna mengurangi beban kerja notulis serta meningkatkan akurasi dan kecepatan dalam menghasilkan laporan rapat. Untuk menjawab kebutuhan tersebut, dikembangkan aplikasi pencatatan rapat secara otomatis berbasis pengenalan suara ini SPERCO (*Speech Recognition*).

SPERCO dirancang untuk membantu notulis dalam merekam dan mentranskripsikan percakapan selama rapat secara otomatis, memungkinkan mereka untuk mengedit hasil transkripsi sesuai kebutuhan. Tujuan pengembangan SPERCO meliputi peningkatan efisiensi pencatatan notulensi, pengurangan kesalahan transkripsi, dan kemudahan dalam menyusun laporan rapat. Dengan semakin banyaknya penggunaan teknologi pengenalan suara, SPERCO diharapkan dapat meningkatkan produktivitas karyawan dan mengurangi

waktu yang diperlukan untuk membuat notulensi. Namun, tantangan seperti kualitas suara, aksen, dan kebisingan latar belakang dapat memengaruhi akurasi transkripsi. Oleh karena itu, SPERCO terus dikembangkan dengan memperhatikan faktor-faktor tersebut agar hasil transkripsi tetap relevan dan akurat sesuai kebutuhan rapat di Perusahaan.

METODE PENELITIAN

Speech Recognition (SPERCO) dikembangkan melalui metode *Voice Activity Detection* (VAD), Tujuan VAD adalah untuk mengekstraksi sinyal suara di mana terdapat aktivitas bicara dari audio mentah, sementara speech segmentation bertujuan untuk memotong audio menjadi segmen-segmen yang lebih kecil berdasarkan titik perubahan pembicara, memastikan bahwa setiap segmen secara akustik berasal dari individu yang



Gambar 1 Alur kerja menggunakan Whisper

sama(Lyu et al., 2024). Seperti pada Gambar 1 Alur kerja menggunakan Whisper. Proses pelaksanaan ini dibagi menjadi beberapa tahap utama sebagai berikut:

1. *Voice Activity Detection* (VAD)

Tahapan pertama dimulai dengan mendeteksi aktivitas audio yang diterima kemudian dilanjutkan pada tahapan VAD yang akan digunakan untuk mendeteksi bagian audio yang memiliki aktivitas suara, sehingga segmen yang tidak relevan (misalnya, keheningan atau noise) dapat dihapus. Teknik ini membantu meningkatkan efisiensi dalam proses transkripsi.

2. *Cut & Merge*

Tahapan selanjutnya yaitu segmen-segmen audio yang terdeteksi melalui VAD akan dipotong dan digabungkan agar memiliki durasi tertentu (kemungkinan untuk menyesuaikan dengan batas model transkripsi).

3. *Batch* & *Padding*

Setelah aktivitas suara terdeteksi, segmen audio diatur agar memiliki panjang yang sesuai untuk diproses dalam *model*. *Padding* digunakan untuk menjaga ukuran *Batch* tetap seragam agar tetap efisiensi dalam pemrosesan.

4. Whisper dan Pyannote *Library*

Selanjutnya akan melibatkan dua proses ini, Whisper dan pustaka Pyannote. Whisper, *model Automatic Speech Recognition* (ASR) dari OpenAI, digunakan untuk menghasilkan transkrip dari audio. Model ini memiliki keunggulan dalam mendeteksi multibahasa dan bekerja dengan baik pada berbagai jenis kualitas audio. Sedangkan pustaka Pyannote digunakan untuk fitur *Speaker Diarization*, yaitu menentukan pembicara dengan mencocokkan *timestamp* dari transkripsi Whisper dengan hasil segmentasi pembicara. Lalu dilanjutkan dengan proses penyelarasan dilakukan untuk memastikan bahwa *timestamp* dari transkripsi dan *Speaker Diarization* selaras secara akurat (Forced Alignment)(M. Y. Wang & Purushotam, n.d.).

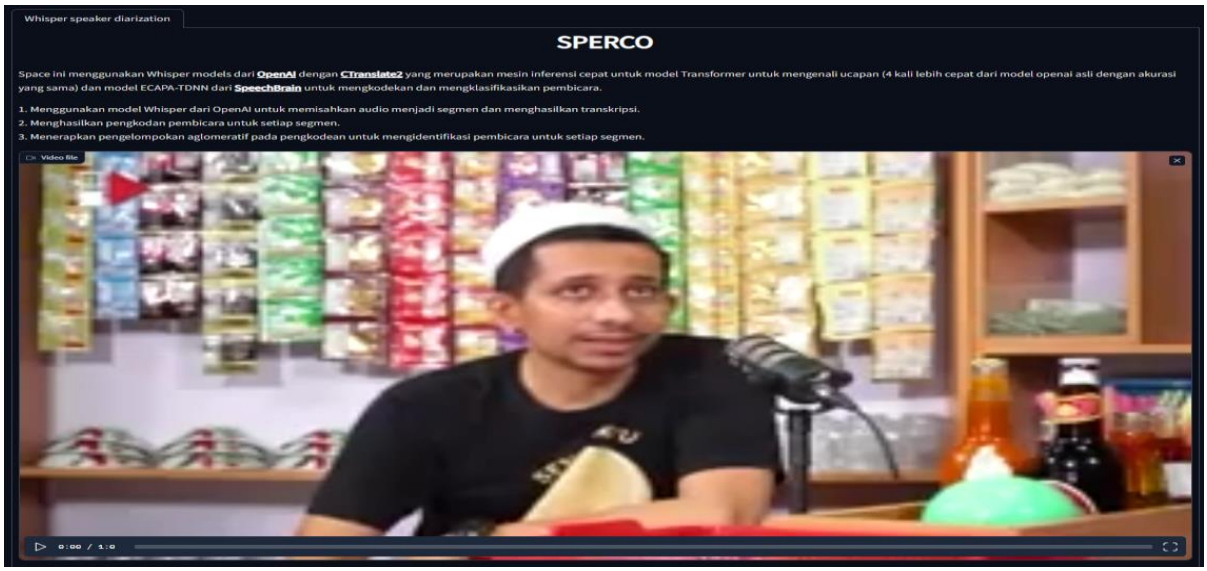
5. Transcript & Speaker Diarization

Hasil akhirnya berupa transkrip dengan *timestamp* yang presisi serta *Speaker Diarization* pembicara (contohnya, siapa berbicara pada waktu tertentu). Hal ini mempermudah analisis lebih lanjut, seperti analisis konten atau pengklasifikasian dialog.

HASIL DAN PEMBAHASAN

Pada tahapan implementasi, SPERCO dirancang untuk melakukan transkrip video audio rapat. Aplikasi akan menghasilkan output ketika video rapat diunggah pada aplikasi lalu menyesuaikan *title* yang dibutuhkan seperti, bahasa yang digunakan, model whisper, dan jumlah *speaker*

Format file yang diunggah dalam bentuk MP4. Format ini digunakan karena format ini didukung oleh hampir semua pemutar video, *browser web*, dan sistem operasi sehingga dapat menyeimbangkan kualitas video dengan ukuran file yang lebih kecil. Format ini juga dapat melakukan kompresi file audio dan video secara terpisah untuk menjaga kualitas(D.P. et al., 2021).



Gambar 2 Sampel video Youtube Podcast

Langkah pertama yaitu mengunggah video. Pada sampel yang digunakan pada pengujian ini yaitu Youtube *podcast* berjudul "PWK - Ternyata Cerita Masa Kecil Habib Ja'far Kocak Banget" dengan mengambil salah satu cuplikan berdurasi 1 menit seperti pada Gambar 2. *Podcast* ini digunakan karena video tersebut berfokus pada dialog pembicaraan dan kualitas audio yang digunakan cukup baik sehingga dapat digunakan sebagai sampel dalam penggunaan pada aplikasi SPERCO.

Langkah selanjutnya adalah memilih bahasa yang digunakan. Pada proyek ini tersedia dalam 4 bahasa yaitu Indonesia, English, Jawa, dan Sunda. Ketika telah memilih bahasa yang diinginkan, selanjut memilih model Whisper yang akan digunakan seperti yang tertera pada Gambar 3 Whisper Models.

Gambar 3 Whisper *Models*

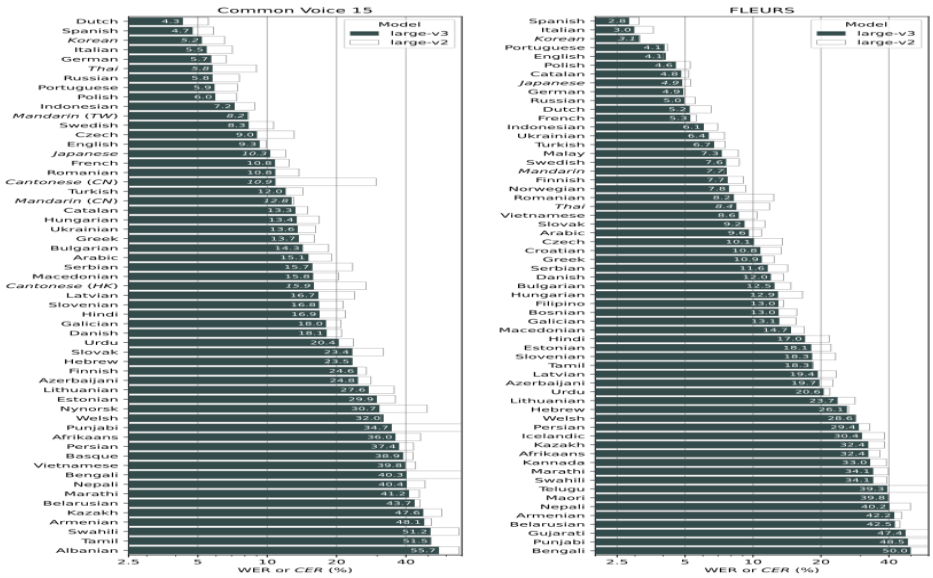
Pemilihan ini bertujuan untuk menentukan keseimbangan antara akurasi, kecepatan, dan kebutuhan perangkat. Untuk aplikasi yang membutuhkan presisi tinggi, model besar seperti *large* atau *medium* ideal. Namun, untuk efisiensi dan kecepatan, model yang lebih kecil seperti *tiny* atau *base* lebih relevan.

Performa Whisper sangat bervariasi bergantung pada bahasanya. Gambar 4 di bawah menunjukkan perincian performa model *large-v3* dan *large-v2* berdasarkan

bahasa, menggunakan WER (*Word Error Rate*) atau CER (*Character Error Rates*) yang dievaluasi pada kumpulan data Common Voice 15 dan Fleurs. Metrik WER/CER tambahan yang sesuai dengan model(Radford et al., 2022).

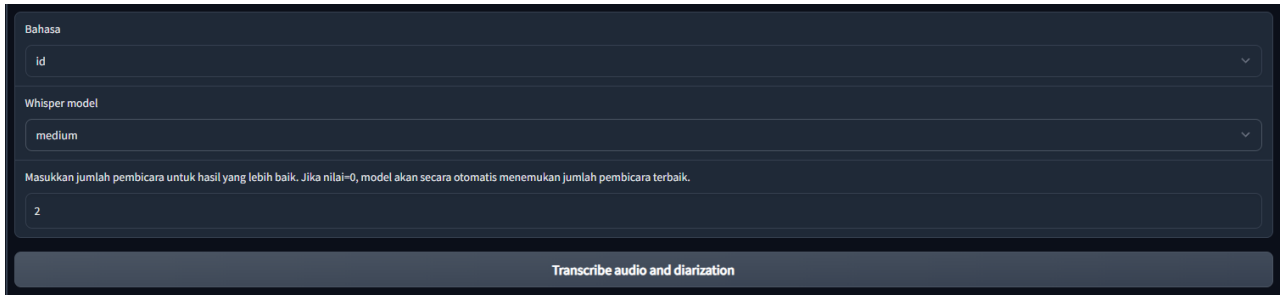
Tabel 1 Whisper Models details

Multilingual model	Parameters	Required VRAM	Relative Speed
Tiny	39 M	~1 GB	~32x
Base	74 M	~1 GB	~16x
Small	244 M	~2 GB	~6x
medium	769 M	~5 GB	~2x
Large	1550 M	~10 GB	~1x

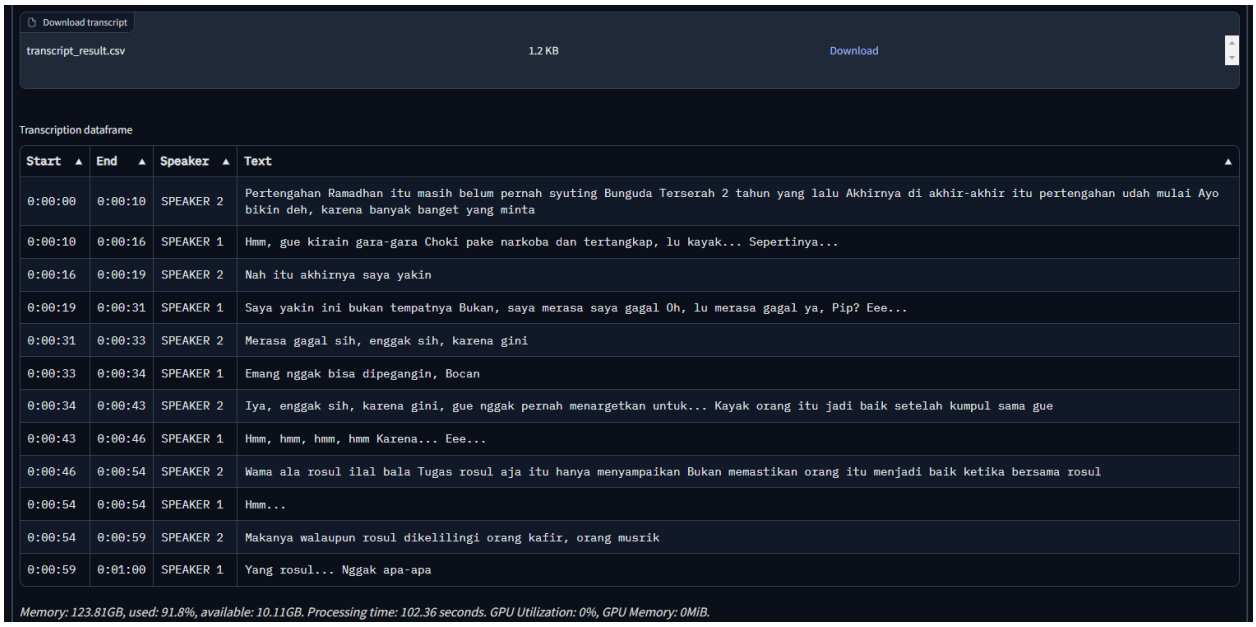


Gambar 4 Pengujian model Whisper pada Common Voice 15 & FLEURS

Penyesuaian ini harus dilakukan berdasarkan spesifikasi dari file yang diunggah. Setelah pemilih model Whisper, dilanjutkan untuk melakukan pemilihan jumlah pembicara pada video audio transkrip. Pemilihan ini dilakukan untuk memberikan hasil yang lebih optimal ketika melakukan transkrip seperti pada Gambar 5. Jika pengguna tidak mengetahui berapa jumlah pembicara pada video yang diunggah, dapat diisi dengan 0 secara otomatis SPERCO akan menemukan jumlah pembicara terbaik. Ketika semua *title* telah diisi, langkah selanjutnya menekan tombol *Transcribe audio and diarization*. Hasil transkrip dapat dilihat pada halaman SPERCO namun jika membutuhkan hasil dalam bentuk format lain, tersedia dalam bentuk *CSV* yang dapat diunduh. Pada halaman ini kita dapat melihat *Transcription dataframe*, mulai dari dialog dimulai, diakhiri, pembicara dan dialog apa yang diucapkan terekam pada halaman SPERCO seperti pada Gambar 6 Data dialog pembicara.



Gambar 5 Pemilihan jumlah pembicara



Gambar 6 Data dialog pembicara

Hasil Pengujian Akurasi Transkripsi

Pengujian akurasi menggunakan metode *Word Error Rate* (WER). WER adalah metrik yang digunakan untuk mengevaluasi kualitas transkripsi yang dihasilkan oleh Sistem *Automatic Speech Recognition* (ASR). Dalam berbagai aplikasi, sering kali penting untuk memperkirakan WER berdasarkan sepasang ujaran suara dan transkripsinya.

Pengujian adaptasi berbasis bahasa menggunakan berbagai dialek China mengungkapkan bahwa penerapan *speech-based in-context learning (SICL)* pada sistem ASR untuk pengenalan kata secara terpisah mampu memberikan penurunan (WER). Temuan ini diverifikasi menggunakan tugas adaptasi pembicara atau pengenalan ucapan berkelanjutan, dan keduanya mencapai pengurangan relatif WER yang signifikan(S. Wang et al., 2023).

Penelitian sebelumnya tentang estimasi WER berfokus pada membangun model yang dilatih dengan mempertimbangkan sistem ASR tertentu salah satunya Whisper. Rumus untuk WER sebagai berikut:

$$WER = \frac{S + D + I}{N}$$

Keterangan:

- Substitusi (S): Jumlah kata yang salah ditranskripsi.
- Penyisipan (I): Jumlah kata tambahan yang tidak ada dalam referensi.
- Penghapusan (D): Jumlah kata yang hilang dari referensi.

WER = 0 menunjukkan transkripsi sempurna, sedangkan nilai lebih tinggi menunjukkan tingkat kesalahan yang lebih besar. Sedangkan untuk mengetahui nilai presentase akurasi transkripsi dapat dihitung dengan rumus berikut:

$$Presentase\ akurasi = (1 - WER) \times 100\ \%$$

Dalam proses pengujian, dilakukan dengan Whisper model tipe *Medium* dan sampel video *podcast* "PWK - Ternyata Cerita Masa Kecil Habib Ja'far Kocak Banget" dengan mengambil salah satu cuplikan berdurasi 1 menit. Dalam melakukan tiga tingkatan kualitas audio, yaitu Kurang Baik, Baik, dan Sangat Baik. Berikut penjelasnnya:

- Kualitas Kurang Baik: Audio banyak mengalami *noise*, seperti suara kendaraan, penonton ataupun suara lainnya.
- Kualitas Baik: Audio terdengar dengan cukup baik, mudah dipahami namun sedikit mengalami kebingungan dalam pengucapan kata atau masih ada gangguan dari suara *noise* sekitar.
- Kualitas Sangat Baik: Audio sangat jelas dan mudah dipahami. Tanpa gangguan suara sekitar.
- Kualitas Ragu: Suara terdengar jelas, namun tidak dapat mendeteksi pembicara dikarenakan dialog scara bersamaan, *noise* ataupun hal lainnya.

Tabel 2 Pengujian akurasi menggunakan Metode *Ground Truth*

No	Ground Truth	Transkrip SPERCO	WER	Akurasi(%)	Kualitas Audio
1	[0:00:00-0:00:10] SPEAKER 2: Pertengahan Ramadhan itu masih belum pernah syuting pemuda tersesat 2	[0:00:00-0:00:10] SPEAKER 2: Pertengahan Ramadhan itu masih belum pernah syuting Bunguda Terserah 2 tahun yang lalu Akhirnya di akhir-akhir itu pertengahan udah	0,07	93%	Tinggi

	tahun yang lalu Akhirnya di akhir-akhir itu pertengahan udah mulai Ayo bikin deh, karena banyak banget yang minta	mulai Ayo bikin deh, karena banyak banget yang minta			
2	[:00:10-0:00:16] SPEAKER 1: Hmm, gue kirain gara-gara Choki pake narkoba dan tertangkap, lu kek... Sepertinya...	[:00:10-0:00:16] SPEAKER 1: Hmm, gue kirain gara-gara Choki pake narkoba dan tertangkap, lu kayak... Sepertinya...	0.14	86%	Sedang
3	[0:00:16-0:00:19] SPEAKER 2: Hehe.. nah itu akhirnya saya yakin.	[0:00:16-0:00:19] SPEAKER 2: Nah itu akhirnya saya yakin.	0,16	84%	Sedang
4	[0:00:19-0:00:31] SPEAKER 2 : Saya yakin ini bukan tempatnya. Bukan, saya merasa ah saya gagal. SPEAKER 1: Oh, lu, lu merasa gagal ya, Bib? Eee...	[0:00:19-0:00:31] SPEAKER 1: Saya yakin ini bukan tempatnya. Bukan, saya merasa saya gagal. Oh, lu merasa gagal ya, Pip? Eee...	0,26	74%	Ragu dan Kurang baik
5	[0:00:31-0:00:33] SPEAKER 2: Eee merasa gagal sih, enggak sih, karena gini.	[0:00:31-0:00:33] SPEAKER 2: Merasa gagal sih, enggak sih, karena gini.	0,12	87,5%	Baik
6	[0:00:33-0:00:34] SPEAKER 1: Emang nggak bisa dipegangin, Bocah	[0:00:33-0:00:34] SPEAKER 1: Emang nggak bisa dipegangin, Bocan.	0,8	80%	Baik
7	[0:00:34-0:00:43] SPEAKER 2:	[0:00:34-0:00:43] SPEAKER 2: Iya, enggak sih, karena gini,	0	100%	Sangat Baik

	Iya, enggak sih, karena gini, gue nggak pernah menargetkan untuk... Kayak orang itu jadi baik setelah kumpul sama gue	gue nggak pernah menargetkan untuk... Kayak orang itu jadi baik setelah kumpul sama gue.			
8	[0:00:43-0:00:46] SPEAKER 2: Hmm, hmm, hmm, hmm. Karena... Eee...	[0:00:43-0:00:46] SPEAKER 1: Hmm, hmm, hmm, hmm. Karena... Eee...	0	100%	Ragu
9	[0:00:46-0:00:54] SPEAKER 2: Wama ala rosul ilal bala. Tugas rasul aja itu hanya menyampaikan. Bukan memastikan orang itu menjadi baik ketika bersama rosul.	[0:00:46-0:00:54] SPEAKER 2: Wama ala rosul ilal bala. Tugas rosul aja itu hanya menyampaikan. Bukan memastikan orang itu menjadi baik ketika bersama rosul.	0,05	95%	Sangat Baik
10	[0:00:54-0:00:54] SPEAKER 1: Hmm...	[0:00:54-0:00:54] SPEAKER 1: Hmm...	0	100%	Sangat Baik
11	[0:00:54-0:00:59] SPEAKER 2: Makanya walaupun rasul dikelilingi orang kafir, orang musyrik	[0:00:54-0:00:59] SPEAKER 2: Makanya walaupun rosul dikelilingi orang kafir, orang musrik.	0,2	80%	Baik
12	[0:00:59-0:01:00] SPEAKER 2: Ya rasul... Nggak apa-apa.	[0:00:59-0:01:00] SPEAKER 1: Yang rosul... Nggak apa-apa.	0,2	80%	Baik

Berdasarkan hasil pengujain pada Tabel 2, SPERCO menampilakan performa diberbagai aspek. Mulai dari skala kualitas audio Sangat Baik dengan tingkat kesalahan kata (WER) yang rendah dan akurasi 100%. Pada kualitas audio Baik, performa masih dapat dirasakan dengan akurasi 80%, Namun ada pada kondisi Ragu dan Kurang baik, akurasi ini dipengaruhi oleh beberapa faktor, seperti Pembicara yang tidak terdeteksi,

meskipun kata-kata yang diucapkan dapat didengar, namun tidak dengan jelas. Kualitas dan lingkungan sekitar sangat mempengaruhi hasil transkrip yang diberikan untuk menginginkan hasil kualitas audio yang lebih akurat.

SIMPULAN

Pengembangan SPERCO pada aplikasi notulensi dengan integrasi Whisper *Model* di Hugging Face menunjukkan akurasi transkripsi otomatis yang baik. Aplikasi ini mengotomatisasi video rapat menjadi teks menggunakan *Speech recognition* pada fitur *Speaker Diarization* untuk mengenali suara individu. Teknologi ini mempermudah pembuatan catatan rapat tanpa campur tangan manual, mendokumentasikan seluruh pembicaraan secara menyeluruh. Aplikasi ini diharapkan dapat membantu di berbagai bidang, meningkatkan produktivitas, mendukung kepatuhan pencatatan rapat, dan memberikan kemudahan akses notula lengkap.

DAFTAR PUSTAKA

D.P., G., Pathania, A., -, A., & -, S. (2021). Authentication Of Digital Mp4 Video Recordings Using File Containers And Metadata Properties. International Journal Of Computer Science Engineering, 10(2), 28–38. Https://Doi.Org/10.21817/Ijcsenet/2021/V10i2/211002004

Lyu, K. M., Lyu, R. Yuan, & Chang, H. T. (2024). Real-Time Multilingual Speech Recognition And Speaker Diarization System Based On Whisper Segmentation. Peerj Computer Science, 10. Https://Doi.Org/10.7717/Peerj-Cs.1973

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2022). Robust Speech Recognition Via Large-Scale Weak Supervision. Http://Arxiv.Org/Abs/2212.04356

Tobin, J., Nelson, P., Macdonald, B., Heywood, R., Cave, R., Seaver, K., Desjardins, A., Jiang, P.-P., & Green, J. R. (2024). Automatic Speech Recognition Of Conversational Speech In Individuals With Disordered Speech. Journal Of Speech, Language, And Hearing Research, 1–10. Https://Doi.Org/10.1044/2024_Jslhr-24-00045

Wang, M. Y., & Purushotam, G. N. (N.D.). Using Deep Learning And Augmented Reality To Improve Accessibility: Inclusive Conversations Using Diarization, Captions, And Visualization.

Wang, S., Yang, C.-H. H., Wu, J., & Zhang, C. (2023). Can Whisper Perform Speech-Based In-Context Learning? Http://Arxiv.Org/Abs/2309.07081

Jones, J., Jiang, W., Synovic, N., Thiruvathukal, G., & Davis, J. (2024). What Do We Know About Hugging Face? A Systematic Literature Review And Quantitative Validation Of Qualitative Claims. Proceedings Of The 18th Acm/Ieee International Symposium On Empirical Software Engineering And Measurement, 13–24. Https://Doi.Org/10.1145/3674805.3686665

Jorg, T., Kämpgen, B., Feiler, D., Müller, L., Düber, C., Mildenberger, P., & Jungmann, F. (2023). Efficient Structured Reporting In Radiology Using An Intelligent Dialogue System Based On Speech Recognition And Natural Language Processing. Insights Into Imaging, 14(1). Https://Doi.Org/10.1186/S13244-023-01392-Y

Ma, H., Peng, Z., Shao, M., Li, J., & Liu, J. (2023). Extending Whisper With Prompt Tuning To Target-Speaker Asr. Http://Arxiv.Org/Abs/2312.08079

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic Speech Recognition: A Survey. Multimedia Tools And Applications, 80(6), 9411–9457. Https://Doi.Org/10.1007/S11042-020-10073-7