BAB V

SIMPULAN DAN SARAN

5.1 Simpulan

Penelitian ini telah berhasil membangun sistem *voice conversational* AI berbasis arsitektur *cascaded* yang mengintegrasikan tiga komponen utama secara sequential: Automatic Speech Recognition menggunakan Whisper Small, Small Language Model menggunakan Gemma 3 1B instruction-tuned, dan Text-to-Speech menggunakan VITS MMS Indonesian. Sistem yang dikembangkan mampu memproses *input audio* pengguna menjadi respons *audio* yang relevan dan natural, dengan arsitektur *cascaded* yang memberikan fleksibilitas untuk melakukan optimasi secara independen pada setiap komponen tanpa mempengaruhi komponen lainnya.

Proses optimasi *inference*-level telah berhasil diterapkan pada setiap komponen sistem untuk meningkatkan efisiensi komputasi pada lingkungan CPU-*only* tanpa memerlukan re-*training* model. Optimasi Whisper melalui konversi ke CTranslate2 dengan INT8 *quantization* menghasilkan penurunan *latency* sebesar 19.87% dan reduksi ukuran model sebesar 48.45%. Optimasi Gemma 3 1B melalui konversi ke format GGUF dengan Q8_0 *quantization* menghasilkan penurunan *latency* sebesar 44.85% dan peningkatan *throughput* sebesar 79.25%. Optimasi VITS melalui konversi ke ONNX *Runtime* dengan *graph optimization* dan INT8 *quantization* menghasilkan *speedup* 1.61x serta penurunan *Real-time Factor* sebesar 38.18%. Kombinasi teknik optimasi ini terbukti efektif dalam meningkatkan performa sistem pada perangkat dengan kapasitas komputasi terbatas.

Evaluasi performa sistem menunjukkan bahwa total *latency end-to-end* berkurang sebesar 30.09% dari 19.76 detik menjadi 13.81 detik, dengan *trade-off* minimal pada kualitas *output*. Akurasi transkripsi mengalami degradasi WER yang sangat minimal yaitu hanya 0.14%, kualitas respons bahasa menunjukkan penurunan *perplexity* sebesar 4.58%, dan *naturalness audio* sintesis tetap terjaga dengan reduksi ukuran model sebesar 22.81%. Hasil ini membuktikan bahwa sistem *voice conversational* AI yang dioptimasi dapat beroperasi secara efisien

pada perangkat dengan daya komputasi terbatas dengan *trade-off* minim, sehingga tetap mempertahankan kualitas *output* yang *acceptable* untuk aplikasi praktis, dan memberikan kontribusi penting bagi implementasi sistem *conversational* AI yang lebih mudah diakses tanpa memerlukan infrastruktur komputasi yang mahal.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilaksanakan, peneliti mengajukan beberapa saran dan rekomendasi untuk pertimbangan dalam studi lanjutan, mengingat adanya limitasi dan temuan yang diperoleh dalam penelitian ini. Rekomendasi-rekomendasi tersebut diuraikan sebagai berikut:

1. Komponen ASR (Whisper) teridentifikasi sebagai kontributor *latency* terbesar dengan proporsi 64.59% terhadap total *latency* sistem setelah optimasi. Penelitian selanjutnya dapat mengeksplorasi teknik optimasi tambahan lainnya, atau penggunaan varian Whisper yang lebih kecil seperti Whisper Tiny atau Whisper Base untuk menurunkan *latency* komponen ini secara lebih signifikan sambil tetap mempertahankan akurasi transkripsi yang *acceptable*.
2. Pengembangan lebih lanjut dapat mempertimbangkan implementasi pada *hardware accelerator* seperti Neural Processing Unit, TPU, atau GPU yang tersedia pada *cloud server* maupun perangkat lokal untuk memberikan peningkatan performa yang lebih signifikan. Selain itu, eksplorasi terhadap model-model yang lebih baru dengan arsitektur yang lebih efisien juga dapat memberikan potensi peningkatan performa yang lebih baik.
3. Penelitian lebih lanjut dapat mengeksplorasi pengaplikasian dari sistem *voice conversational AI* yang telah dioptimasi ini dalam berbagai bidang seperti *customer service automation*, *educational assistant*, atau *accessibility tools* untuk individu, sehingga memberikan manfaat praktis dari hasil penelitian ini.