

***INFERENCE-LEVEL OPTIMIZATION PADA SISTEM VOICE
CONVERSATIONAL AI BERBASIS CPU MELALUI METODE
QUANTIZATION DAN RUNTIME MODEL CONVERSION***



SKRIPSI

Kenny Budiarmo Lawson

00000081065

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

***INFERENCE-LEVEL OPTIMIZATION PADA SISTEM VOICE
CONVERSATIONAL AI BERBASIS CPU MELALUI METODE
QUANTIZATION DAN RUNTIME MODEL CONVERSION***



Diajukan sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana Komputer

Kenny Budiarto Lawson

00000081065

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG**

2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Kenny Budiarmo Lawson

Nomor Induk Mahasiswa : 00000081065

Program Studi : Sistem Informasi

Skripsi dengan judul: *INFERENCE-LEVEL OPTIMIZATION PADA SISTEM VOICE CONVERSATIONAL AI BERBASIS CPU MELALUI METODE QUANTIZATION DAN RUNTIME MODEL CONVERSION*

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 18 Desember 2025



Kenny Budiarmo Lawson

LEMBAR PERNYATAAN PENGGUNAAN BANTUAN KECERDASAN ARTIFISIAL (AI)

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Kenny Budiarmo Lawson
NIM : 00000081065
Program Studi : Sistem Informasi
Judul Skripsi : *Inference-Level Optimization pada Sistem Voice conversational AI Berbasis CPU Melalui Metode Quantization dan Runtime Model Conversion*

Dengan ini saya menyatakan secara jujur menggunakan bantuan Kecerdasan Artifisial (AI) dalam pengerjaan Tugas/Project/~~Tugas Akhir~~ sebagai berikut:

- ✓ Menggunakan AI sebagaimana diizinkan untuk membantu dalam menghasilkan ide-ide utama saja
- ✓ Menggunakan AI sebagaimana diizinkan untuk membantu menghasilkan teks pertama saja
- ✓ Menggunakan AI untuk menyempurnakan sintaksis dan tata bahasa untuk pengumpulan tugas
- ☐ Karena tidak diizinkan: Tidak menggunakan bantuan AI dengan cara apa pun dalam pembuatan tugas

Saya juga menyatakan bahwa:

- (1) Menyerahkan secara lengkap dan jujur penggunaan perangkat AI yang diperlukan dalam tugas melalui Formulir Penggunaan Perangkat Kecerdasan Artifisial (AI)
- (2) Saya mengakui bahwa saya telah menggunakan bantuan AI dalam tugas saya baik dalam bentuk kata, paraphrase, penyertaan ide atau fakta penting yang disarankan oleh AI dan saya telah menyantumkan dalam sitasi serta referensi
- (3) Terlepas dari pernyataan di atas, tugas ini sepenuhnya merupakan karya saya sendiri

Tangerang, 18 Desember 2025
Yang Menyatakan,



Kenny Budiarmo Lawson

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Kenny Budiarto Lawson
Nomor Induk Mahasiswa : 00000081065
Program Studi : Sistem Informasi
Jenjang : S1
Jenis Karya : Skripsi
Judul Karya Ilmiah : *Inference-Level Optimization pada Sistem Voice conversational AI Berbasis CPU Melalui Metode Quantization dan Runtime Model Conversion*

Menyatakan dengan sesungguhnya bahwa saya bersedia* (**pilih salah satu**):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - ☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - ☐ Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 18 Desember 2025



Kenny Budiarto Lawson

* Pilih salah satu opsi tanpa harus menghapus opsi yang tidak dipilih

** Jika tidak bisa membuktikan LoA jurnal/ HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN

KATA PENGANTAR

Puji Syukur kepada Tuhan Yang Maha Esa atas berkat dan karunia-Nya dalam selesainya penulisan Laporan Skripsi ini dengan judul: “*Inference-Level Optimization* pada Sistem *Voice conversational* AI Berbasis CPU melalui Metode *Quantization* dan *Runtime Model Conversion*.” Saya menyadari bahwa sulit untuk menyelesaikan Laporan Skripsi ini tanpa bantuan dan bimbingan dari berbagai pihak. Maka dari itu, saya mengucapkan terima kasih kepada:

1. Dr. Ir. Andrey Andoko, M.Sc., Ph.D, selaku Rektor Universitas Multimedia Nusantara.
2. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti S. Kom., M. Kom., selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara.
4. Dr. Erick Fernando S. Kom., M. S. I., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga dan teman-teman saya yang telah memberikan bantuan dukungan material dan moral dalam proses pengerjaan, sehingga penulis dapat menuntaskan tugas akhir ini.

Semoga laporan ini dapat memberikan dampak positif bagi berbagai pihak yang terlibat dalam penyusunan laporan, dan dapat memberikan kontribusi sebagai wawasan bagi pembaca, pihak kampus, serta penelitian di masa depan.

Tangerang, 18 Desember 2025



Kenny Budiarto Lawson

INFERENCE-LEVEL OPTIMIZATION PADA SISTEM VOICE CONVERSATIONAL AI BERBASIS CPU MELALUI METODE QUANTIZATION DAN RUNTIME MODEL CONVERSION

Kenny Budiarto Lawson

ABSTRAK

Voice conversational AI semakin dibutuhkan seiring meningkatnya kebutuhan pengguna akan interaksi yang alami, efisien, dan bebas kendali manual dalam berbagai aplikasi seperti asisten virtual, layanan pelanggan, dan sistem edukasi. Namun, arsitektur *cascaded* yang menggabungkan *Automatic Speech Recognition* (ASR), *Small Language Model* (SLM), dan *Text-to-Speech* (TTS) masih menghadapi kendala berupa *latency* tinggi dan beban komputasi besar, terutama ketika dijalankan pada lingkungan *CPU-only* yang umum digunakan pada *server* berbiaya rendah maupun perangkat lokal. Penelitian ini menerapkan teknik optimasi pada tahap *inference* tanpa melakukan pelatihan ulang model, meliputi konversi model Whisper Small menggunakan CTranslate2 dengan INT8 *quantization*, GGUF *quantization* untuk Gemma 3 1B, serta konversi VITS MMS ke ONNX dengan optimasi graf dan *static* INT8 *quantization*. Evaluasi dilakukan menggunakan dataset Common Voice Indonesian untuk ASR dan kumpulan *prompt* teks yang disusun secara manual untuk SLM dan TTS, dengan metrik pengujian berupa *Word Error Rate*, *Character Error Rate*, *latency*, *Real-time Factor*, *perplexity*, *tokens-per-second*, dan ukuran model. Hasil penelitian menunjukkan bahwa optimasi Whisper menggunakan CTranslate2 INT8 menurunkan *latency* sebesar 19,87% dengan peningkatan *error rate* yang minimal; optimasi Gemma menggunakan GGUF Q8_0 menghasilkan penurunan *latency* sebesar 44,85% dan peningkatan *throughput* 26,78%; sedangkan optimasi VITS memberikan percepatan 1,61x dengan penurunan *latency* sebesar 37,74%. Secara keseluruhan, sistem *end-to-end* mengalami penurunan *latency* sebesar 30,09% dari 19,76 detik menjadi 13,81 detik, disertai dengan reduksi ukuran model pada semua komponen, menunjukkan bahwa kombinasi teknik optimasi *inference-level* efektif meningkatkan kinerja sistem *voice conversational AI* pada lingkungan *CPU-only*.

Kata kunci: *Cascaded architecture, inference optimization, model runtime conversion, quantization, voice conversational AI*

INFERENCE-LEVEL OPTIMIZATION IN CPU-BASED CONVERSATIONAL AI VOICE SYSTEM USING QUANTIZATION AND RUNTIME MODEL CONVERSION METHOD

Kenny Budiarmo Lawson

ABSTRACT (English)

Conversational AI is increasingly needed as users need natural, efficient, and manual-free interactions in various applications such as virtual assistants, customer service, and education systems. However, cascaded architectures that combine Automatic Speech Recognition (ASR), Small Language Model (SLM), and Text-to-Speech (TTS) still face obstacles in the form of high latency and large computational burdens, especially when running in CPU-only environments commonly used on low-cost servers or local devices. This study applies optimization techniques at the inference stage without retraining the model, including the conversion of the Whisper Small model using CTranslate2 with INT8 quantization, GGUF quantization for Gemma 3 1B, and the conversion of VITS MMS to ONNX with graph optimization and static INT8 quantization. The evaluation was conducted using the Indonesian Common Voice dataset for ASR and a manually compiled collection of prompt texts for SLM and TTS, with test measurements in the form of Word Error Rate, Character Error Rate, latency, Real-time Factor, perplexity, tokens-per-second, and model size. The results show that Whisper optimization using CTranslate2 INT8 reduces latency by 19.87% with minimal increase in error rate; Gemma optimization using GGUF Q8_0 results in a 44.85% decrease in latency and a 26.78% increase in throughput; while VITS optimization provides a $1.61\times$ acceleration with a 37.74% decrease in latency. Overall, the end-to-end system experiences a 30.09% decrease in latency from 19.76 seconds to 13.81 seconds, accompanied by a reduction in model size across all components, indicating that the combination of inference-level optimization techniques effectively improves the performance of AI conversational voice systems in CPU-only environments.

Keywords: Cascaded architecture, inference optimization, model runtime conversion, quantization, voice conversational AI

DAFTAR ISI

HALAMAN PERNYATAAN TIDAK PLAGIAT	ii
LEMBAR PERNYATAAN PENGGUNAAN BANTUAN KECERDASAN ARTIFISIAL (AI)	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
KATA PENGANTAR.....	v
ABSTRAK	vi
ABSTRACT (English)	vii
DAFTAR ISI.....	viii
DAFTAR TABEL	xi
DAFTAR RUMUS	xii
DAFTAR GAMBAR.....	xiii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	5
1.3 Batasan Masalah.....	5
1.4 Tujuan dan Manfaat Penelitian	6
1.4.1 Tujuan Penelitian.....	6
1.4.2 Manfaat Penelitian	7
1.5 Sistematika Penulisan	7
BAB II LANDASAN TEORI	10
2.1 Penelitian Terdahulu.....	10
2.2 Teori yang Berkaitan	17
2.2.1 <i>Artificial Intelligence (AI)</i>	17
2.2.2 <i>Deep learning</i>	18
2.2.3 <i>Inference-Level Optimization</i>	19
2.2.4 <i>Small Language Model (SLM) / Large Language Model (LLM)</i>	20
2.2.5 <i>Speech-to-Text (STT) / Automatic Speech Recognition (ASR)</i> ...	21
2.2.6 <i>Text-to-Speech (TTS)</i>	22
2.3 <i>Framework/Algoritma yang digunakan</i>	23

2.3.1	Whisper	23
2.3.2	Gemma 3 1B	24
2.3.3	VITS MMS.....	26
2.3.4	<i>Quantization</i>	28
2.3.5	<i>ONNX Runtime Optimization</i>	31
2.3.6	<i>CTranslate2 Optimization</i>	32
2.3.7	<i>GGUF Quantization Format</i>	33
2.4	Rumus Metrik Evaluasi	34
2.5	<i>Tools/software yang digunakan</i>	37
2.5.1	Python	37
2.5.2	Visual Studio Code.....	37
2.5.3	<i>Hardware yang Digunakan</i>	38
BAB III METODOLOGI PENELITIAN		39
3.1	Gambaran Umum Objek Penelitian.....	39
3.2	Metode Penelitian	40
3.2.1	Alur Penelitian.....	41
3.3	Teknik Pengumpulan Data.....	45
3.3.1	<i>Dataset Evaluasi ASR (Whisper)</i>	46
3.3.2	<i>Dataset Evaluasi SLM (Gemma 3 1B) dan TTS (VITS)</i>	47
3.4	Teknik Optimasi Model	48
3.4.1	Optimasi Model ASR (Whisper).....	48
3.4.2	Optimasi Model SLM (Gemma 3 1B)	49
3.4.3	Optimasi Model TTS (VITS)	50
3.5	Teknik Evaluasi Model	51
3.5.1	Evaluasi Model ASR (Whisper).....	51
3.5.2	Evaluasi Model SLM (Gemma 3 1B).....	53
3.5.3	Evaluasi Model TTS (VITS).....	55
BAB IV ANALISIS DAN HASIL PENELITIAN		57
4.1	Persiapan Data.....	57
4.2.1	Persiapan <i>Dataset Audio</i> untuk Evaluasi ASR/STT.....	57
4.2.2	Persiapan <i>Dataset Teks</i> untuk Evaluasi SLM	59
4.2.3	Persiapan <i>Dataset Teks</i> untuk Evaluasi TTS	60

4.2	Implementasi Model <i>Baseline</i>	60
4.2.1	Implementasi <i>Baseline</i> Whisper untuk ASR/STT.....	61
4.2.2	Implementas <i>Baseline</i> Gemma 3 1B untuk SLM	62
4.2.3	Implementasi <i>Baseline</i> VITS untuk TTS	63
4.3	Optimasi Model	65
4.3.1	Optimasi Model Whisper dengan CTranslate2 dan <i>Quantization</i>	65
4.3.2	Optimasi Model Gemma 3 1B dengan INT8 <i>Quantization</i>	66
4.3.3	Optimasi Model VITS dengan ONNX <i>Runtime</i> dan <i>Quantization</i>	68
4.4	Evaluasi Performa Model	70
4.4.1	Evaluasi Performa Model Whisper	70
4.4.2	Evaluasi Performa Model Gemma 3 1B.....	73
4.4.3	Evaluasi Performa Model VITS	75
4.5	Integrasi Sistem <i>End-to-end Pipeline</i>	77
4.5.1	Perancangan <i>Deployment</i> Sistem	77
4.5.2	Hasil <i>Interface Deployment</i>	84
4.6	Analisis dan Interpretasi Hasil.....	86
4.6.1	Hasil Pembangunan Sistem <i>Voice conversational AI</i> Berbasis <i>Arsitektur Cascaded</i>	86
4.6.2	Penerapan dan Optimasi Proses <i>Inference</i> pada Setiap Komponen.....	87
4.6.3	Hasil Evaluasi	89
BAB V SIMPULAN DAN SARAN		93
5.1	Simpulan.....	93
5.2	Saran.....	94
DAFTAR PUSTAKA.....		95
GLOSARIUM.....		105
LAMPIRAN.....		107

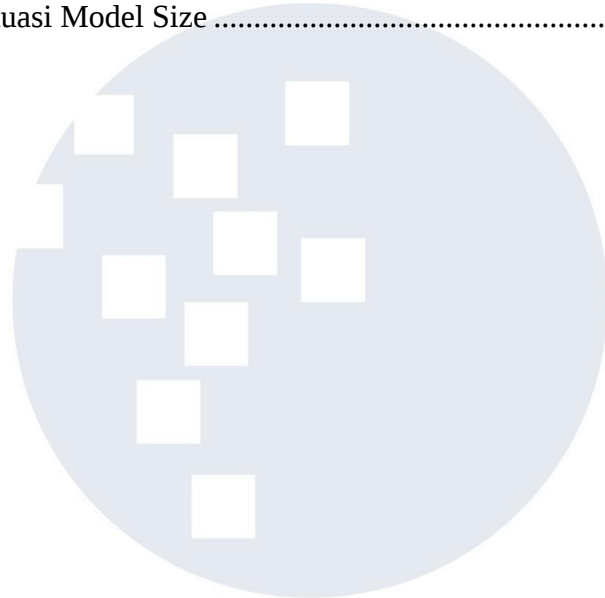
DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	10
Tabel 3. 1 Informasi <i>Dataset</i> Evaluasi ASR	46
Tabel 3. 2 Informasi <i>Dataset</i> Evaluasi SLM dan TTS.....	47
Tabel 3. 3 Metrik Evaluasi Model ASR.....	51
Tabel 3. 4 Metrik Evaluasi SLM.....	53
Tabel 3. 5 Metrik Evaluasi Model TTS	55
Tabel 4. 1 Evaluasi Model Whisper	70
Tabel 4. 2 Evaluasi Model Gemma 3 1B	73
Tabel 4. 3 Evaluasi Model VITS.....	75
Tabel 4. 4 Penerapan Optimasi Masing-Masing Model.....	87
Tabel 4. 5 Hasil Evaluasi Performa Setelah Optimasi Model.....	89
Tabel 4. 6 Informasi <i>Latency End-to-end Pipeline</i>	90



DAFTAR RUMUS

Rumus 2. 1 Evaluasi <i>Word Error Rate</i>	34
Rumus 2. 2 Evaluasi <i>Character Error Rate</i>	35
Rumus 2. 3 Evaluasi <i>Real-time Factor</i>	35
Rumus 2. 4 Evaluasi <i>Latency</i>	36
Rumus 2. 5 Evaluasi <i>Token Generation Rate</i>	36
Rumus 2. 6 Evaluasi <i>Model Size</i>	36



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

Gambar 2. 1 Arsitektur Model Whisper.....	23
Gambar 2. 2 Arsitektur Model VITS	27
Gambar 3. 1 Kerangka Sistem <i>Voice conversational AI</i>	39
Gambar 3. 2 Alur Penelitian.....	41
Gambar 4. 1 Kode Python Pengambilan <i>Dataset Audio</i>	57
Gambar 4. 2 Hasil CSV <i>Dataset Audio</i>	58
Gambar 4. 3 Hasil Persiapan <i>Dataset</i> Teks SLM	59
Gambar 4. 4 Hasil Persiapan <i>Dataset</i> Teks TTS	60
Gambar 4. 5 Proses Implementasi <i>Baseline</i> Whisper	61
Gambar 4. 6 Proses Implementasi <i>Baseline</i> Gemma 3 1B.....	62
Gambar 4. 7 Proses Implmentasi <i>Baseline</i> VITS.....	64
Gambar 4. 8 Proses Optimasi Model Whisper.....	65
Gambar 4. 9 Proses Optimasi Model Gemma 3 1B.....	66
Gambar 4. 10 Konfigurasi Model Gemma 3 1B	67
Gambar 4. 11 Proses Knoversi <i>Runtime</i> Model VITS	68
Gambar 4. 12 Proses Graph Optimization <i>Runtime</i> Model VITS.....	68
Gambar 4. 13 Proses <i>Quantization</i> Model VITS	69
Gambar 4. 14 Proses <i>Load</i> Setiap Komponen	78
Gambar 4. 15 Kode Pemrosesan ASR	79
Gambar 4. 16 Kode Pemrosesan SLM.....	80
Gambar 4. 17 Proses Penambahan Prosody Injection VITS MMS	82
Gambar 4. 18 Kode Pemrosesan TTS	83
Gambar 4. 19 <i>User Interface</i> Sistem <i>Voice conversational AI</i>	84
Gambar 4. 20 <i>Interface</i> Hasil Pemrosesan Percakapan	85

UIN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran A Turnitin Similarity Report	107
Lampiran B Formulir Konsultasi Skripsi.....	108
Lampiran C Formulir Penggunaan Perangkat Kecerdasan Buatan (AI).....	109

