PENGEMBANGAN TEKNOLOGI TEXT-TO-SPEECH (TTS) BAHASA INDONESIA MENGGUNAKAN VOICE CLONING BERBASIS KONVERSI GRAPHEME-TO-PHONEME (G2P)

(Skripsi)

Oleh

MUHAMMAD AL FAHREZI

NPM 1955061009

FAKULTAS TEKNIK

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2026

PENGEMBANGAN TEKNOLOGI TEXT-TO-SPEECH (TTS) BAHASA INDONESIA MENGGUNAKAN VOICE CLONING BERBASIS KONVERSI GRAPHEME-TO-PHONEME (G2P)

Oleh

Muhammad Al Fahrezi

1955061009

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar

SARJANA TEKNIK

Pada

Program Studi Teknik Informatika

Jurusan Teknik Elektro

Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2026

ABSTRAK

PENGEMBANGAN TEKNOLOGI TEXT-TO-SPEECH (TTS) BAHASA INDONESIA MENGGUNAKAN VOICE CLONING BERBASIS KONVERSI GRAPHEME-TO-PHONEME (G2P)

OLEH

MUHAMMAD AL FAHREZI

Berbagai kondisi medis seperti *stroke*, *cerebral palsy*, cedera otak, penyakit neurodegeneratif, dan kanker kepala–leher dapat menyebabkan kehilangan kemampuan bicara sehingga memerlukan teknologi bantu komunikasi. Text-to-Speech (TTS) menjadi solusi yang memungkinkan teks diubah menjadi suara sintesis untuk mendukung komunikasi pengguna. Penelitian ini mengembangkan sistem TTS Bahasa Indonesia dengan menerapkan voice cloning berbasis Grapheme-to-Phoneme (G2P) untuk meningkatkan ketepatan pelafalan. Sistem dibangun menggunakan arsitektur Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) dengan modul G2P berbasis BERT untuk menghasilkan representasi fonem kontekstual. Evaluasi menunjukkan model G2P mencapai akurasi 0,9907 dan Phoneme Error Rate (PER) 2,15%. Namun, integrasi G2P pada sistem TTS belum meningkatkan kualitas suara, di mana sistem tanpa G2P memperoleh Mean Opinion Score (MOS) 3,80 dan dengan G2P sebesar 2,89. Hasil ini menunjukkan perlunya optimasi lebih lanjut pada integrasi G2P dalam model VITS untuk meningkatkan kualitas suara sintetis Bahasa Indonesia.

Kata kunci: *Text-to-Speech*, *Grapheme-to-Phoneme*, *Voice Cloning*, BERT, VITS, Bahasa Indonesia

ABSTRACT

DEVELOPMENT OF AN INDONESIAN TEXT-TO-SPEECH (TTS) SYSTEM USING VOICE CLONING BASED ON GRAPHEME-TO-PHONEME (G2P) CONVERSION

by

MUHAMMAD AL FAHREZI

*Various medical conditions such as stroke, cerebral palsy, brain injury, neurodegenerative diseases, and head–neck cancer can lead to speech impairment, thereby requiring assistive communication technologies. Text-to-Speech (*TTS*) systems provide a solution by converting text into synthesized speech to support user communication. This study develops an Indonesian Text-to-Speech system by implementing voice cloning with a Grapheme-to-Phoneme (*G2P*) approach to improve pronunciation accuracy. The system is built using the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (*VITS*) architecture, integrated with a* BERT *based G2P module to generate contextual phoneme representations. Evaluation results show that the* G2P *model achieves an accuracy of 0.9907 and a Phoneme Error Rate (*PER*) of 2.15%. However, the integration of* G2P *into the* TTS *system does not improve speech quality, as the system without* G2P *achieves a Mean Opinion Score (*MOS*) of 3.80, while the system with* G2P *obtains a* MOS *of 2.89. These findings indicate that further optimization is required in integrating the* G2P *module into the* VITS *model to enhance the quality of synthesized Indonesian speech.*

Keywords: *Text-to-Speech*, *Grapheme-to-Phoneme*, *Voice Cloning*, BERT, VITS, Indonesian *Language*

Judul Skripsi	: PENGEMBANGAN TEKNOLOGI TEXT-TO-SPEECH (TTS) BAHASA INDONESIA MENGGUNAKAN VOICE CLONING BERBASIS KONVERSI GRAPHEME-TO-PHONEME (G2P)
Nama Mahasiswa	: *Muhammad Al Fahrezi*
Nomor Pokok Mahasiswa	: 1955061009
Program Studi	: Teknik Informatika
Jurusan	: Teknik Elektro
Fakultas	: Teknik

Menyetujui

1.Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Dr. Ir. M. Komarudin, S.T., M.T	**Yessi Mulyani, S.T., M.T.**
NIP. 196812071997031006	NIP. 197312262000122001

2. Mengetahui

Ketua Jurusan	Ketua Program Studi
Teknik Elektro	Teknik Informatika
Herlinawati, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 197103141999032001	NIP. 197312262000122001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Ir. Komarudin, S.T., M.T.** .....

Sekretaris : **Yessi Mulyani, S.T., M.T.** .....

Penguji : **Ir. Titin Yulianti, S.T., M.Eng.** .....

2. Dekan Fakultas Teknik

Dr. H. Ahmad Herison, S.T., M.T.
NIP. 196910302000031001

Tanggal Lulus Ujian Skripsi: 2 Februari 2026

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini, menyatakan bahwa skripsi saya dengan judul "Pengembangan Teknologi *Text-To-Speech* (TTS) Bahasa Indonesia Menggunakan *Voice Cloning* Berbasis Konversi *Grapheme-To-Phoneme* (G2P)" dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung.

Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 2 Februari 2026

Pembuat Pernyataan,

Muhammad Al Fahrezi

NPM. 1955061009

RIWAYAT HIDUP

Penulis Bernama Muhammad Al Fahrezi Lahir pada tanggal 20 Januari 2001. Penulis merupakan anak sulung dari pasangan Rano Kalestian dan Ibu Siti Aisyah. Penulis mengawali di SD Negeri 08 Pagi Jakarta. Selanjutnya, penulis melanjutkan pendidikan di MTS Nurul Falah, lalu melanjutkan ke SMA Negeri 2 Pringsewu dan tamat pada tahun 2019. Kemudian pada tahun yang sama penulis diterima sebagai mahasiswa baru pada Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung melalui jalur SMMPTN.

Selama menjadi mahasiswa, penulis aktif dalam beberapa kegiatan. Penulis melaksankan Kuliah kerja Nyata (KKN) periode II tahun 2022 di Desa Sidomulyo, Kecamatan Sumberejo, Kabupaten Tanggamus, Provinsi Lampung. Pada Tahun 2023 mengikuti Program Magang dan Studi Independen Bersertifikat Kampus Merdeka di perusahaan Soca AI yang berada di Bandung untuk posisi AI Engineer.

MOTTO

“Sesungguhnya bersama kesulitan ada kemudahan. Maka apabila kamu telah selesai (dari sesuatu urusan), tetaplah bekerja keras (untuk urusan yang lain). Dan hanya kepada Tuhanmulah kamu berharap”

(Q.S Al-Insyirah: 6-8)

“Sesungguhnya Allah tidak akan mengubah keadaan suatu kaum sampai mereka mengubah keadaan yang ada pada diri mereka sendiri.”

(QS. Ar-Ra’d: 11)

“Dan bahwa manusia hanya memperoleh apa yang telah diusahakannya.”

(QS. An-Najm: 39)

“Begitu kamu berhenti belajar, kamu mulai mati.”

(Albert Einstein)

“Pendidikan adalah senjata paling ampuh yang bisa kamu gunakan untuk mengubah dunia.”

(Nelson Mandela)

PERSEMBAHAN

Sujud syukur kupersembahkan ke hadirat Allah SWT, Tuhan Yang Maha Agung dan Maha Tinggi. Atas rahmat, kasih sayang, dan takdir-Mu, Engkau tuntun langkahku hingga mampu menjadi pribadi yang terus belajar, berpikir, berilmu, beriman, dan bersabar dalam setiap proses kehidupan.

Kupersembahkan karya ilmiah ini teruntuk:

"Karya ini kupersembahkan dengan penuh cinta, hormat, dan rasa terima kasih yang tak terhingga kepada kedua orang tuaku tercinta. Terima kasih atas setiap doa yang tak pernah terputus, atas pengorbanan yang tak pernah terhitung, serta atas cinta yang selalu menjadi tempatku kembali ketika dunia terasa berat. Segala pencapaian ini tidak akan pernah berarti tanpa restu dan ridha kalian. Semoga setiap huruf dalam karya ini menjadi saksi atas perjuangan dan ketulusan kalian dalam membesarkanku."

"Untuk kedua adikku tersayang, terima kasih telah menjadi bagian dari perjalanan ini. Semoga apa yang kakak raih hari ini dapat menjadi penyemangat bagi kalian untuk terus belajar, bermimpi setinggi mungkin, dan berani menghadapi setiap tantangan. Percayalah, tidak ada usaha yang sia-sia selama kita terus berdoa dan berikhtiar."

"Semoga keberhasilan ini menjadi langkah awal menuju masa depan yang lebih baik, membuka jalan untuk cita-cita yang lebih besar, dan menjadi amal kebaikan yang diridhai Allah SWT. Aamiin."

SANWANCANA

Segala puji dan syukur senantiasa penulis panjatkan ke hadirat Allah SWT, Tuhan semesta alam, Yang Maha Pengasih lagi Maha Penyayang. Atas limpahan rahmat, taufik, dan hidayah-Nya, penulis diberikan kekuatan, kesabaran, serta keteguhan hati dalam menempuh setiap proses hingga akhirnya skripsi yang berjudul *"*PENGEMBANGAN TEKNOLOGI *TEXT-TO-SPEECH* (TTS) BAHASA INDONESIA MENGGUNAKAN *VOICE CLONING* BERBASIS KONVERSI *GRAPHEME-TO-PHONEME* (G2P*)"* ini dapat terselesaikan. Penyusunan skripsi ini merupakan salah satu syarat untuk memperoleh gelar Sarjana Teknik pada Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung. Namun lebih dari itu, perjalanan ini adalah proses pembelajaran panjang tentang ketekunan, tanggung jawab, dan keikhlasan dalam berusaha. Dalam setiap tahap penelitian dan penulisan, penulis menyadari bahwa karya ini tidak akan terwujud tanpa dukungan, bimbingan, serta doa dari berbagai pihak. Oleh karena itu, dengan penuh rasa hormat dan ketulusan, penulis menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Hi. Ahmad Herison, S.T., M.T., selaku Dekan Fakultas Teknik Universitas Lampung (FT Unila) periode 2025–2029.
2. Bapak Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc., selaku Dekan Fakultas Teknik Universitas Lampung periode 2021–2025.
3. Ibu Herlinawati, S.T., M.T., selaku Ketua Jurusan Teknik Elektro Universitas Lampung, atas arahan dan dukungan yang diberikan selama penulis menempuh pendidikan.
4. Ibu Yessi Mulyani, S.T., M.T., selaku Ketua Program Studi Teknik Informatika sekaligus Pembimbing Pendamping, yang telah memberikan nasihat,

motivasi, serta dengan penuh kesabaran membimbing dan mengarahkan penulis dalam proses penyusunan skripsi ini.

5. Bapak Dr. Ir. M. Komarudin, S.T., M.T., selaku Pembimbing Utama, yang dengan kesabaran dan ketelatenan telah meluangkan waktu, tenaga, serta pikiran untuk membimbing penulis hingga skripsi ini terselesaikan dengan baik.
6. Ibu Ir. Titin Yulianti, S.T., M.Eng., selaku Dosen Penguji, yang telah memberikan kritik, saran, dan masukan yang membangun demi penyempurnaan skripsi ini.
7. Seluruh Dosen Program Studi Teknik Informatika yang telah membagikan ilmunya kepada penulis.
8. Seluruh jajaran staf administrasi atas bantuannya dalam menyelesaikan urusan administrasi di Jurusan Teknik Elektro Universitas Lampung.
9. Teman-teman seperjuangan di Program Studi Teknik Informatika yang telah menjadi bagian dari perjalanan ini. Terima kasih atas kebersamaan, diskusi, kerja sama, serta dukungan yang saling menguatkan dalam menghadapi setiap tantangan akademik.
10. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah membantu baik secara langsung maupun tidak langsung dalam penyusunan skripsi ini.

Semoga Allah SWT membalas segala kebaikan dan bantuan yang telah diberikan kepada penulis. Penulis berharap skripsi ini dapat membawa manfaat bagi para pembaca serta menjadi bekal penulis dalam mengembangkan dan menerapkan ilmu yang telah dipelajari.

Bandar Lampung, 2 Februari 2026
Penulis

Muhammad Al Fahrezi
NPM 1955061009

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Berbagai kondisi medis diketahui menjadi penyebab utama munculnya kebutuhan teknologi bantu komunikasi, seperti *stroke*, *cerebral palsy*, *cedera otak*, penyakit neurodegeneratif, serta kanker kepala dan leher yang dapat mengakibatkan kehilangan kemampuan bicara[1]. Teknologi bantu komunikasi ini berperan penting dalam membantu individu dengan gangguan bicara untuk tetap dapat menyampaikan pesan secara verbal. Dalam konteks ini, teknologi *Text-to-Speech* (TTS) menjadi salah satu solusi yang efektif karena memungkinkan teks diubah menjadi suara sintesis, sehingga mendukung komunikasi, kemandirian, dan partisipasi sosial pengguna.

Dalam aplikasi TTS, salah satu proses yang berperan terhadap kualitas hasil Dari TTS adalah konversi *Grapheme-to-Phoneme* (G2P)[2]. G2P adalah proses konversi dari grafem, yaitu unit terkecil dari sistem penulisan seperti huruf dalam alfabet, menjadi fonem, yang merupakan unit suara terkecil dalam bahasa lisan[3]. Proses ini sangat penting karena setiap bahasa memiliki aturan fonologis yang berbeda, dan akurasi konversi ini berpengaruh langsung terhadap kualitas suara yang dihasilkan. Misalnya, dalam Bahasa Indonesia, satu huruf atau kombinasi huruf bisa mewakili satu atau beberapa bunyi yang berbeda tergantung pada konteksnya, sehingga pemahaman yang mendalam terhadap struktur fonologis bahasa sangat diperlukan untuk memastikan hasil TTS yang akurat. Kesalahan dalam proses G2P dapat mengakibatkan distorsi dalam pengucapan, intonasi, dan bahkan arti dari kata yang dihasilkan, yang dapat mengurangi efektivitas dan kenyamanan penggunaan sistem TTS

Bahasa Indonesia, sebagai salah satu bahasa dengan jumlah penutur terbesar di dunia, memiliki kompleksitas dan keunikan fonologis yang tidak selalu ditemukan dalam bahasa lain[4]. Misalnya, struktur suku kata yang relatif sederhana namun kaya dengan variasi vokal, konsonan, dan diftong memberikan tantangan tersendiri dalam proses konversi G2P. Meskipun demikian, perhatian terhadap pengembangan teknologi pemrosesan bahasa untuk Bahasa Indonesia masih tergolong minim, terutama dalam konteks konversi G2P yang akurat untuk aplikasi TTS. Kebanyakan penelitian G2P yang ada saat ini lebih berfokus pada bahasa-bahasa dengan penutur yang lebih dominan di dunia teknologi, seperti Bahasa Inggris atau Mandarin, sementara Bahasa Indonesia belum mendapatkan porsi perhatian yang sepadan. Di sisi lain, perkembangan teknologi voice cloning, yang memungkinkan penciptaan suara sintetis yang tidak hanya menyerupai, tetapi juga mereplikasi karakteristik unik suara individu tertentu, semakin meningkatkan kebutuhan sistem G2P yang presisi.

Bidirectional Encoder Representations from Transformers (BERT) telah menjadi pilar utama dalam berbagai aplikasi *Nature Language Processing* (NLP). Ketiga model ini, masing-masing dengan arsitektur dan mekanisme pembelajaran yang berbeda, telah terbukti memberikan hasil yang mengesankan dalam berbagai tugas NLP, termasuk dalam konversi G2P. LSTM, dengan kemampuannya menangkap ketergantungan jangka panjang dalam urutan data, telah banyak digunakan dalam tugas-tugas sekuensial [5]. Sementara itu, BERT, dengan pendekatan transformer-nya yang memanfaatkan pemahaman konteks dalam dua arah [6], telah merevolusi berbagai aplikasi NLP dengan menghasilkan representasi bahasa yang lebih kaya dan kontekstual.

Namun, meskipun ketiga model ini telah menunjukkan potensi yang besar, hingga saat ini belum ada kajian komprehensif yang membandingkan kinerja mereka secara langsung dalam konteks konversi G2P untuk Bahasa Indonesia. Padahal, mengingat keunikan fonologis Bahasa Indonesia, pemilihan model yang tepat dapat berdampak signifikan terhadap akurasi dan kualitas hasil konversi G2P. Dalam konteks ini, penelitian ini bertujuan untuk tidak hanya mengembangkan model konversi G2P menggunakan LSTM, dan BERT, tetapi juga melakukan evaluasi mendalam terhadap kinerja masing-masing model.

Selain itu, untuk mencapai reproduksi suara yang lebih natural dan realistis dalam aplikasi voice cloning, penelitian ini memanfaatkan model Conditional Variational Autoencoder (CVAE) with Adversarial Learning for End-to-End Text-to-Speech (VITS). CVAE, sebagai salah satu pendekatan generatif yang canggih, memiliki kemampuan untuk menangkap distribusi kompleks dari data suara dan menghasilkan variasi output yang lebih kaya dan beragam. Dengan menggabungkan CVAE dengan Adversarial Learning, yang menggunakan jaringan adversarial untuk meningkatkan kualitas dan realisme suara yang dihasilkan, model ini mampu menciptakan suara sintetis yang tidak hanya akurat dari segi fonetis tetapi juga mendekati karakteristik suara manusia secara alami[7].

Sistem VITS digunakan dalam penelitian ini merupakan salah satu terobosan terbaru dalam teknologi TTS. Berbeda dengan pendekatan TTS tradisional yang sering memisahkan proses pelatihan akustik dan prosodi, VITS mengintegrasikan semua tahap pembuatan suara dalam satu kerangka kerja end-to-end[8]. Ini memungkinkan model untuk belajar secara langsung dari data mentah dan menghasilkan suara dengan intonasi, ritme, dan ekspresi yang lebih natural.

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan teknologi TTS Bahasa Indonesia, serta membuka peluang lebih lanjut dalam pengembangan aplikasi voice cloning yang dapat digunakan untuk berbagai keperluan, seperti asisten virtual, pengisi suara, dan lain-lain.

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini adalah

1. Bagaimana mengembangkan Teknologi TTS berbasis bahasa Indonesia menggunakan BERT
2. Bagaimana menggunakan *voice cloning* berbasis G2P untuk pengembangan teknologi TTS bahasa Indonesia
3. Bagaimana kinerja model G2P dalam mengonversi teks menjadi representasi fonem?

4. Bagaimana pengaruh hasil konversi Grapheme-to-Phoneme G2P terhadap kualitas pencuapan kata yang dihasilkan oleh model VITS pada sistem Text-to-Speech TTS?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah

1. Mengembangkan teknologi TTS berbasis bahasa Indonesia.
2. Menerapkan voice cloning berbasis G2P untuk pengembangan teknologi TTS dalam bahasa Indonesia.
3. Mengukur dan menganalisis kinerja model Grapheme-to-Phoneme
4. Mengukur dan menganalisis pengaruh hasil konversi G2P terhadap kualitas pencuapan kata yang dihasilkan oleh model VITS dalam sistem TTS.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini yaitu

1. Memperluas pemahaman tentang penerapan *deep learning* dalam G2P untuk Bahasa Indonesia bagi Peneliti
2. Diharapkan dapat Meningkatkan kualitas layanan berbasis suara dalam Bahasa Indonesia, seperti asisten virtual dan sistem navigasi untuk masyarakat umum
3. Diharapkan dapat Membantu pengembang memilih model deep learning yang paling efektif untuk konversi G2P bagi peneliti di bidang pengembangan teknologi TTS

1.5 Batasan Masalah

Adapun batasan masalah pada penelitian ini sebagai berikut:

1. Penelitian ini terbatas pada penggunaan model deep learning yaitu BERT untuk G2P konversi
2. Penelitian ini berfokus pada aturan fonologis standar Bahasa Indonesia tanpa memperhatikan dialek atau logat lokal yang mungkin memengaruhi pengucapan.
3. Penelitian ini menggunakan dataset yang terbatas pada teks dan suara dalam Bahasa Indonesia yang tersedia secara publik atau yang dikumpulkan secara khusus untuk penelitian ini.

4. Penelitian ini tidak mencakup aspek lain dari teknologi TTS atau voice cloning, seperti sintesis prosodi atau deteksi emosi, yang juga dapat mempengaruhi kualitas suara yang dihasilkan.

1.6 Sistematika Penulisan Skripsi

Sistematika penulisan skripsi ini antara lain:

BAB I PENDAHULUAN

Bab ini merupakan pendahuluan berisi tentang latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah dari penelitian yang dilakukan terkait Pengembangan TTS dengan G2P.

BAB II TINJAUAN PUSTAKA

Bab ini merupakan Tinjauan Pustaka berisi definisi maupun istilah yang diambil dari berbagai sumber (buku, jurnal, dan sebagainya). BAB II berisi tentang pengertian Text to Speech, Voice Clone, VITS, dan lain lain

BAB III METODOLOGI PENELITIAN

Bab ini merupakan metodologi penelitian berisi tempat dan waktu penelitian, jadwal penelitian, alat dan bahan penelitian dan tahapan penelitian

BAB IV HASIL DAN PEMBAHASAN

Bab ini merupakan hasil dan pembahasan berisi tentang hasil penelitian yang dilakukan serta pengolahan data dari hasil penelitian.

BAB V KESIMPULAN

Bab ini merupakan kesimpulan berisi kesimpulan dan saran berdasarkan penelitian.

II. TINJAUAN PUSTAKA

2.1 *Artificial Intelligence* (AI)

Artificial Intelligence (AI) adalah bidang ilmu komputer yang berfokus pada pengembangan sistem yang dapat meniru kecerdasan manusia. AI mencakup berbagai pendekatan dan metode yang dirancang untuk memungkinkan komputer melakukan tugas-tugas yang biasanya memerlukan kecerdasan manusia, seperti pemecahan masalah, pengambilan keputusan, dan pemahaman bahasa[8]. AI tidak hanya melibatkan penciptaan sistem yang mampu melakukan tugas tertentu seperti bermain catur atau mendiagnosis penyakit, tetapi juga mencakup sistem yang dapat berinteraksi dengan manusia dan dunia nyata secara lebih kompleks.

Pada dasarnya, AI dapat didefinisikan melalui beberapa pendekatan utama, yaitu bertindak seperti manusia (*Turing Test*), berpikir seperti manusia (*cognitive modeling*), berpikir rasional (*laws of thought*), dan bertindak rasional (*rational agent*). Empat teknik dasar dalam AI yang sering diidentifikasi adalah *Reasoning* (Penalaran), *Planning* (Perencanaan), *Learning* (Pembelajaran), dan *Searching* (Pencarian). Keempat teknik ini membentuk landasan dari banyak aplikasi AI modern.

1. *Reasoning* (Penalaran)

Reasoning atau penalaran dalam AI adalah kemampuan untuk membuat keputusan atau menarik kesimpulan berdasarkan informasi yang ada.

- *Rule-Based Reasoning*: Sistem berbasis aturan menggunakan aturan-aturan yang telah didefinisikan sebelumnya untuk menarik kesimpulan. Sistem ini sering digunakan dalam *expert systems*[9].

- *Case-Based Reasoning*: Dalam penalaran berbasis kasus, keputusan dibuat dengan mencari kesamaan dengan kasus-kasus sebelumnya dan menerapkan solusi yang serupa [10].
- Fuzzy Logic: Penalaran fuzzy memungkinkan penanganan ketidakpastian dan pernyataan yang tidak jelas, beroperasi dalam rentang nilai daripada nilai biner ya/tidak [11].

2. Pencarian (*Searching*)

Pencarian merupakan proses dasar dalam AI yang mencakup beberapa metode untuk menemukan solusi terbaik dari suatu permasalahan. Terdapat beberapa jenis pencarian, di antaranya adalah:

- *Blind Search*: Ini adalah metode pencarian yang tidak menggunakan informasi tambahan tentang domain masalah untuk mempersempit ruang pencarian. Metode ini sering digunakan dalam situasi di mana tidak ada informasi lain yang dapat diandalkan untuk menavigasi proses pencarian.
- *Heuristic Search*: Berbeda dengan *blind search*, *heuristic search* memanfaatkan informasi tambahan atau pengetahuan tentang masalah untuk membimbing proses pencarian menuju solusi yang lebih efisien. Heuristik sering digunakan dalam algoritma seperti A* dan *Greedy Best-First Search* [9].
- *Probabilistic Search*: Metode ini melibatkan pendekatan pencarian berdasarkan probabilitas, seperti algoritma Monte Carlo atau pencarian berbasis jaringan Bayesian, yang digunakan untuk menghadapi masalah dengan ketidakpastian tinggi[12].

3. Perencanaan (*Planning*)

Perencanaan dalam AI adalah proses menentukan langkah-langkah atau tindakan untuk mencapai tujuan tertentu. Ini mencakup:

- *Goal Stack Planning*: Teknik ini melibatkan penggunaan tumpukan tujuan yang diatur berdasarkan prioritas untuk mencapai hasil yang diinginkan [13].

- *Constraint Posting*: Dalam metode ini, solusi direncanakan dengan menetapkan dan menyelesaikan batasan atau *constraint* yang ada dalam masalah.

4. Pembelajaran (*Learning*)

Pembelajaran dalam AI adalah proses di mana sistem belajar dari data. Terdapat beberapa pendekatan pembelajaran yang umum digunakan:

- *Supervised Learning*: Model pembelajaran ini melibatkan pelatihan dengan data yang sudah dilabeli, di mana model belajar dari contoh input-output yang diberikan [14].
- *Unsupervised Learning*: Dalam pembelajaran ini, model bekerja dengan data yang tidak dilabeli, mencoba menemukan pola atau struktur tersembunyi dalam data [14].

Reinforcement Learning: Model ini belajar dengan cara trial and error, di mana tindakan model dievaluasi berdasarkan reward yang diterima dari lingkungannya[14].

2.2 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah cabang dari AI yang berfokus pada interaksi antara komputer dan manusia menggunakan bahasa alami

Deep learning telah menjadi pendekatan dominan dalam NLP. Pembelajaran mendalam menggunakan sejumlah besar data mentah yang tidak terstruktur, baik teks maupun suara, untuk melatih model sehingga menjadi semakin akurat dalam memahami dan menghasilkan bahasa alami. Ada beberapa subkategori model dalam pembelajaran mendalam yang penting untuk NLP

1. Model *Sequence-to-Sequence* (seq2seq)

Model seq2seq didasarkan pada jaringan saraf berulang (*Recurrent Neural Networks*, RNNs) dan sering digunakan dalam tugas-tugas seperti penerjemahan mesin. Model ini mengubah urutan teks dari satu bahasa misalnya bahasa Jerman menjadi urutan teks dalam bahasa lain misalnya bahasa Inggris. RNN memungkinkan model untuk menangani data sekuensial dengan mengingat

informasi yang relevan dari urutan sebelumnya, membuatnya ideal untuk menangani teks yang memiliki struktur berurutan. Model seq2seq juga dapat ditingkatkan dengan mekanisme perhatian untuk fokus pada bagian tertentu dari input selama proses penerjemahan.

2. Model Transformer

Model transformer telah merevolusi NLP dengan memperkenalkan mekanisme perhatian diri (self-attention) yang memungkinkan model untuk menangkap hubungan antara semua bagian dari teks, terlepas dari jarak posisi antar token. Model ini menggunakan tokenisasi bahasa, di mana setiap kata atau subkata dipetakan ke dalam token, dan self-attention untuk menghitung ketergantungan antar token dalam teks. Transformer dapat dilatih secara efisien menggunakan pembelajaran mandiri (self-supervised learning) pada dataset teks yang sangat besar. Salah satu tonggak penting dalam pengembangan model transformer adalah representasi encoder dua arah dari Google, yang dikenal sebagai BERT (Bidirectional Encoder Representations from Transformers). BERT telah menjadi fondasi bagi banyak aplikasi NLP, termasuk mesin pencari Google.

2.3 Deep Learning

Deep Learning merupakan bagian dari *machine learning* yang berfokus pada penggunaan jaringan saraf tiruan dengan banyak lapisan atau biasa disebut deep neural networks untuk mempelajari data secara otomatis. Arsitektur Deep Learning dirancang untuk menangkap pola kompleks dalam data melalui berbagai lapisan pemrosesan non-linear. Deep learning bekerja seperti mengajarkan komputer untuk memproses data dengan cara yang terinspirasi otak manusia. Metode ini telah merevolusi banyak bidang, termasuk pengenalan gambar, pemrosesan bahasa alami (NLP), pengenalan suara, dan banyak lagi[14]

Melatih model Deep Learning memerlukan teknik dan algoritma optimasi yang kompleks untuk memastikan bahwa model dapat mempelajari representasi yang bermakna dari data. Beberapa teknik utama meliputi:

Backpropagation: Backpropagation adalah algoritma dasar untuk melatih jaringan saraf tiruan, di mana kesalahan pada output dihitung dan secara iteratif disebarkan kembali ke setiap lapisan untuk memperbarui bobot model. Proses ini memungkinkan jaringan untuk meminimalkan fungsi loss dan meningkatkan akurasi prediksi [14].

Optimizers: Optimizer adalah algoritma yang digunakan untuk memperbarui bobot model selama pelatihan berdasarkan gradien yang dihitung. Beberapa optimizer yang umum digunakan dalam Deep Learning adalah Stochastic Gradient Descent (SGD), Adam, dan RMSprop. Adam, misalnya, populer karena kemampuannya untuk mengadaptasi tingkat pembelajaran (learning rate) selama pelatihan, yang membantu dalam mempercepat konvergensi.

Seiring dengan kemajuan teknologi dan pemahaman yang lebih mendalam tentang jaringan saraf, Deep Learning telah menunjukkan potensi luar biasa dalam berbagai domain. Teknologi ini tidak hanya diterapkan dalam penelitian akademis, tetapi juga dalam aplikasi praktis yang telah mengubah cara berinteraksi dengan dunia digital dan fisik. Berikut adalah beberapa aplikasi utama yang telah diadopsi di berbagai bidang:

Pengenalan Gambar dan Objek: Deep Learning, khususnya CNN, telah merevolusi bidang pengenalan gambar dengan kemampuannya untuk secara otomatis mengekstraksi fitur-fitur penting dari gambar[14].

Dalam NLP, model Deep Learning seperti RNN, LSTM, dan Transformer telah digunakan untuk tugas-tugas seperti penerjemahan bahasa, analisis sentimen, dan pembuatan teks. Model seperti BERT dan GPT telah menunjukkan kinerja yang sangat baik dalam berbagai tugas NLP dengan memanfaatkan representasi kontekstual yang kaya dari teks [6].

2.4 *Text to Speech* (TTS)

Sistem *Text-To-Speech* (TTS) adalah suatu teknologi yang secara otomatis menghasilkan sinyal ucapan dengan mentranskripsikan grafem ke fonem untuk kalimat yang diucapkan. Keunikan TTS dibandingkan dengan mesin pembicara

konvensional terletak pada kemampuannya untuk secara otomatis mengucapkan kata-kata baru. Produksi ucapan oleh mesin, dikenal sebagai TTS, merupakan suatu sistem yang memungkinkan mesin untuk menghasilkan suara manusia secara otomatis. Proses ini dilakukan dengan cara mentranskripsikan atau mengubah tulisan menjadi suara sehingga mesin dapat mengucapkan kalimat yang ditentukan. Dengan demikian, TTS memungkinkan mesin untuk memberikan output suara yang meniru atau mensimulasikan suara manusia [16].

Umumnya, langkah-langkah dalam fungsi sistem TTS mencakup NLP dalam bentuk modul yang mengubah teks menjadi fonem, menghasilkan transkripsi fonetik serta informasi mengenai intonasi dan ritme yang dikenal dengan istilah prosodi. Selain itu, terdapat Digital Signal Processing (DSP) sebagai modul yang mengubah fonem menjadi ucapan, mengonversi informasi fonetis yang diterimanya menjadi sinyal suara[15].

Sistem TTS mensintesis gelombang suara mentah dari teks yang diberikan melalui beberapa komponen. Dengan perkembangan pesat jaringan saraf dalam, pipeline sistem TTS telah disederhanakan menjadi pemodelan generatif dua tahap, selain dari pra-pemrosesan teks seperti normalisasi teks dan fonemisasi. Pada tahap pertama, sistem TTS menghasilkan representasi ucapan menengah dari teks yang telah diproses sebelumnya. Representasi ini berupa Mel-Spectogram[16]dan Fitur linguistik[17] yang telah di proses sebelumnya. Tahap kedua menghasilkan gelombang suara mentah yang dikondisikan berdasarkan representasi Intermediate tersebut[17][18].

2.5 Voice Clone

Voice cloning adalah proses teknologi yang digunakan untuk menciptakan salinan suara manusia dengan menggunakan algoritma dan model pembelajaran mesin. Tujuannya adalah untuk mereplikasi suara seseorang sehingga suara yang dihasilkan oleh sistem tersebut sangat mirip dengan suara aslinya. Teknologi ini memanfaatkan kecerdasan buatan, khususnya dalam bidang pengolahan sinyal suara dan *deep learning*.

Voice cloning adalah tugas pembelajaran untuk mensintesis suara dari pembicara yang tidak dikenal berdasarkan hanya beberapa sampel. Meskipun metode *voice cloning* saat ini mencapai hasil yang menjanjikan dalam sintesis TTS untuk suara baru, tetapi pendekatan-pendekatan tersebut masih kurang dalam kemampuan untuk mengontrol ekspresivitas audio yang disintesis.

2.6 *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech* (VITS)

Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS) adalah sebuah model TTS yang menggunakan pendekatan *end-to-end* untuk menghasilkan suara yang lebih alami dibandingkan model TTS dua tahap yang ada[6]. Model ini mengadopsi inferensi variansional yang ditingkatkan dengan *normalizing flows* dan proses pelatihan adversarial, yang meningkatkan daya ekspresif dari pemodelan generatif. Model VITS mengatasi masalah dalam model TTS dua tahap yang memerlukan pelatihan atau penyetelan berurutan untuk menghasilkan kualitas yang tinggi. VITS menggunakan pendekatan *end-to-end* dengan *Conditional Variational Autoencoder* (CVAE) untuk menghubungkan dua modul sistem TTS melalui variabel laten, memungkinkan pembelajaran yang efisien dan menghasilkan suara berkualitas tinggi[6].

2.6.1 *Conditional Variational Autoencoder* (CVAE)

CVAE adalah varian dari VAE yang memodelkan distribusi probabilistik dengan kondisi tertentu. Dalam model ini, distribusi laten z bersyarat pada input tertentu (disebut sebagai kondisi), yang dalam kasus TTS ini adalah teks yang diinputkan. CVAE bertujuan untuk memaksimalkan *variational lower bound* atau *evidence* lower bound (ELBO) dari *log-likelihood* data [7].

$$Log\ p\theta(x|c)\ \geq \mathrm{Eq}\phi(\mathrm{z}|\mathrm{x})\left[log\ p\theta(x|z) - log\ \frac{q\phi(\mathrm{z}|\mathrm{x})}{p\theta(z|c)}\right] \quad (1)$$

Keterangan:

log $p_\theta(x \mid c)$ adalah kemungkinan model menghasilkan data audio $x$berdasarkan kondisi teks c.

$\mathbb{E}_{q_\phi(z|x)}[\cdot]$ adalah rata-rata nilai yang dihitung dari sampel variabel laten $z$yang diambil dari encoder.

z adalah variabel laten yang menyimpan representasi tersembunyi dari data audio.

$q_\phi(z \mid x)$adalah distribusi posterior, yaitu distribusi yang dipelajari oleh encoder dari data audio.

$p_\theta(x \mid z)$adalah probabilitas rekonstruksi, yang menunjukkan seberapa baik decoder dapat menghasilkan kembali audio $x$dari variabel laten z.

$\log \frac{q_\phi(z|x)}{p_\theta(z|c)}$ merupakan komponen KL divergence, yang mengukur perbedaan antara distribusi laten dari audio dan distribusi laten yang diprediksi dari teks.

$p_\theta(z \mid c)$adalah distribusi prior yang dikondisikan pada teks.

$$\text{Lrecon} = ||\text{xmel} - \hat{}\text{xmel}||\ 1 \qquad (2)$$

Keterangan :

$\mathcal{L}_{\text{recon}}$adalah *loss* rekonstruksi, yang digunakan untuk mengukur perbedaan antara sinyal akustik asli dan hasil rekonstruksi model.

x_{mel}adalah *mel-spectrogram* asli yang diekstraksi dari sinyal audio masukan.

$\hat{x}_{mel}$adalah *mel-spectrogram* hasil rekonstruksi yang dihasilkan oleh *decoder* atau generator model.

$\|\cdot\|_1$menyatakan norma L1, yaitu jumlah nilai absolut dari selisih antara dua mel-spectrogram.

Loss ini mendorong model untuk menghasilkan keluaran yang mendekati mel-spectrogram asli secara detail dan stabil.

CVAE menggunakan *reconstruction loss* yang dihitung di domain *mel-spectrogram* untuk memastikan bahwa prediksi spektrum audio mendekati data target. L1 loss

antara *spektrum mel* yang dihasilkan dan spektrum *mel* target digunakan untuk meningkatkan kualitas ucapan.

$$LKL = Ez \sim q\phi(z \mid xlin)[logq\phi(z \mid xlin) - logp\theta(z \mid ctext, A)] \quad (3)$$

$\mathcal{L}_{\mathrm{KL}}$adalah nilai loss KL divergence yang mengukur perbedaan antara distribusi posterior dan distribusi prior

z adalah variabel laten yang merepresentasikan informasi akustik.

$q_{\phi}(z \mid x_{lin})$ adalah distribusi posterior yang dihasilkan oleh encoder dari fitur akustik x_{lin}.

x_{lin}adalah fitur akustik masukan, berupa *mel-spectrogram* atau linear spectrogram.

Φ menyatakan parameter *encoder posterior*.

$p_{\theta}(z \mid c_{text}, A)$adalah distribusi *prior* yang dikondisikan pada representasi teks dan *alignment*.

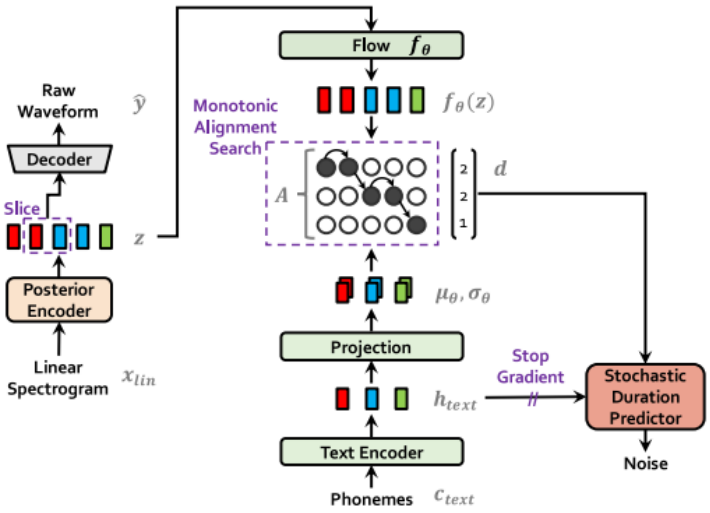
Ctext adalah representasi *embedding* teks hasil *encoder* teks.

A adalah matriks *alignment monotonic* antara token teks dan frame akustik.

θ menyatakan parameter *prior encoder* dan *normalizing flow*.

$\mathbb{E}_{z\sim q_{\phi}(z|x_{lin})}[\cdot]$ menyatakan ekspektasi terhadap sampel variabel laten $z$yang diambil dari distribusi posterior.

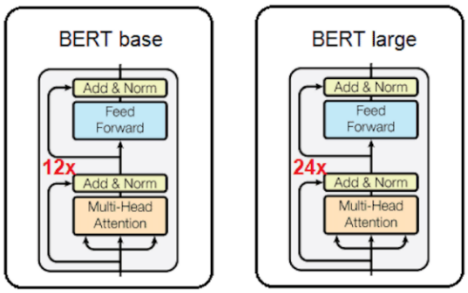
Selain itu, CVAE menghitung KL *divergence* antara distribusi *posterior* dan *prior* untuk memastikan bahwa distribusi yang dipelajari tidak menyimpang terlalu jauh dari distribusi laten yang diinginkan.



Gambar 2. 1 Training Procedure [5]

2.7 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) adalah model pemrosesan *Natural Language Processing* (NLP) yang dikembangkan oleh Google AI. BERT adalah model berbasis transformer yang dilatih dengan cara bidirectional, yang dapat memahami kontekst kata dalam kalimat dari dua arah[6]. Berbeda dengan model sebelumnya yang hanya membaca teks dalam satu arah, sehingga memberikan pemahaman yang lebih kaya dan akurat tentang konteks.



Gambar 2. 2 Arsitektur BERT [6]

Arsitektur BERT terdiri dari beberapa lapisan (layers) yang berfungsi untuk memproses dan memahami teks secara mendalam:

1. Embedding Layer: Setiap kata dalam input teks diubah menjadi representasi vektor dengan dimensi tertentu melalui embedding. Ini termasuk embedding untuk kata itu sendiri (word embedding), posisi kata dalam kalimat (positional embedding), dan segmen (segment embedding). Embedding ini kemudian dijumlahkan untuk membentuk representasi akhir dari setiap kata dalam input.
2. Encoder Layers: BERT terdiri dari beberapa lapisan encoder yang identik, yang masing-masing terdiri dari dua sub-layer utama:
3. Self-Attention Layer: Meskipun tidak menjelaskan Transformer, penting untuk diketahui bahwa lapisan ini memungkinkan BERT memahami hubungan antar kata dalam kalimat secara bersamaan, baik kata sebelum maupun setelah kata target.
4. Feedforward Neural Network: Setiap self-attention layer diikuti oleh feedforward neural network (FFNN) yang bersifat fully connected, yang memproses output dari self-attention layer untuk meningkatkan kemampuan representasi teks.

Dalam proses pelatihannya, BERT menggunakan teknik *Masked Language Model* (MLM) di mana beberapa kata dalam input sengaja disembunyikan (*masked*) dan model dilatih untuk memprediksi kata-kata yang hilang tersebut[6]. Teknik ini membantu BERT memahami hubungan antara kata-kata dalam konteks yang lebih luas. Selain itu, BERT juga menggunakan teknik *Next Sentence Prediction* (NSP) yang melatih model untuk memprediksi apakah dua kalimat memiliki keterikatan satu sama lain dan membantu dalam berbagai tugas seperti tanya jawab dan inferensi teks [6].

2.7.1 IndoBERT

IndoBERT adalah adaptasi model BERT yang secara khusus dilatih menggunakan korpus berbahasa Indonesia. Pengembangan IndoBERT bertujuan untuk mengatasi keterbatasan model multibahasa yang sering kali kurang optimal dalam memahami karakteristik linguistik bahasa Indonesia, seperti imbuhan, reduplikasi, dan struktur kalimat khas. IndoBERT dilatih menggunakan data teks berbahasa Indonesia dalam jumlah besar yang bersumber dari berbagai domain, termasuk berita, artikel daring,

dan sumber teks lainnya. Dengan pelatihan ini, IndoBERT mampu menghasilkan representasi kata dan kalimat yang lebih sesuai dengan konteks bahasa Indonesia.

2.8 *Grapheme-to-Phoneme* (G2P)

Grapheme-to-Phoneme (G2P) adalah proses konversi dari simbol grafem menjadi fonem. G2P berfungsi untuk menghasilkan transkripsi fonetik dari teks tertulis, yang sangat penting dalam berbagai aplikasi seperti sintesis suara, pengenalan suara, dan pembelajaran bahasa. Fonem merupakan unit terkecil dalam suatu bahasa yang dapat membedakan arti[3]. Dalam Bahasa Indonesia, fonem diidentifikasi melalui analisis bunyi-bunyi bahasa yang bisa mempengaruhi arti kata. Misalnya, perbedaan bunyi antara /e/ dan /ə/ dalam kata "bebek" dan "bebeg" dapat mengubah makna kata tersebut sepenuhnya. Oleh karena itu, setiap fonem memiliki peran penting dalam struktur kata dan kalimat, serta dalam komunikasi sehari-hari

International Phonetic Alphabet (IPA) adalah sistem alfabet yang dikembangkan untuk menyediakan cara yang konsisten dan tidak ambigu dalam merepresentasikan bunyi bahasa dari berbagai bahasa di dunia[19]. Tujuan utamanya adalah menciptakan satu set simbol yang dapat digunakan untuk menuliskan bunyi-bunyi dari berbagai bahasa di dunia dengan cara yang seragam dan dapat dipahami secara internasional. Setiap simbol dalam IPA mewakili satu bunyi tertentu, sehingga menghilangkan ambiguitas yang sering terjadi dalam penulisan dengan alfabet standar. Simbol-simbol dalam IPA dibagi menjadi beberapa kategori seperti konsonan, vokal, diakritik, dan simbol suprasegmental. IPA digunakan dalam berbagai bidang seperti linguistik, fonetik, pendidikan bahasa, dan logopedi untuk membantu memahami dan menganalisis pengucapan kata secara tepat.

IPA terdiri dari beberapa kategori simbol yang digunakan untuk merepresentasikan berbagai jenis bunyi:

1. Konsonan: Simbol untuk bunyi yang dihasilkan dengan menyempitkan atau menutup saluran vokal pada titik tertentu. Misalnya, /p/ untuk bunyi (p) seperti pada kata "pat", dan /t/ untuk bunyi (t) seperti pada kata "tap".

2. Vokal: Simbol untuk bunyi yang dihasilkan dengan aliran udara bebas melalui saluran vokal tanpa hambatan signifikan. Misalnya, /i/ untuk bunyi (i) seperti pada kata "bit", dan /a/ untuk bunyi (a) seperti pada kata "sat".

Tabel 2. 1 IPA Konsonan

IPA	Alfabet
/b/	b
/tʃ/	c
/d/	d
/f/	f
/g/	g
/h/	h
/dʒ/	j
/k/	k
/l/	l
/m/	m
/n/	n
/p/	p
/r/	r
/s/	s
/t/	t

/v/	v
/w/	w
/x/	x
/j/	y
/z/	z
/ŋ/	ng
/ɲ/	ny
/ʃ/	sy

Tabel 2. 2 IPA Vokal

IPA	Alfabet
/a/	a
/e/	e
/ə/	e
/i/	i
/o/	o
/u/	u
/ɛ/	e

Tabel 2.3 IPA Diftong

IPA	Alfabet

/ai/	ai
/au/	au
/oi/	oi

2.9 Evaluasi Model

2.9.1 *Phoneme Error Rate* (PER)

Phoneme Error Rate (PER) adalah salah satu metrik evaluasi yang digunakan untuk mengukur akurasi sistem pengenalan suara dan konversi G2P. PER mirip dengan WER, tetapi fokusnya adalah pada tingkat kesalahan di tingkat fonem, bukan di tingkat kata

$$PER = \frac{S + D + I}{N} \quad (4)$$

Dimana:

S adalah jumlah substitusi, fonem salah yang dikenali,

D adalah jumlah penghapusan, fonem yang hilang dari transkripsi,

I adalah jumlah penyisipan, fonem yang tidak ada di referensi tetapi muncul di hasil transkripsi,

N adalah jumlah total fonem dalam transkripsi referensi

2.9.2 Mean Opinion Score (MOS)

$$MOS = \frac{1}{n} \sum_{i=1}^{N} x1 \quad 5$$

Keterangan:

- N= jumlah responden
- x_i= skor yang diberikan responden ke-i

Mean opinion score (MOS) adalah metode evaluasi subjektif yang digunakan untuk mengukur persepsi manuasia terhadap audio atau ucapan maupun layanan multimedia lainnya. Yang mana manusia selaku penilai memberikan skor atau nilai dari skala 1 sampai 5, semakin tinggi nilai yang didapaat maka semakin tinggi juga kualitas sinyal audio tersebut, lalu hasil akhirnya dirata ratakan menjadi nilai MOS [20].MOS telah lama dijadikan standart karena secara langsung merepresentasikan pengalaman pendengar, sehingga sering digunakan untuk membandingkan kualitas *codec*, *speech enhancement*, maupun system TTS

Tabel 2. 4 Skala Penilaian MOS

Skor	**Kualitas Persepsi**
5	Sangat baik
4	Baik
3	Cukup
2	Buruk
1	Sangat buruk

2.9.3 Classification Report

Classification report merupakan metode evaluasi yang digunakan dalam machine learning berbasis klasifikasi karena mampu memberikan gambaran kinerja hasil model secara lebih rinci dibandingkan hanya akurasi. Metrik nya seperti precision, recall, F1-Score, dan support untuk setiap kelas.

	precision	recall	f1-score	support
Kelas 0	0.85	0.80	0.82	100
Kelas 1	0.78	0.83	0.80	90
accuracy			0.81	190
macro avg	0.82	0.82	0.81	190
weighted avg	0.82	0.81	0.81	190

Gambar 2. 3 Classification Report

1. *Precision* adalah ketepatan model dalam memprediksi suatu kelas. Semakin tinggi *precision*, semakin kecil kemungkinan model memberikan prediksi salah pada kelas itu.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Dimana:

- TP (True Positive): data yang sebenarnya bernilai positif dan diprediksi positif oleh model
- FP (False Positive): data yang seharusnya bernilai negatif namun diprediksi sebagai positif

2. *Recall* adalah kemampuan model dalam menemukan seluruh data yang benar dari suatu kelas. Nilai *recall* yang tinggi menandakan model jarang melewatkan data penting pada kelas tersebut.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

Dimana:

- TP (True Positive): data yang sebenarnya bernilai positif dan diprediksi positif oleh model

- False Negatif: Data yang sebenarnya bernilai positif malah diprediksi sebagai negatif.

3. F1-*score* merupakan ukuran gabungan yang menyeimbangkan precision dan *recall*. Metrik diperlukan untuk evaluasi yang lebih stabil, terutama pada dataset yang tidak seimbang, karena F1-*score* tidak hanya melihat ketepatan prediksi tetapi juga kemampuan model menangkap data kelas tersebut secara menyeluruh.

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \tag{8}$$

4. *Accuracy* adalah persentase prediksi benar secara keseluruhan. Metrik ini menunjukkan seberapa baik model mampu mengklasifikasikan data dengan benar.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

2.10 Perangkat Lunak (*Software*)

2.10.1 Python

Python adalah bahasa pemrograman tingkat tinggi yang bersifat serba guna, mudah dipahami, dan populer. Dikembangkan oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991, Python telah menjadi salah satu bahasa pemrograman yang paling banyak digunakan di dunia.

Beberapa fitur utama Python termasuk sintaksis yang bersih dan mudah dibaca, dukungan untuk paradigma pemrograman yang beragam seperti pemrograman berorientasi objek, fungsional, dan pemrograman terstruktur. Python juga dilengkapi dengan berbagai pustaka dan modul standar yang kaya, serta memiliki komunitas pengguna yang besar yang aktif dalam pengembangan dan pemeliharaan.

Karena fleksibilitasnya, Python digunakan dalam berbagai bidang termasuk pengembangan web, analisis data, Artificial intelligence, Machine Learning pengembangan perangkat lunak, pemrograman sistem, dan masih banyak lagi. Keunggulan Python dalam hal keterbacaan dan produktivitas membuatnya menjadi pilihan populer bagi pengembang pemula maupun yang berpengalaman.

2.10.2 Google Colaboratory

Google Colab adalah platform *cloud-based* yang memungkinkan pengguna untuk menulis dan mengeksekusi kode Python dalam lingkungan *Jupyter Notebook* yang terpusat di *cloud*. Fitur utama yang membuat *Google Colab* populer adalah akses gratisnya ke *Graphics Processing Unit* (GPU) dan *Tensor Processing Unit*. (TPU), yang memungkinkan pengguna untuk melakukan pelatihan model pembelajaran mesin dengan cepat dan efisien.

GPU (*Graphics Processing Unit*) adalah jenis prosesor khusus yang dirancang untuk melakukan operasi terkait grafis secara efisien. GPU awalnya digunakan untuk mengolah tampilan grafis pada perangkat komputer, tetapi kemudian dikembangkan untuk melakukan komputasi umum, termasuk dalam bidang ilmu data, kecerdasan buatan, dan pengolahan gambar. GPU memiliki banyak inti pemrosesan yang bekerja secara paralel, sehingga sangat efektif untuk tugas-tugas yang memerlukan komputasi besar-besaran.

TPU (Tensor Processing Unit) adalah jenis prosesor khusus yang dikembangkan oleh Google untuk mempercepat operasi pada jaringan saraf tiruan (neural networks). TPU didesain khusus untuk mengoptimalkan tugas-tugas yang memerlukan pengolahan tensor, seperti yang banyak terjadi dalam pembelajaran mesin dan kecerdasan buatan. TPU memiliki struktur yang sangat efisien untuk melakukan operasi matriks dan vektor, yang sering digunakan dalam pembelajaran mesin, sehingga memberikan peningkatan kinerja yang signifikan dalam proses pelatihan model. Notebook dapat dijalankan dalam waktu hingga 3 jam 20 menit dan runtime berhenti apabila tidak terdapat aktivitas pada notebook dalam jangka waktu tertentu.

2.10.3 Pytorch

PyTorch adalah sebuah kerangka kerja pemrograman yang dikembangkan oleh Facebook pada tahun 2016, dan sering digunakan untuk membangun model kecerdasan buatan (AI) seperti jaringan saraf tiruan. Keunggulan utama dari PyTorch adalah kemampuannya yang fleksibel dan mudah digunakan, terutama untuk peneliti atau pengembang yang ingin mencoba berbagai eksperimen. Salah satu fitur andalan PyTorch adalah dynamic computational graph, yang memungkinkan pengguna untuk membuat dan mengubah struktur komputasi secara dinamis saat program berjalan. Hal ini membuat PyTorch lebih intuitif dan mudah dipahami dibandingkan beberapa kerangka kerja lain, terutama ketika pengguna perlu melakukan debugging atau mengembangkan model dengan alur komputasi yang kompleks. Selain itu, PyTorch juga memiliki dukungan untuk komputasi berbasis tensor, yang memungkinkan pengolahan data dalam bentuk array berdimensi tinggi, serta mudah diintegrasikan dengan GPU untuk mempercepat proses pelatihan model. Secara keseluruhan, PyTorch sangat populer di kalangan peneliti AI karena kemudahan penggunaannya, fleksibilitas, dan performa tinggi yang ditawarkan

2.10.4 Coqui TTS

Coqui TTS merupakan sebuah toolkit open-source untuk sintesis suara (text-to-speech) yang dikembangkan sebagai kelanjutan dari proyek TTS Mozilla dan ditujukan untuk mendukung kegiatan penelitian serta pengembangan aplikasi di bidang pemrosesan bahasa alami dan kecerdasan buatan. Toolkit ini menyediakan berbagai arsitektur model modern, seperti VITS, Tacotron, Tacotron2, Glow-TTS, serta model generatif berbasis multilingual voice cloning seperti XTTS, yang mampu menghasilkan keluaran suara dengan tingkat naturalitas yang tinggi. Coqui TTS dilengkapi dengan pipeline yang komprehensif, meliputi tahap persiapan dataset, proses pelatihan dan fine-tuning model, hingga tahap inferensi audio, sehingga memungkinkan penyesuaian sistem sintesis suara terhadap bahasa dan karakteristik penutur tertentu. Keunggulan Coqui TTS terletak pada fleksibilitas,

ketersediaan model pra-latih, serta dukungannya terhadap pengembangan sistem TTS untuk bahasa dengan sumber daya terbatas. Namun demikian, kualitas suara yang dihasilkan sangat dipengaruhi oleh kualitas dan kuantitas data pelatihan yang digunakan, serta pemanfaatan fitur voice cloning perlu memperhatikan aspek etika, khususnya terkait privasi dan persetujuan penggunaan suara dalam penerapan di lingkungan nyata.

2.11 Cara Membuat AI

AI merupakan proses langkah demi langkah untuk mengembangkan dan menerapkan proyek AI untuk memecahkan masalah. Memberikan kerangka kerja yang tepa untuk membantu mencapai tujuan dalam proyek AI. Adapun langkah-langkah untuk membuat AI berdasarkan referensi dari Yessi Mulyani (2023):

Gambar 2. 4 Tahapan Cara Membuat AI

1. Menentukan Tujuan dan Identifikasi Masalah

Langkah pertama adalah memahami tujuan yang ingin dicapai dan mengidentifikasi masalah yang ingin diselesaikan. Masalah ini bisa berupa persoalan bisnis, teknis, atau sosial yang dapat diatasi menggunakan AI.

2. Memilih dan Menentukan Jenis AI yang Digunakan

Setelah masalah ditentukan, langkah berikutnya adalah memilih jenis AI yang paling sesuai. Apakah itu machine learning, deep learning, natural language processing (NLP), atau jenis AI lainnya.

3. Mempelajari AI yang Sudah Ditentukan

Langkah ini melibatkan studi mendalam tentang model atau algoritma yang relevan dengan jenis AI yang dipilih. Penting untuk memahami cara kerja model tersebut agar bisa diterapkan dengan benar pada kasus yang dihadapi.

4. Pemrograman atau Pembuatan Software

Setelah memilih model, langkah selanjutnya adalah melakukan pemrograman dan membangun software. Proses ini melibatkan coding dan implementasi algoritma yang telah dipelajari ke dalam sistem.

5. Penyempurnaan dan Optimisasi

Setelah software selesai, perlu dilakukan penyempurnaan dan optimisasi. Ini bisa berupa pengujian, debugging, serta peningkatan performa model AI agar lebih efisien dan akurat dalam menyelesaikan tugasnya.

6. Implementasi Software ke dalam Komputer atau Mesin

Langkah terakhir adalah mengintegrasikan software AI yang telah sempurna ke dalam sistem komputer atau mesin. Dengan ini, komputer atau mesin dapat berfungsi secara cerdas sesuai dengan tujuan awal yang telah ditentukan.

2.12 Penelitian Terdahulu

Kim, Kong, dan Son (2021) melakukan penelitian mengenai TTS end-to-end dengan mengusulkan metode baru bernama Variational Inference with Adversarial Learning for End-to-End Text-to-Speech. Penelitian ini bertujuan untuk menghasilkan model TTS single-stage yang mampu menghasilkan suara sintesis yang lebih natural dan berkualitas tinggi dibandingkan sistem TTS dua tahap tradisional dengan pelatihan terpisah. Data eksperimen menggunakan dua dataset suara yaitu LJ Speech yang berisi 13.100 klip audio rekaman satu pembicara dan VCTK yang terdiri dari sekitar 44.000 klip dengan banyak pembicara. Metode yang digunakan mengombinasikan CVAE dengan adversarial training, serta menerapkan normalizing flows dan stochastic duration predictor untuk menangkap variasi ritme

dan ekspresi suara dari teks input. Evaluasi menggunakan MOS dari manusia menunjukkan bahwa metode ini memberikan kualitas suara yang lebih natural dan efisien dibandingkan dengan model TTS terbaik yang tersedia secara publik, sekaligus mendekati kualitas rekaman asli.

Jia, Zen, Shen, Zhang, dan Wu melakukan penelitian untuk meningkatkan kualitas TTS dengan memperkenalkan PnG BERT, sebuah model augmented BERT yang dirancang sebagai *encoder* dalam arsitektur TTS. Model ini memanfaatkan representasi fonem dan grafem secara bersamaan serta menyertakan penjajaran pada tingkat kata untuk menangkap hubungan antara tulisan dan pengucapan. PnG BERT dilatih secara *self-supervised* pada korpus teks besar dan kemudian di-*fine-tune* dalam tugas TTS. Dalam eksperimen, model ini digunakan sebagai pengganti *encoder* pada model TTS berbasis durasi, dan hasil evaluasi subjektif menunjukkan bahwa TTS dengan PnG BERT menghasilkan prosodi dan pengucapan yang lebih natural dan akurat dibandingkan baseline, bahkan tidak berbeda signifikan dari rekaman suara manusia profesional dalam preferensi pendengar.

Zhang, Song, Tan, dkk. melakukan penelitian untuk meningkatkan performa TTS dengan memperkenalkan *Mixed-Phoneme* BERT, sebuah varian dari model BERT yang memanfaatkan representasi gabungan *phoneme* dan *sup-phoneme* sebagai input. Dalam studi ini, representasi *sup-phoneme* diperoleh dengan menerapkan *Byte-Pair Encoding* pada urutan *phoneme* sehingga model dapat menangkap konteks dan informasi semantik yang lebih kaya dibandingkan hanya menggunakan *phoneme* tunggal saja. Model *Mixed-Phoneme* BERT pertama kali dilatih pada korpus teks besar yang terdiri dari jutaan kalimat dari berita dan *Wikipedia*, kemudian hasilnya di-*fine-tune* pada tugas *TTS* menggunakan dataset *LJSpeech* yang berisi sekitar 24 jam audio dan teks transkripsi. Untuk mengevaluasi performa, peneliti membandingkan kualitas suara yang dihasilkan oleh beberapa konfigurasi sistem *TTS* termasuk *FastSpeech 2* dasar, *FastSpeech 2* dengan *Mixed-Phoneme BERT* yang tidak *pre-trained*, dan *FastSpeech 2* dengan *Mixed-Phoneme* BERT yang *pre-trained* melalui uji dengar subjektif seperti *MOS* dan *Comparison MOS* (*CMOS*). Hasil menunjukkan bahwa penggunaan *Mixed-Phoneme* BERT yang telah

pre-trained secara signifikan meningkatkan kualitas suara yang dihasilkan *CMOS* +0,30 dibanding *FastSpeech 2*, dengan *prosody* yang lebih ekspresif dan *natural*, serta mempercepat *inference* sekitar tiga kali dibandingkan model *PnG BERT* sebelumnya.

Li, Han, Jiang, dan Mesgarani melakukan penelitian untuk meningkatkan kualitas *TTS* dengan memperkenalkan *Phoneme-Level* BERT, sebuah model BERT yang dilatih khusus pada level *phoneme* dengan tugas tambahan memprediksi *grapheme* yang bersesuaian bersama dengan prediksi *phoneme* yang di-*mask*. Tujuan penelitian ini adalah untuk mengatasi keterbatasan model *pre-training* berbasis kata atau *sup-phoneme* yang kurang efisien ketika hanya *phoneme* yang dibutuhkan dalam TTS, serta untuk menghadirkan representasi linguistik yang lebih relevan dengan tugas TTS. *PL*-BERT dilatih secara *self-supervised* untuk mengenali hubungan *phoneme* dan *grapheme*, kemudian digunakan sebagai *encoder* dalam model TTS — pada penelitian ini dibandingkan dengan model *Style*TTS sebagai *baseline*. Evaluasi dilakukan melalui pengujian subjektif menggunakan MOS pada data uji, di mana model dengan PL-BERT menunjukkan peningkatan MOS yang signifikan dibandingkan dengan *baseline Style*TTS tanpa *PL*-BERT, menunjukkan bahwa PL-BERT mampu menghasilkan suara sintetis dengan *prosody* dan *naturalness* yang lebih baik.

Nguyen, Pham, dan Nguyen menggunakan *XPhone*BERT, sebuah model *pre-trained multilingual* yang dirancang khusus untuk mempelajari representasi *phoneme* dalam tugas TTS. Model ini menggunakan arsitektur yang sama dengan *BERT-base* dan dilatih dengan pendekatan RoBERTa *pre-training* pada korpus berukuran besar yang terdiri dari sekitar 330 juta kalimat pada level *phoneme* dari hampir 100 bahasa dan *locale*. *XPhone*BERT kemudian digunakan sebagai *encoder phoneme* dalam sistem TTS *end-to-end* untuk memperkuat representasi fonetik sebelum proses sintesis suara. Evaluasi dilakukan dengan membandingkan model TTS berbasis VITS yang memakai *XPhone*BERT dan yang tidak memakai *XPhone*BERT pada dua *setting* data: seluruh data pelatihan dan data terbatas (misalnya 5% data). Hasil evaluasi menunjukkan bahwa TTS dengan *XPhone*BERT

mampu meningkatkan kualitas *naturalness* suara menurut *MOS* serta memperbaiki metrik objektif seperti *Mel-Cepstrum Distance* dan *Root Mean Square Error of F0* dibandingkan *baseline* tanpa *pre-training*, termasuk pada kondisi data pelatihan yang sangat terbatas, menunjukkan bahwa *XPhoneBERT* efektif dalam menghasilkan suara sintetis yang lebih *natural* dan berkualitas tinggi lintas bahasa.

Tabel 2.5 Penelitian Terdahulu

No	Peneliti	Data	Algoritma	Hasil
1	Kim, Kong, & Son[7]	LJ Speech 13.100 klip audio, VCTK klip multi speaker	Conditional Variational Autoencoder with Adversarial Learning	Hasilnya MOS: Data LJ Speech: 4.43 ± 0.05 VITS 4.46 ± 0.05 *ground truth*; VCTK: ~4.38 VITS ~3.8 baseline
2	Linh The Nguyen, Thinh Pham, dan Dat Quoc Nguyen[21]	Data *Pre-training*: 330 juta kalimat berbasis fonem berasal dari Wiki40B dan Wikipedia Data TTS: LJ*Speech* dan bahasa Vietnam 18 jam audio dengan 12.300 klip	XPhoneBERT	Bahasa Inggris data penuh: MOS meningkat dari 4.00 menjadi 4.14 Bahasa Vietnam data penuh: MOS meningkat dari 3.74 menjadi 3.89
3	Zhang, Song, Tan, dkk[22]	Data *Pre-trained* 83 juta kalimat teks tidak berlabel, berasal dari WMT News Corpus dan Wikipedia Data fine-tuning TTS: LJSpeech	Mixed-Phoneme BERT	Mean Opinion Score (MOS): 3,75 menjadi 4,04
4	Li, Han, Jiang & Mesgarani [23]	Data *Pre-Trained* 74 juta kalimat berasal dari	*Phoneme-Level* BERT	StyleTTS dengan PL-BERT nilai MOS 4,24

No	Peneliti	Data	Algoritma	Hasil
		Wikipedia bahasa Inggris Data TTS: LJ*Speech*		StyleTTS tanpa BERT nilai MOS 4,19
5	Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, Yonghui Wu[24]	Data *Pre-trained:* 131 juta kalimat berasal dari korpus wikipedia Data TTS: Dataset *proprietary* google 243 jam rekaman suara	PnG BERT	MOS meningkat dari 4,41 menjadi 4,47

III. METODE PENELITIAN

3.1 Waktu dan Tempat

Penelitian dan pembuatan skripsi ini adalah dilakukan pada:

Waktu : April 2025 – Desember 2025.

Tempat : Universitas Lampung.

3.2 Jadwal Penelitian

Berikut adalah susunan jadwal yang telah direncanakan untuk penelitian ini

Tabel 3.1 Jadwal penelitian

Aktivitas	2025								
	Apr	Mei	Jun	Jul	Agust	Sept	Okt	Nov	Des
Studi Literatur									
Pengumpulan Data									
Pengolahan Data									
Perancangan Model									
Pelatihan Model									
Evaluasi Hasil									
Perbandingan Hasil									
Pengerjaan Laporan									

3.3 Alat dan Bahan Penelitian

3.3.1 Alat

alat yang digunakan dalam penelitian ini dengan spesifikasi berikut

Tabel 3.2 Alat

No.	Alat	Spesifikasi	Kegunaan
1	Laptop	Processor Intel(R) Core(TM) i9-13900H CPU RAM 16 GB GPU GTX 4060 8 GB OS Windows 11	Untuk Pengembangan dan pelatihan model.
2	Python	Versi 3.11	Implementasi program dan pengolahan data.
3	Google Colab	GPU Tesla T4	Pelatihan model berbasis GPU.
4	Notepad		Pengeditan file teks sederhana untuk dataset
5	Visual Studio Code		Penulisan dan pengelolaan kode program.
6	Adobe Enhance Audio		Pembersihan dan peningkatan kualitas audio.

3.3.2 Bahan

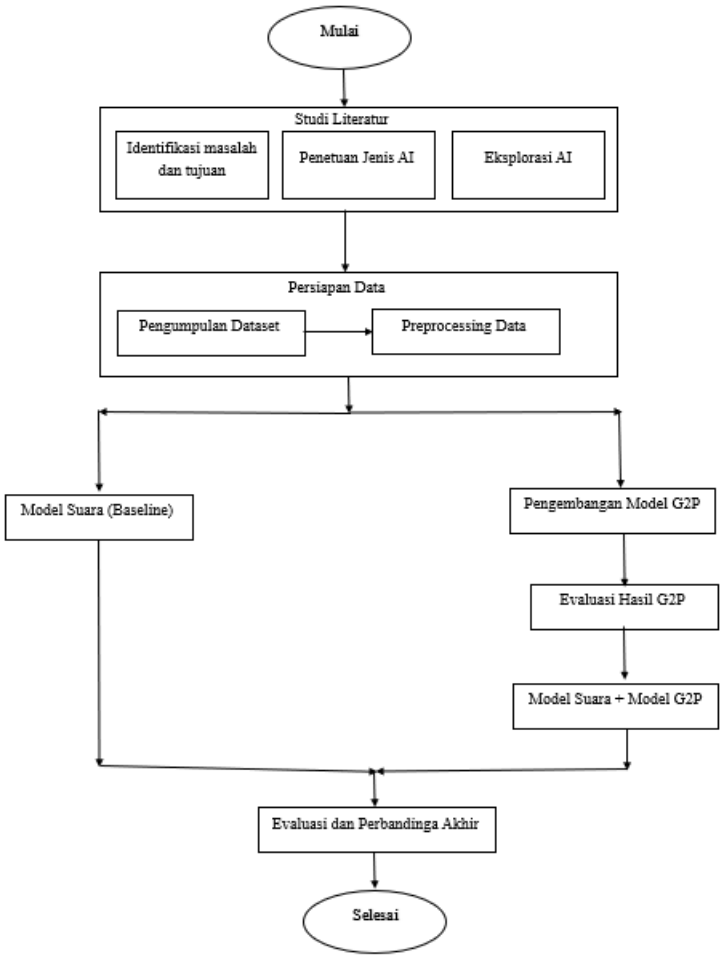
Dalam penelitian ini bahan penelitian yang digunakan yaitu :

1. Data audio yang direkam dari suara peneliti sendiri, agar data sesuai yang diinginkan
2. Kamus yang berisi fonem-fonem Bahasa Indonesia beserta transkripsinya. Peneliti mendapatkan *lexicon* fonem pada situs Github https://github.com/open-

dict-data/ipa-dict/blob/master/data/ma.txt, yang ditambahkan kata lain sesuai kebutuhan

3.4 Tahapan Penelitian

Tahapan Penelitian ini yang digunakan yaitu sebagai berikut :



Gambar 3.1 Tahapan penelitian

3.4.1 Studi Literatur

Tahap studi literatur dilakukan untuk memperoleh dasar teori, metode, serta pendekatan yang relevan dalam pengembangan sistem TTS berbasis model G2P. Studi ini bertujuan untuk memahami konsep dasar kecerdasan buatan, meninjau berbagai penelitian sebelumnya, dan menentukan arah penelitian yang tepat. Berdasarkan diagram alur penelitian Gambar 3.1, studi literatur mencakup tiga

langkah utama, yaitu identifikasi masalah dan tujuan, penentuan jenis AI yang digunakan, dan eksplorasi teknologi AI yang sesuai.

3.4.1.1 Identifikasi Masalah dan Tujuan

Langkah pertama dalam studi literatur adalah mengidentifikasi permasalahan utama yang diselesaikan. Dalam konteks penelitian ini, permasalahan yang diangkat adalah bagaimana meningkatkan kualitas pelafalan pada sistem TTSbahasa Indonesia melalui integrasi model G2P.

Studi literatur dilakukan dengan meninjau:

- Penelitian terdahulu mengenai sistem TTS berbasis deep learning seperti Tacotron, FastSpeech, dan VITS, yang menunjukkan peningkatan signifikan dalam naturalitas suara namun masih memiliki kelemahan dalam akurasi pelafalan.
- Kajian tentang model G2P sebagai solusi untuk mengatasi kesalahan fonetik pada model TTS berbasis teks langsung.
- Studi bahasa dan fonetik Indonesia untuk memahami struktur fonem dan kaidah pelafalan yang perlu diterapkan dalam model G2P.

Dari hasil penelusuran, ditetapkan bahwa pengembangan model G2P terintegrasi VITS dapat menjadi pendekatan potensial untuk meningkatkan akurasi fonetik dan kejelasan pelafalan sistem TTS Bahasa Indonesia.

3.4.1.2 Penentuan Jenis AI yang Digunakan

Berdasarkan hasil eksplorasi literatur, jenis kecerdasan buatan yang sesuai untuk penelitian ini adalah deep learning berbasis model sekuensial.

Pendekatan ini dipilih karena mampu mempelajari relasi antara dua urutan simbol berbeda dalam hal ini, grafem dan fonem yang merupakan inti dari konversi G2P.

Beberapa pendekatan model yang menjadi acuan dalam studi literatur meliputi:

- Transformer dan BERT, yang menawarkan kemampuan kontekstual lebih kuat melalui mekanisme attention, dan telah terbukti unggul dalam berbagai tugas NLP termasuk pelafalan dan transliterasi.
- VITS, sebagai model end-to-end yang menggabungkan representasi fonem dengan sintesis suara, memberikan naturalitas tinggi pada hasil suara.

Dari berbagai alternatif tersebut, penelitian ini memilih kombinasi model G2P berbasis BERT untuk konversi teks ke fonem, dan VITS sebagai model TTS utama. Kombinasi ini dipilih karena keduanya mewakili arsitektur modern yang telah banyak digunakan dalam sistem TTS generasi terbaru.

3.4.1.3 Eksplorasi Teknologi AI

Tahap selanjutnya adalah eksplorasi terhadap teknologi pendukung, pustaka, dan framework yang digunakan dalam implementasi. Tujuannya adalah untuk memastikan model dapat dikembangkan secara efisien dan *reproducible*.

Eksplorasi meliputi:

- Framework utama: *PyTorch* dan *Coqui TTS*, karena keduanya mendukung implementasi VITS dan kompatibel dengan model berbasis Transformer.
- Pendekatan data: mengikuti format dataset LJSpeech, yang umum digunakan dalam pelatihan TTS, dengan struktur direktori wavs/ dan berkas metadata.csv.
- Pra-pemrosesan audio: studi tentang teknik *denoising* dan *normalization* berbasis AI untuk memastikan kualitas data audio tetap tinggi.
- Evaluasi model: mengacu pada metrik-metrik yang digunakan dalam penelitian sejenis, yaitu PER untuk G2P dan MOS untuk penilaian kualitas suara.

Melalui eksplorasi ini, diperoleh rancangan awal pipeline sistem yang terdiri atas tiga modul utama:

1. Modul G2P untuk konversi teks ke fonem,
2. Modul TTS (VITS) untuk sintesis suara, dan
3. Modul evaluasi untuk menilai performa dan kualitas hasil suara.

Hasil eksplorasi ini menjadi landasan teknis dalam tahap implementasi yang dijelaskan pada bagian selanjutnya.

3.4.2 Pengumpulan Data

Pada penelitian ini, data yang dikumpulkan terdiri dari dua jenis utama, yaitu data audio untuk pelatihan model TTS dan data teks untuk pelatihan model G2P berbasis homograf.

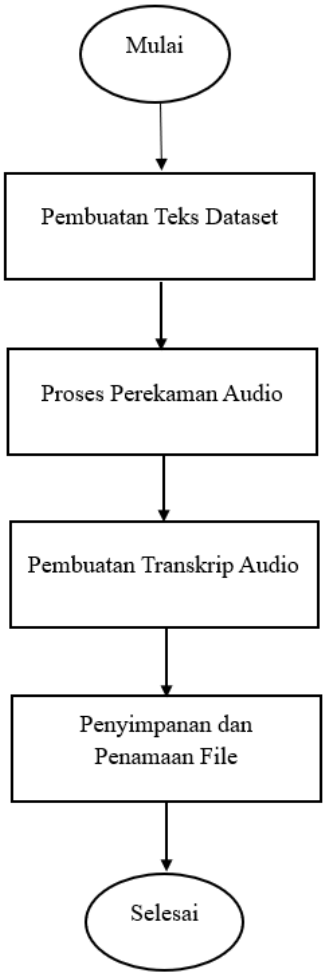
3.4.2.1 Pengumpulan Data Audio untuk TTS

Pengumpulan data audio dilakukan melalui proses perekaman suara peneliti sendiri dengan membaca kalimat dalam Bahasa Indonesia. Pendekatan ini dipilih untuk menjaga konsistensi karakter suara, menghindari permasalahan lisensi, serta memberikan kontrol penuh terhadap kualitas rekaman.

Sebelum proses perekaman, teks kalimat disusun dengan bantuan ChatGPT untuk menghasilkan variasi kalimat yang beragam dengan panjang antara 5 hingga 30 kata. Variasi tema meliputi kehidupan sehari-hari, teknologi, pendidikan, sosial, dan budaya guna memastikan cakupan kosakata yang representatif.

Proses perekaman dilakukan menggunakan mikrofon internal smartphone Xiaomi 12T dengan aplikasi perekam bawaan. Rekaman dilakukan di ruangan tertutup untuk meminimalkan gangguan suara latar. Setiap kalimat dibacakan dengan intonasi stabil, artikulasi jelas, dan tempo konsisten. Jika terjadi kesalahan pengucapan, rekaman diulang hingga sesuai dengan teks.

Seluruh file audio direkam dalam format WAV dengan *sample rate* 44.100 Hz. Dataset yang diperoleh terdiri dari 2.029 klip audio dengan rata-rata 8,41 kata per audio dan total 5.028 kata unik.



Gambar 3. 2 Pengumpulan Data Audio untuk Teks.

3.4.2.2 Pengumpulan Data G2P

Pengumpulan data G2P difokuskan pada pembuatan dataset berbasis kalimat yang mengandung kata homograf. Dataset ini bertujuan untuk melatih model agar mampu menentukan pelafalan yang tepat berdasarkan konteks kalimat.

a. Dataset Homograf Berbasis Kalimat

Dataset homograf berbasis kalimat dikembangkan untuk melatih model G2P dalam memahami perbedaan pelafalan kata berdasarkan konteks semantik kalimat. Dalam Bahasa Indonesia terdapat sejumlah kata yang memiliki penulisan sama namun makna dan pelafalannya berbeda (homograf). Model perlu diberikan contoh kontekstual agar mampu memprediksi pelafalan yang tepat sesuai makna yang dimaksud.

Pada penelitian ini, pembuatan dataset dilakukan dengan memanfaatkan ChatGPT untuk menghasilkan kalimat secara sistematis dan bervariasi. Pendekatan ini dipilih karena mampu mempercepat proses penyusunan data serta menghasilkan variasi konteks yang luas dalam waktu relatif singkat. Setiap kata homograf dibuat dalam berbagai skenario kalimat agar mencerminkan penggunaan alami dalam Bahasa Indonesia.

Kata homograf yang digunakan dalam penelitian ini meliputi:

- Apel
- Per
- Memerah
- Seri
- Serang
- Jeli

Untuk setiap kata, dibuat beberapa variasi kalimat yang mewakili makna berbeda. Sebagai contoh, kata *"apel"* dapat bermakna buah atau kegiatan upacara. Perbedaan makna tersebut memengaruhi pelafalan fonem, khususnya pada vokal /e/. Oleh karena itu, setiap kemunculan kata target dalam kalimat diberi anotasi pelabelan sense yang membedakan kategori makna dan pelafalan.

Format penyimpanan dataset menggunakan struktur JSONL (JSON Lines), di mana setiap baris merepresentasikan satu data kalimat yang terdiri atas atribut berikut:

- text: kalimat lengkap
- index: posisi kata target dalam kalimat
- sense: kategori makna kata homograf
- sense_id: label numerik pelafalan

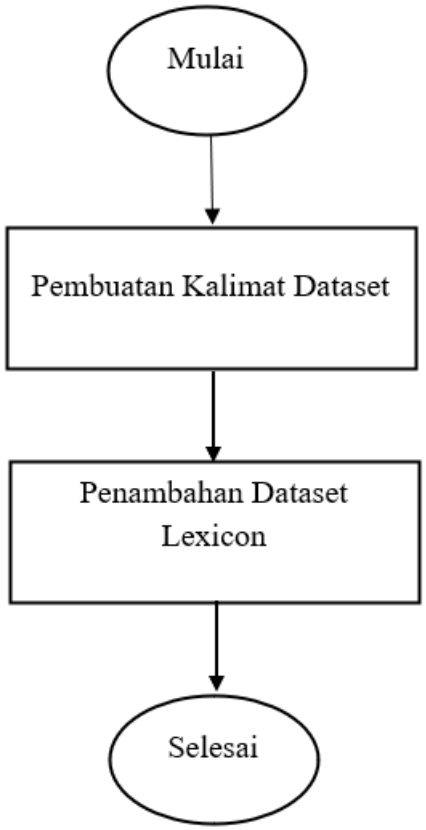
Atribut *index* digunakan untuk menandai posisi kata homograf di dalam kalimat, sehingga sistem dapat mengetahui lokasi kata target saat proses pelatihan. Sementara itu, atribut *sense* dan *sense_id* berfungsi sebagai label klasifikasi yang diprediksi oleh model.

Total dataset berbasis kalimat yang dihasilkan berjumlah 27.601 baris data, yang kemudian digunakan dalam proses normalisasi teks, target marking, serta pembagian data latih, validasi, dan uji. Dataset ini menjadi komponen utama dalam

pelatihan model G2P untuk mempelajari hubungan antara konteks kalimat dan pelafalan fonem yang tepat.

b. Dataset Lexicon

Dataset lexicon yang digunakan dalam penelitian ini diambil dari repositori id_word2phoneme yang tersedia di HuggingFace Dataset https://huggingface.co/datasets/bookbot/id_word2phoneme ini merupakan leksikon kata fonem yang awalnya berasal dari kumpulan data IPA, telah dimodifikasi untuk kebutuhan Bahasa Indonesia dengan menghapus huruf aksen dan menyederhanakan representasi fonem agar sesuai dengan standar Bahasa Indonesia. Dataset ini mencakup pasangan grafem dan fonem dalam notasi IPA untuk lebih dari 27 ribu baris pasangan data, sehingga menyediakan representasi pelafalan fonem yang luas dan bermanfaat sebagai referensi dasar pelafalan dalam sistem G2P. Struktur data yang berupa pasangan kata dan fonem memudahkan integrasi dalam model G2P karena dapat digunakan sebagai sumber ground truth untuk belajar pemetaan antara teks dan fonem yang benar.



Gambar 3. 3 Pengumpulan Data untuk G2P

3.4.3 Pra-pemrosesan Data

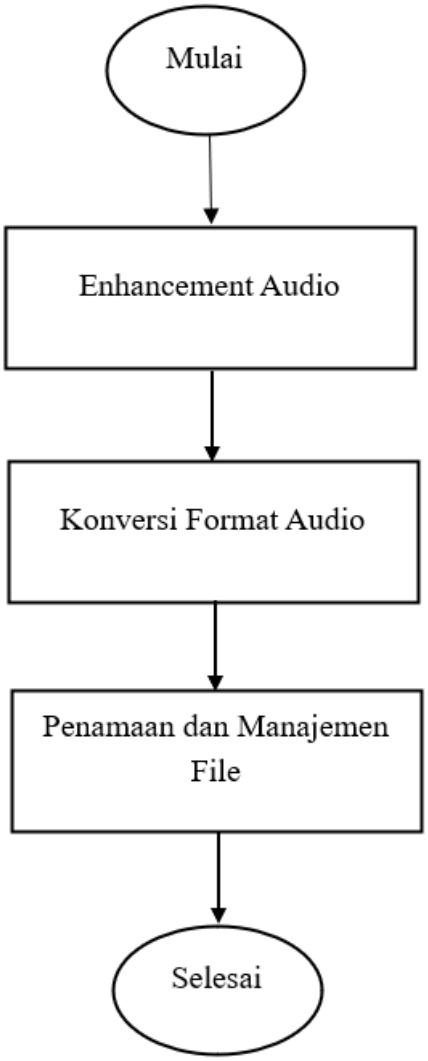
Tahap pra-pemrosesan data dilakukan untuk memastikan seluruh dataset berada dalam format yang seragam, bersih, dan siap digunakan dalam proses pelatihan model. Prapemrosesan dilakukan secara terpisah untuk data audio TTS dan data teks G2P karena keduanya memiliki karakteristik dan kebutuhan pemrosesan yang berbeda.

3.4.3.1 Pra-pemrosesan Data Audio untuk TTS

Pada tahap pra-pemrosesan data audio, seluruh file audio hasil rekaman diunggah ke layanan Adobe Enhance Speech untuk dilakukan pembersihan suara. Setiap file diproses secara otomatis untuk mengurangi noise latar dan meningkatkan kejernihan vokal. File hasil enhancement kemudian diunduh kembali dalam format WAV.Audio yang telah dibersihkan selanjutnya dikonversi menggunakan library Python librosa. Proses konversi dilakukan dengan mengubah format menjadi WAV mono, menyesuaikan *sample rate* menjadi 22.050 Hz, serta menetapkan bit depth 16-bit agar sesuai dengan kebutuhan model.

Penelitian ini menggunakan format audio WAV mono karena WAV bersifat *lossless* sehingga kualitas suara tetap terjaga dan tidak mengalami penurunan akibat kompresi, sedangkan mono dipilih karena proses TTS umumnya hanya membutuhkan satu kanal suara sehingga lebih efisien dalam penyimpanan dan mempercepat pemrosesan. *Sample rate* 22.050 Hz digunakan karena sudah cukup untuk merepresentasikan rentang frekuensi penting pada suara manusia dengan baik serta memberikan keseimbangan antara kualitas audio dan beban komputasi dibanding *sample rate* yang lebih tinggi. Sementara itu, bit depth 16-bit dipilih karena merupakan standar umum pada speech processing yang mampu menangkap detail amplitudo suara secara stabil, sehingga sesuai untuk kebutuhan ekstraksi fitur seperti Mel-Spectrogram dan pelatihan model TTS.

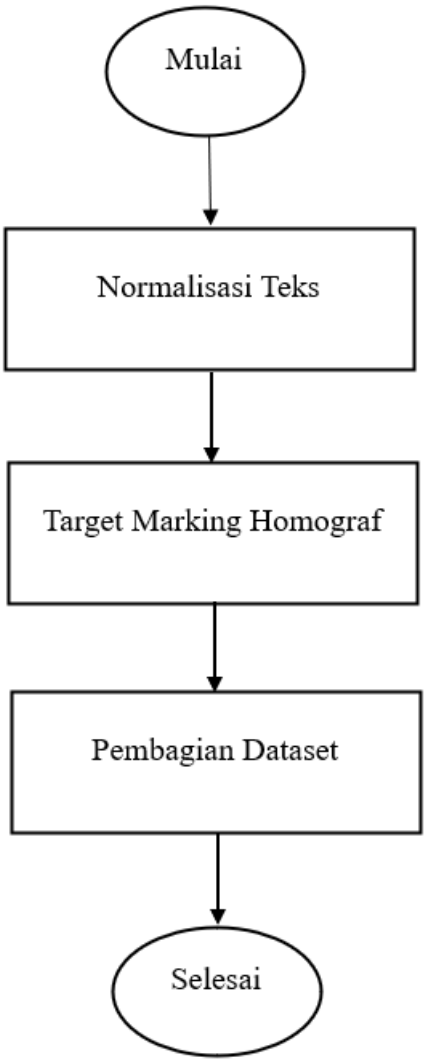
Setelah konversi, seluruh file audio disimpan dalam folder wavs/ dengan penamaan yang konsisten dan terurut. Selanjutnya dibuat file metadata.csv yang berisi pasangan nama file audio dan teks transkrip dengan format <nama_file>|<teks>|<teks>.



Gambar 3. 4 Pra-pemrosesan Data untuk Audio

3.4.3.2 Pra-pemrosesan Data G2P

Pada tahap pra-pemrosesan data G2P, seluruh kalimat dalam dataset terlebih dahulu dinormalisasi. Proses normalisasi dilakukan dengan mengubah teks menjadi huruf kecil serta mengonversi angka, satuan ukuran, mata uang, tanggal, dan waktu ke bentuk penulisan lengkap. Setelah normalisasi, dilakukan proses target marking untuk menandai kata homograf. Token [TGT] disisipkan sebelum dan sesudah kata target berdasarkan indeks yang telah ditentukan. Sebelum penyisipan, sistem memvalidasi indeks kata untuk memastikan posisi sesuai dengan panjang kalimat. Dataset yang telah dinormalisasi dan diberi penanda kemudian dibagi menjadi data latih, validasi, dan data uji dengan rasio 80:10:10.



Gambar 3.5 Pra-pemrosesan Data untuk G2P.

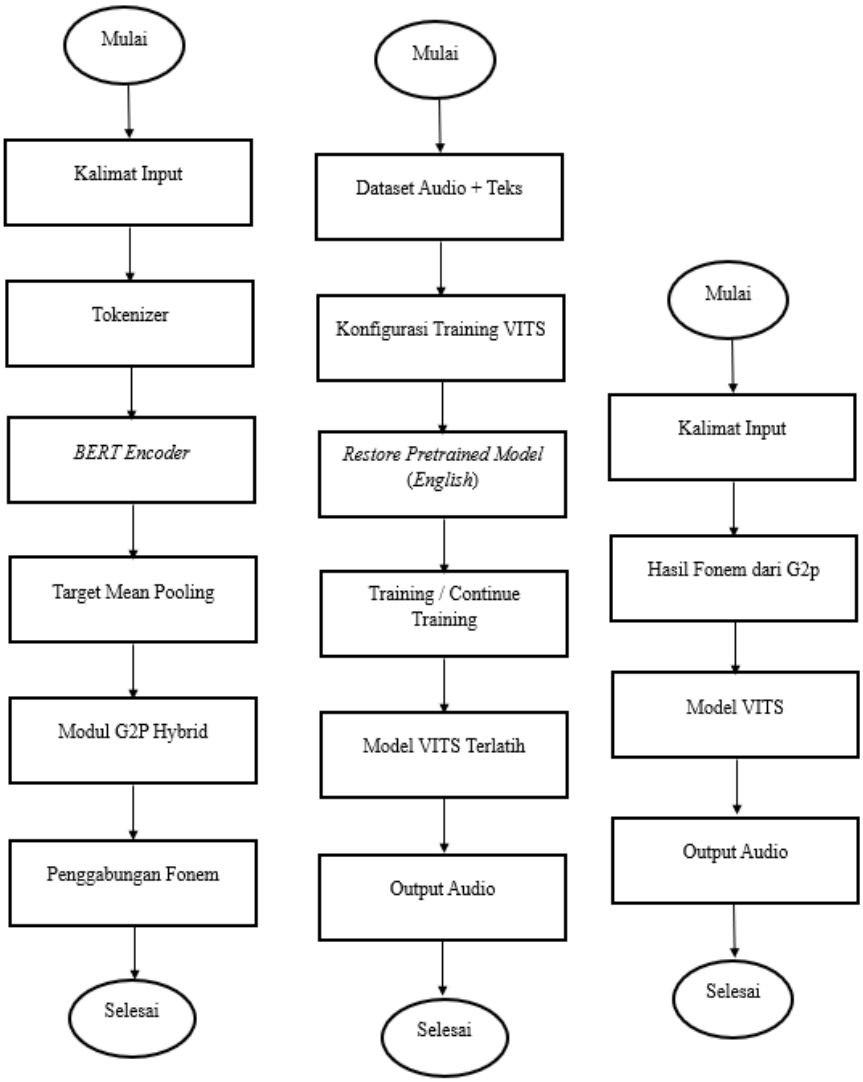
3.4.4 Pengembangan Model

Tahap pengembangan model dilakukan setelah proses pengumpulan data dan preprocessing selesai. Tujuan utama dari tahap ini adalah membangun sistem TTS yang dilengkapi dengan modul G2P untuk meningkatkan akurasi pelafalan. Secara umum, tahapan ini mencakup pembangunan model baseline TTS, pengembangan model G2P, serta integrasi keduanya ke dalam pipeline sistem berbasis VITS.

Model baseline TTS dirancang sebagai sistem pembanding yang tidak menggunakan modul G2P. Baseline ini digunakan untuk mengukur performa dasar sistem TTS, sehingga peningkatan yang dihasilkan setelah penambahan G2P dapat dianalisis secara lebih objektif, terutama pada aspek ketepatan fonetik dan kualitas audio yang dihasilkan.

Selanjutnya, dikembangkan model G2P yang berfungsi untuk mengubah teks menjadi urutan fonem IPA. Model ini menggunakan pendekatan deep learning dengan arsitektur BERT, yang memiliki kemampuan memahami konteks kata secara menyeluruh.

Output dari model G2P berupa urutan fonem yang kemudian digunakan sebagai input bagi model VITS dalam proses sintesis suara. Integrasi dilakukan dengan membangun pipeline otomatis yang menjalankan konversi teks menjadi fonem terlebih dahulu sebelum tahap pembangkitan audio oleh VITS.



Gambar 3.6 Model TTS, Model G2P dan Model TTS dengan G2P.

Berdasarkan gambar 3.6 tersebut memperlihatkan perbandingan antara model TTS *baseline*, model G2P, dan sistem integrasi keduanya. Pada model TTS *baseline*, teks

dalam bentuk grafem langsung diproses oleh VITS tanpa konversi fonem. Model G2P secara terpisah bertugas mengklasifikasikan *sense* homograf menggunakan BERT dan mengonversi seluruh kata menjadi fonem melalui pendekatan *hybrid*. Pada sistem yang diusulkan, output fonem dari model G2P digunakan sebagai input model VITS, sehingga pelafalan homograf dapat disesuaikan dengan konteks kalimat sebelum proses sintesis suara dilakukan. Integrasi ini bertujuan meningkatkan akurasi pelafalan pada hasil audio yang dihasilkan.

3.4.5 Evaluasi dan Penyesuaian Model

Tahap evaluasi dilakukan untuk menilai kinerja sistem dari dua aspek utama, yaitu akurasi model G2P dan kualitas keluaran suara TTS. Tujuan evaluasi ini adalah memastikan bahwa integrasi G2P ke dalam pipeline VITS benar-benar memberikan peningkatan performa dibandingkan model baseline. Evaluasi dilakukan secara terpisah pada masing-masing komponen agar hasil analisis lebih terarah dan mudah dibandingkan.

Evaluasi model G2P dilakukan menggunakan confusion matrix untuk melihat kesesuaian antara fonem prediksi dan fonem referensi, sekaligus mengidentifikasi fonem yang paling sering tertukar. Selain itu, menggunaakn PER agar mengetahui kesalahan rata rata dari fonemnya. Sementara itu, evaluasi sistem TTS dilakukan menggunakan metode subjektif MOS, di mana beberapa pendengar memberikan penilaian skala 1–5 berdasarkan naturalitas, kejernihan pelafalan, dan kenyamanan suara. Hasil evaluasi kemudian dianalisis untuk melihat peningkatan kualitas setelah integrasi G2P

V. KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Penelitian ini berhasil mengembangkan sistem TTS Bahasa Indonesia berbasis arsitektur VITS dengan menggunakan dataset sebanyak 2.029 klip audio yang direkam secara mandiri dan diproses mengikuti standar format LJSpeech. Sistem *baseline* mampu menghasilkan suara sintetis yang dapat dipahami dengan baik, dengan nilai rata-rata MOS sebesar 3,80 berdasarkan penilaian 6 responden terhadap 40 sampel audio uji. Nilai tersebut menunjukkan bahwa sistem yang dikembangkan telah mencapai kualitas suara pada kategori baik.
2. Integrasi modul G2P berbasis BERT ke dalam *pipeline* TTS berhasil dilakukan melalui pengembangan frontend phonemizer Bahasa Indonesia pada Coqui TTS. Model G2P dilatih menggunakan 27.601 data berbasis kalimat dan berhasil mencapai nilai PER sebesar 2,15% dengan total 38 kesalahan dari 1.765 fonem referensi pada data uji. Hasil ini menunjukkan bahwa sistem konversi G2P telah berjalan secara akurat sebelum digunakan dalam proses sintesis suara.
3. Berdasarkan hasil pengujian kinerja model G2P pada dua konfigurasi yang diterapkan, diperoleh nilai *accuracy* sebesar 0,9907 serta *precision*, *recall*, dan F1-score masing-masing sebesar 0,9903 pada kedua konfigurasi. Evaluasi PER menghasilkan 2,15% dengan 50 kalimat yang di uji dengan dua konfigurasi model yang berbeda. Hasil tersebut menunjukkan bahwa model G2P memiliki tingkat akurasi yang sangat tinggi dan kinerja yang konsisten, serta tidak terdapat perbedaan yang signifikan antara konfigurasi pertama dan konfigurasi kedua.
4. Berdasarkan hasil pengujian kualitas suara menggunakan penilaian subjektif, sistem TTS tanpa penerapan G2P memperoleh nilai sebesar 3,80, sedangkan sistem TTS dengan penerapan G2P memperoleh nilai 2,8

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan, beberapa saran yang dapat diberikan untuk pengembangan lanjutan adalah sebagai berikut:

1. Peneliti selanjutnya disarankan untuk melakukan optimasi lebih lanjut pada integrasi modul G2P dengan model TTS berbasis VITS, khususnya dalam penyesuaian representasi fonem agar lebih sesuai dengan karakteristik fonetik Bahasa Indonesia dan kebutuhan model sintesis suara.
2. Disarankan untuk menggunakan dataset audio dengan jumlah yang lebih besar, kualitas rekaman yang lebih tinggi, serta melibatkan lebih dari satu penutur agar model TTS dan voice cloning dapat mempelajari variasi suara dan meningkatkan naturalitas hasil sintesis.
3. Evaluasi kualitas suara pada penelitian berikutnya disarankan untuk melibatkan jumlah pendengar yang lebih banyak dan beragam, serta mengombinasikan metode subjektif MOS dengan metrik objektif tambahan agar hasil evaluasi menjadi lebih komprehensif dan representatif.
4. Penelitian selanjutnya disarankan untuk melakukan pelatihan lebih lama

DAFTAR PUSTAKA

[1] J. L. F. Alan M. Jette, Carol Mason Spicer, *The Promise of Assistive Technology to Enhance Activity and Work Participation*. 2017. doi: 10.17226/24740.

[2] P. Taylor, *Text to speech synthesis*, vol. 0. Cambridge University Press.

[3] A. Chaer, *Linguistik Umum*. 2014.

[4] M. Widodo, G. E. Putrawan, and R. Perdana, "How Is a Foreign Language Pronounced? A Case Study of Indonesian as a Foreign Language Among Speakers of Other Languages," *J. Lang. Teach. Res.*, vol. 14, no. 1, pp. 231–238, 2023, doi: 10.17507/jltr.1401.24.

[5] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[7] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *Proc. Mach. Learn. Res.*, vol. 139, pp. 5530–5540, 2021.

[8] L. F. Huang, *Artificial intelligence*, vol. 4. 2010. doi: 10.1109/ICCAE.2010.5451578.

[9] T. S. Levitt and E. J. Horvitz, "Review of Heuristics: Intelligent Search Strategies for Computer Problem Solving," vol. 8, no. 1, 1987.

[10] J. Main, T. Dillon, and S. Shiu, "A Tutorial on Case-Based Reasoning," pp. 1–28, 2001.

[11] L. I. Kuncheva, "Fuzzy sets," vol. 353, pp. 79–115, 2000, doi: 10.1007/978-3-7908-1850-5_4.

[12] M. Borovcnik, H.-J. Bentz, and R. Kapadia, *A Probabilistic Perspective*. 1991. doi: 10.1007/978-94-011-3532-0_2.

[13] R. E. Fikes and N. I. J. Niisson, “Strips: a New Approach To the Application of Theorem Proving To Problem Solving,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2, pp. 608–620, 1971.

[14] A. C. Ian Goodfellow, Yoshua Bengio, “Deep Learning,” *MIT Press*, vol. 521, no. 7553, p. 785, 2017, doi: 10.1016/B978-0-12-391420-0.09987-X.

[15] J. John R. Deller, John G. Proakis, and John H. L. Hansen, “Discrete Time Processing of Speech Signals”.

[16] J. Shen *et al.*, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 4779–4783, 2018, doi: 10.1109/ICASSP.2018.8461368.

[17] A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” pp. 1–15, 2016. Available: http://arxiv.org/abs/1609.03499

[18] N. Kalchbrenner *et al.*, “Efficient neural audio synthesis,” *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 6, pp. 3775–3784, 2018.

[19] I. P. Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

[20] S. Leonte, A. Pastrav, C. Zamfirescu, and E. Puschita, “Voice Quality Evaluation in a Mobile Cellular Network: In Situ Mean Opinion Score Measurements,” *Sensors*, vol. 24, no. 20, 2024, doi: 10.3390/s24206630.

[21] L. T. Nguyen, T. Pham, and D. Q. Nguyen, “XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-Augus, pp. 5506–5510, 2023, doi: 10.21437/Interspeech.2023-444.

[22] G. Zhang *et al.*, “Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, pp. 456–460, 2022, doi: 10.21437/Interspeech.2022-621.

[23] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, “Phoneme-Level Bert for Enhanced Prosody of Text-To-Speech with Grapheme Predictions,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2023-June, pp. 3–7, 2023, doi: 10.1109/ICASSP49357.2023.10097074.

[24] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, “PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 5, pp. 3631–3635, 2021, doi: 10.21437/Interspeech.2021-1757.