BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi kecerdasan buatan atau *Artificial Intelligence* (AI) telah mengalami kemajuan pesat, terutama dengan munculnya model-model generatif seperti *Generative Pre-trained Transformer* (GPT) yang mampu memahami konteks dan menghasilkan respons yang sangat mirip manusia dalam percakapan [1], [2]. Kemajuan ini mendorong transformasi interaksi manusia dengan mesin, di mana kebutuhan akan sistem AI yang dapat berkomunikasi secara alami dan efektif semakin meningkat di berbagai bidang seperti layanan pelanggan, pendidikan, dan Kesehatan [3]. Namun, tantangan muncul karena interaksi manusia yang kompleks dan beragam, sehingga organisasi dan individu memerlukan solusi *conversational* AI yang mampu menangani percakapan multi-tahap dengan konteks yang berkelanjutan dan respons yang relevan [1], [4]. Selain itu, kebutuhan akan interaksi suara yang lebih alami dan responsif mendorong pengembangan *conversational voice* AI yang menggabungkan pengenalan suara, pemahaman bahasa, dan sintesis suara secara terpadu [5]. Dengan demikian, *conversational voice* AI menjadi solusi penting untuk mengatasi keterbatasan komunikasi tradisional antara manusia dan mesin, meningkatkan efisiensi dan kualitas interaksi dalam berbagai aplikasi praktis [6]. Fenomena ini menandai era baru di mana AI tidak hanya sebagai alat, tetapi juga sebagai mitra interaktif dalam kehidupan sehari-hari dan dunia kerja [2], [7].

Teknologi *voice conversational* AI dirancang untuk memungkinkan interaksi alami antara manusia dan mesin melalui percakapan suara. Sistem ini umumnya dibangun dengan arsitektur *cascaded* yang terdiri dari tiga komponen utama, yaitu *Automatic Speech Recognition* (ASR) atau *Speech-to-Text* (STT), *Large Language Model* (LLM) atau *Small Language Model* (SLM), dan *Text-to-Speech* (TTS) [8]. Proses dimulai ketika *input* suara dari pengguna diubah menjadi teks oleh modul ASR atau STT. Selanjutnya, teks tersebut diproses oleh LLM atau SLM untuk

memahami konteks, menghasilkan respons, atau melakukan *reasoning* berbasis bahasa alami. Hasil keluaran berupa teks respons kemudian diubah kembali menjadi suara melalui modul TTS, sehingga pengguna dapat menerima balasan dalam bentuk *audio*. Arsitektur *cascaded* ini banyak digunakan karena fleksibilitas dan kemudahan integrasinya, serta memungkinkan pengembangan dan optimasi setiap komponen secara terpisah [9]. Namun, pendekatan ini juga memiliki tantangan seperti potensi hilangnya informasi paralinguistik (intonasi dan emosi) saat konversi antar-*modality*, serta akumulasi *error* dari satu tahap ke tahap berikutnya [8]. Meskipun demikian, arsitektur *cascaded* tetap menjadi fondasi utama dalam pengembangan *voice conversational voice* AI modern, beberapa contohnya untuk aplikasi layanan pelanggan, asisten virtual, maupun sistem edukasi berbasis suara.

Dalam pengembangan sistem percakapan berbasis suara, setiap komponen utama system memiliki berbagai opsi model yang umum digunakan di penelitian-penelitian sebelumnya. Untuk ASR atau STT, model-model populer meliputi GMM-HMM, Deep Neural Network-HMM (DNN-HMM), wav2vec, QuartzNet, FastConformer, dan Whisper [10], [11], [12], [13], [14]. Pada komponen *SLM*, *Large Language Models* (LLM) seperti GPT-3, Llama, dan Gemma sering digunakan karena kemampuannya dalam memahami konteks dan menghasilkan respons yang relevan [15], [16]. Sementara itu, untuk TTS, model yang banyak dipakai antara lain Tacotron, FastSpeech, HiFi-GAN, dan VITS, yang menawarkan kualitas suara alami dan efisiensi tinggi [17], [18], [19], [20].

Dalam penelitian ini, dipilih model Whisper untuk ASR, Gemma 3 1B untuk SLM, dan VITS untuk TTS. Whisper dipilih karena terbukti unggul dalam berbagai pengujian lintas bahasa dan *domain*, dengan performa yang konsisten lebih baik dibandingkan model lain seperti XLSR-53, QuartzNet, dan FastConformer, terutama dalam hal *Word Error Rate* (WER) yang lebih rendah [11], [13], [14]. Gemma 3 1B dipilih sebagai SLM karena model ini, meskipun berukuran lebih kecil dari GPT-3, mampu melampaui model BERT dan LLM besar lain dalam tugas pemahaman bahasa, serta efisien untuk *deployment* di lingkungan dengan sumber daya terbatas [15]. Untuk TTS, VITS dipilih karena arsitektur *end-to-end*-nya

menghasilkan suara yang alami, mendukung multi-*speaker*, dan dapat dioptimasi untuk *parameter* yang lebih efisien, serta telah terbukti unggul dalam berbagai metrik kualitas suara dibandingkan model TTS konvensional [18], [19], [20].

Setiap model yang dipilih memiliki keunggulan dalam kualitas dan akurasi, namun juga menghadapi tantangan signifikan ketika diimplementasikan pada perangkat dengan sumber daya terbatas. Whisper, meskipun unggul dalam akurasi dan generalisasi lintas *domain*, memiliki ukuran model dan kebutuhan komputasi yang cukup besar sehingga kurang efisien untuk *deployment* di perangkat berbasis CPU atau *edge device*. Tantangan ini dapat diatasi melalui optimasi seperti kompresi model, *quantization*, dan format model yang lebih ringan, yang terbukti mampu menurunkan kebutuhan memori dan mempercepat inferensi tanpa penurunan akurasi yang signifikan [21], [22]. Pada SLM, meskipun model seperti Gemma 3 1B dirancang untuk efisiensi, keterbatasan komputasi pada lingkungan CPU menjadikan *quantization* sebagai strategi utama untuk meningkatkan kelayakan inferensi model pada kebutuhan *deployment* dengan sumber daya terbatas. Untuk TTS, VITS dikenal menghasilkan suara alami dan efisien, tetapi model ini tetap cukup berat untuk perangkat *low-cost*, sehingga optimasi seperti konversi ke ONNX *runtime* dan *quantization* sangat penting untuk mempercepat inferensi dan menurunkan konsumsi memori [23], [24].

Meskipun berbagai penelitian telah membuktikan efektivitas optimasi pada komponen individual sistem *voice conversational* AI, terdapat *research gap* signifikan dalam implementasi optimasi *end-to-end* pada lingkungan CPU-*only*. Mayoritas studi terdahulu berfokus pada optimasi *single-component* atau *deployment* berbasis *hybrid cloud-edge* dengan akselerasi GPU, sementara integrasi sistematis optimasi *inference-level* pada seluruh komponen *pipeline* dengan teknik *quantization* yang berbeda masih terbatas. Penelitian-penelitian sebelumnya telah menunjukkan hasil optimasi yang baik pada level komponen individual, namun dokumentasi mengenai integrasi antar-komponen dan *trade-off* performa dalam satu sistem terpadu masih belum komprehensif. Kekosongan riset ini menjadi krusial mengingat kebutuhan industri akan *deployment voice* AI pada infrastruktur CPU-*only* dikarenakan sebagian besar infrastruktur perusahaan yang tidak

didukung GPU atau TPU, serta untuk memenuhi regulasi privasi data seperti HIPAA, GDPR, dan PCI-DSS yang mensyaratkan pemrosesan lokal tanpa pengiriman data ke *cloud* [25], [26]. Oleh karena itu, diperlukan penelitian yang mengeksplorasi optimasi multi-komponen dalam arsitektur *cascaded* untuk membuktikan kelayakan *deployment* pada lingkup CPU-*only*.

Penelitian ini berfokus pada metode optimasi *inference* yang diterapkan pada setiap komponen utama sistem, yaitu Whisper untuk ASR, Gemma 3 1B sebagai SLM, dan VITS untuk TTS. Untuk Whisper, teknik *quantization* terbukti secara signifikan mengurangi ukuran model dan mempercepat proses inferensi tanpa penurunan akurasi yang berarti, sehingga sangat cocok untuk *deployment* pada perangkat dengan sumber daya terbatas [27], [28], [29]. Implementasi format model yang lebih efisien seperti GGUF, serta penggunaan *inference framework* seperti ctranslate2, memungkinkan eksekusi model secara optimal di CPU dengan konsumsi memori dan *latency* yang jauh lebih rendah [30]. Pada komponen SLM, penerapan optimasi melalui konversi model Gemma 3 1B ke format GGUF dengan Q8_0 *quantization* memungkinkan pengurangan presisi numerik bobot model sehingga kompleksitas komputasi dan kebutuhan memori pada tahap inferensi dapat ditekan, tanpa memerlukan proses pelatihan ulang serta dengan tetap menjaga kualitas keluaran model pada tingkat yang dapat diterima untuk kebutuhan aplikasi. Untuk TTS, ekspor model VITS ke ONNX *Runtime* dan penerapan *quantization* telah terbukti meningkatkan efisiensi komputasi dan menurunkan *latency* inferensi secara signifikan, sehingga memungkinkan pemrosesan suara secara *real-time* di perangkat *low-cost* tanpa degradasi kualitas suara yang berarti [31]. Dengan menerapkan metode optimasi ini, sistem dapat berjalan lebih efisien dan responsif pada perangkat atau *server* dengan daya komputasi rendah, sekaligus mempertahankan kualitas hasil yang kompetitif.

Melalui rangkaian optimasi yang dilakukan, diharapkan penelitian dapat memberikan kontribusi, mencakup implementasi dan validasi strategi optimasi multi-komponen yang mengintegrasikan CTranslate2 INT8 *quantization* untuk ASR, GGUF *quantization* untuk SLM, dan ONNX *Runtime dynamic quantization* untuk TTS dalam satu *pipeline* terpadu. Penelitian ini membuktikan bahwa optimasi

inference-level mampu mencapai pengurangan *end-to-end latency* yang signifikan dengan tetap mempertahankan kualitas *output*, sekaligus memvalidasi kapabilitas *deployment* pada sistem berbasis CPU atau *local hosting* tanpa akselerasi *hardware* khusus. Secara teoritis, penelitian ini memperkaya literatur mengenai optimasi arsitektur *cascaded* dan implementasi model ringan untuk *voice conversational* AI, khususnya dalam lingkup dengan keterbatasan sumber daya komputasi. Secara praktis, hasil penelitian diharapkan dapat dimanfaatkan oleh industri telekomunikasi, layanan kesehatan, dan layanan finansial untuk membangun solusi *voice assistant* yang memenuhi *compliance* regulasi privasi data, hemat infrastruktur, dan dapat diakses secara luas tanpa ketergantungan pada *cloud services* atau *hardware* mahal.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian yang telah diuraikan, berikut merupakan beberapa rumusan masalah dalam penelitian ini:

1. Bagaimana merancang sistem *voice conversational* AI berbasis arsitektur *cascaded*?
2. Bagaimana metode optimasi *inference* dengan metode *quantization* dan *runtime conversion* diterapkan pada komponen ASR, SLM, dan TTS?
3. Bagaimana hasil evaluasi akurasi, kualitas, dan performa sistem setelah dilakukan optimasi pada setiap komponen?

1.3 Batasan Masalah

Dalam menjaga fokus penelitian terkait optimasi *inference* pada sistem *voice conversational* AI berbasis arsitektur *cascaded*, batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Sistem *voice conversational* AI yang dibangun hanya mencakup tiga komponen utama, yaitu *Automatic Speech Recognition* (ASR), *Small Language Model* (SLM), dan *Text-to-Speech* (TTS), tanpa mencakup fitur tambahan lainnya.

2. Model yang digunakan terbatas pada Whisper untuk ASR, Gemma 3 1B untuk SLM, dan VITS untuk TTS, sehingga penelitian tidak membahas model lain di luar ruang lingkup tersebut.
3. Optimasi yang dilakukan berfokus pada level *inference*, meliputi *quantization*, dan percepatan *runtime*, tanpa melakukan pelatihan ulang (*re-training*) atau perubahan arsitektur internal dari model-model tersebut.
4. Evaluasi performa difokuskan pada metrik-metrik tertentu, yaitu *latency*, efisiensi penggunaan sumber daya komputasi, akurasi transkripsi, kualitas respons bahasa, dan kualitas sintesis suara.
5. Pengujian dilakukan dalam lingkungan komputasi berdaya rendah berbasis CPU, sehingga hasil penelitian tidak mencakup performa pada perangkat dengan akselerator GPU atau *high-performance computing*.
6. Cakupan bahasa yang diuji terbatas pada Bahasa Indonesia, sehingga penelitian tidak membahas performa model dalam bahasa lain atau variasi dialek daerah secara khusus.
7. Keluaran penelitian berupa model dan *pipeline* sistem yang telah dioptimasi, yang dapat diintegrasikan ke berbagai aplikasi *voice conversational* AI, namun tidak mencakup pengembangan aplikasi industri secara penuh.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Tujuan dari dilakukannya penelitian ini adalah sebagai berikut:

1. Membangun sistem *voice conversational* AI berbasis arsitektur *cascaded* yang terdiri dari komponen ASR, SLM, dan TTS.
2. Menerapkan dan mengoptimalkan proses *inference* pada setiap komponen sistem agar lebih efisien dan sesuai untuk digunakan pada perangkat dengan kapasitas komputasi terbatas.
3. Mengevaluasi performa sistem setelah dilakukan optimasi melalui pengukuran akurasi, kualitas *output*, serta efisiensi penggunaan sumber daya komputasi.

1.4.2 Manfaat Penelitian

Manfaat dari adanya penelitian ini adalah sebagai berikut:

1. Memberikan pendekatan yang lebih efisien dalam mengembangkan sistem *voice conversational* AI yang mampu berjalan dengan baik pada perangkat atau *server* berdaya rendah, sehingga dapat mendukung implementasi teknologi percakapan suara tanpa memerlukan infrastruktur komputasi yang mahal.
2. Mendorong peningkatan kualitas layanan berbasis suara pada organisasi, industri, maupun sektor publik, karena sistem yang lebih cepat dan responsif dapat digunakan untuk mendukung layanan informasi, otomasi operasional, dan peningkatan pengalaman pengguna.
3. Menyediakan dasar penerapan *voice assistant* yang mudah diakses, baik untuk kebutuhan internal maupun eksternal organisasi, sehingga teknologi ini dapat dimanfaatkan untuk menjawab kebutuhan masyarakat, meningkatkan efisiensi kerja, serta memperluas akses terhadap layanan berbasis suara.
4. Menjadi landasan bagi penelitian selanjutnya dalam pengembangan dan optimasi *sistem conversational* AI, khususnya pada pendekatan optimasi *inference* untuk komponen ASR, SLM, dan TTS, sehingga dapat memperkaya literatur dan mendorong penelitian lanjutan terkait efisiensi model dan desain sistem percakapan suara.

1.5 Sistematika Penulisan

BAB I PENDAHULUAN

Bab 1 merupakan bagian pendahuluan yang memuat beberapa elemen utama, yaitu latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, serta manfaat penelitian. Bagian ini tidak hanya memberikan gambaran umum mengenai konteks topik yang dibahas, tetapi juga menguraikan permasalahan inti yang menjadi dasar pelaksanaan penelitian. Pemaparan mengenai

permasalahan tersebut menjadi dasar penting dalam menentukan arah penelitian dan ruang lingkup yang dianalisis.

BAB II LANDASAN TEORI

Bab ini menyajikan pembahasan yang lebih komprehensif mengenai literatur dan teori-teori yang berkaitan dengan penelitian, termasuk algoritma, perangkat, serta *framework* yang digunakan. Seluruh teori yang dijelaskan merujuk pada sumber-sumber terpercaya seperti buku akademik dan jurnal ilmiah yang telah terverifikasi kredibilitasnya. Bab ini bertujuan memberikan landasan konseptual yang kuat bagi metode dan pendekatan penelitian, serta menunjukkan relevansi literatur yang ada dalam menjawab permasalahan yang diangkat.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan pendekatan penelitian yang digunakan serta langkah-langkah yang ditempuh dalam pelaksanaan penelitian. Di dalamnya dijabarkan proses yang menjadi dasar dalam perancangan dan pengembangan sistem, mulai dari tahapan yang dilakukan, metode yang digunakan, hingga alur kerja penelitian secara keseluruhan. Penjelasan mengenai cara pengumpulan informasi yang dibutuhkan juga disampaikan untuk menunjukkan bagaimana data atau hasil pengujian diperoleh dan digunakan dalam mendukung tujuan penelitian. Selain itu, bab ini memberikan gambaran umum mengenai alur penelitian yang dijalankan dari awal hingga akhir penelitian.

BAB IV ANALISIS DAN HASIL PENELITIAN

Bab ini menyajikan hasil penerapan metodologi penelitian yang telah dirancang, serta pembahasan terhadap temuan-temuan yang diperoleh. Penjelasan dimulai dari pemaparan hasil proses yang dilakukan selama penelitian, kemudian dilanjutkan dengan pengolahan dan interpretasi informasi untuk memastikan bahwa data yang diperoleh relevan dan dapat dipertanggungjawabkan. Selanjutnya, bab ini menguraikan analisis yang dilakukan untuk memahami hasil tersebut secara menyeluruh, menjelaskan langkah-langkah yang digunakan dalam mengevaluasi

temuan penelitian. Pembahasan juga mencakup penjelasan mengenai bagaimana hasil yang diperoleh menjawab rumusan masalah dan tujuan penelitian.

BAB V SIMPULAN DAN SARAN

Bab ini memuat rangkuman dari hasil penelitian yang telah dilakukan, dengan menyoroti pokok-pokok temuan serta implikasinya terhadap pengembangan pengetahuan dan praktik di bidang terkait. Bab ini juga menjelaskan berbagai keterbatasan penelitian yang dapat memengaruhi hasil dan ruang lingkup kajian, sehingga memberikan gambaran mengenai aspek-aspek yang belum dapat dibahas secara menyeluruh. Berdasarkan keseluruhan proses dan temuan yang diperoleh, bab ini menyampaikan saran-saran yang diharapkan dapat memberikan manfaat bagi pengembangan akademik maupun penerapan praktis, sebagai kontribusi penelitian untuk inovasi di bidang yang relevan.