Teknika 19 (3): 809-819

JURNAL TEKNIKA
ISSN: 0854-3143
e-ISSN: 2622-3481
Journal homepage: http://jurnal.polsri.ac.id/index.php/teknika
Journal Email: teknika@polsri.ac.id

Text to Speech Bahasa Jawa dialek Solo-Jogja dengan Metode VITS

Putri Syakira Wirdiani[1], Muhammad Fikry[*2], Yusra[3], Febi Yanto[4], Pizaini[5]
[1,*2,3,4,5] Program Studi Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
*Email Penulis Korespondensi: macfikry@gmail.com

Abstrak

Pengembangan TTS di Indonesia masih berfokus pada Bahasa Indonesia dan bahasa asing, sementara bahasa daerah seperti Jawa dialek Solo-Jogja belum banyak tersentuh, padahal memiliki banyak penutur dan nilai budaya tinggi. Penelitian ini mengembangkan model TTS untuk dialek tersebut menggunakan metode Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS). Metode ini dipilih karena kemampuannya mengintegrasikan inferensi variasional, aliran normalisasi, dan pelatihan adversarial secara end-to-end, sehingga menghasilkan suara sintetis dengan kualitas lebih alami. Dataset berisi 450 pasangan teks dan audio dari penutur asli, dibersihkan manual dan disusun dalam format LJSpeech. Sebanyak 428 data digunakan untuk pelatihan dan 22 untuk evaluasi. Model dilatih menggunakan Coqui TTS di Google Colab dengan fonemizer eSpeak. Setelah pelatihan, model terbaik digunakan untuk menyintesis 50 kalimat uji yang dinilai oleh lima penutur asli menggunakan metode MOS. Rata-rata skor yang diperoleh adalah 4,088, melampaui standar minimum 4,0. Meski begitu, masih ada kekurangan dalam kejelasan fonem dan kealamian jeda. Hasil ini menunjukkan potensi besar TTS untuk pelestarian bahasa daerah dan pengembangan teknologi serupa untuk bahasa lokal lainnya.

Kata kunci— *Bahasa Jawa, Text-to-Speech, VITS, MOS, Sintesis Suara*

Abstract

TTS development in Indonesia still largely focuses on Bahasa Indonesia and foreign languages, while regional languages like the Javanese Solo-Jogja dialect remain underexplored—despite having many speakers and rich cultural value. This study develops a TTS model for the dialect using the Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS) method. VITS is chosen for its ability to integrate variational inference, normalizing flows, and adversarial training in an end-to-end manner, allowing for more natural-sounding synthetic speech. The dataset consisted of 450 text-audio pairs from native speakers, manually cleaned and formatted in LJSpeech. A total of 428 samples were used for training and 22 for evaluation. The model was trained using Coqui TTS in Google Colab with the help of the eSpeak phonemizer. After training, the best-performing model synthesized 50 test sentences, which were rated by five native speakers using the MOS method. The average score was 4.088, exceeding the minimum quality standard of 4.0. However, issues in phoneme clarity

Received: June 13, 2025; Online Available: July 22, 2025

and pause naturalness remain. These results highlight the strong potential of TTS in preserving local languages and advancing similar technologies for other regional dialects.

Keywords— *Javanese language, Mean Opinion Score, Text to Speech, VITS, Voice Synthesis*

1. PENDAHULUAN

Sebagian besar masyarakat Indonesia menjadikan bahasa daerah sebagai bahasa ibu. Bahasa daerah memiliki peran penting sebagai sarana budaya, penghubung intra-etnis, penguat hubungan sosial, serta sebagai media untuk memahami sejarah dan warisan leluhur melalui komunikasi lisan maupun tulisan. Selain itu, bahasa daerah juga berfungsi sebagai identitas, ciri khas, alat komunikasi, dan instrumen budaya yang telah bertahan selama berabad-abad bahkan ribuan tahun [1]. Salah satu bahasa daerah dengan jumlah penutur terbesar di Indonesia adalah bahasa Jawa. Bahasa ini digunakan oleh suku Jawa yang mayoritas tinggal di Jawa Tengah, Yogyakarta, dan Jawa Timur, serta beberapa wilayah lain seperti Banten dan Jawa Barat [2].

Bahasa Jawa memiliki beragam dialek, namun dialek Solo-Jogja sering dianggap sebagai standar dalam penggunaan bahasa Jawa. Hal ini terlihat pada buku pembelajaran bahasa Jawa, seperti "Tantri Basa Kelas 4," yang ditulis menggunakan dialek tersebut [3]. Kabupaten Temanggung dan Magelang juga dianggap menggunakan dialek Solo-Jogja sebagai dialek standar dalam berkomunikasi [4]. Selain itu, terdapat penelitian [5] yang membahas khotbah Jum'at di Cilacap menjadikan dialek Solo-Jogja sebagai acuan bahasa Jawa baku.

Keanekaragaman bahasa daerah di Indonesia menghadapi tantangan besar, sehingga pelestariannya menjadi tanggung jawab bersama. Hal ini berdasar pada penelitian Lembaga Ilmu Pengetahuan Indonesia (LIPI), dari 746 bahasa daerah yang ada, 169 di antaranya terancam punah. Hal ini disebabkan oleh jumlah penutur yang sangat sedikit, mayoritas sudah lanjut usia, tidak adanya generasi penerus, serta lokasi penutur yang terpencil [2].

Saat ini, penggunaan bahasa daerah dalam kehidupan sehari-hari semakin menurun. Oleh sebab itu, kemajuan teknologi komunikasi diharapkan dapat membantu upaya pelestarian budaya yang hampir punah akibat perkembangan zaman [6]. Hal ini menunjukkan upaya pengembangan teknologi untuk bahasa daerah, termasuk bahasa Jawa, perlu untuk dikembangkan.

Perkembangan teknologi informasi dan komunikasi yang berlangsung saat ini, membawa perubahan besar dalam sistem komunikasi di Indonesia. Di masa lalu, menyampaikan informasi sangat terbatas karena teknologi yang tersedia sangat sederhana. Namun, perkembangan zaman telah membuat sistem komunikasi menjadi lebih mudah [7]. Kemajuan ini memberikan kenyamanan dan meningkatkan aksesibilitas berkomunikasi dalam berbahasa, sekaligus membawa banyak manfaat, seperti mempermudah pekerjaan dan memenuhi berbagai kebutuhan. Salah satu teknologi yang terus berkembang seiring waktu adalah kecerdasan buatan (*Artificial Intelligence*), yang kini menjadi bagian tak terpisahkan dari berbagai aspek kehidupan [8].

Salah satu aplikasi AI yang semakin populer adalah *Text to Speech* (TTS), yang memungkinkan komputer untuk mengubah teks tertulis menjadi suara yang dapat didengar [16]. TTS adalah teknologi sintesis suara yang mengonversi teks dalam suatu bahasa menjadi ucapan sesuai dengan teks dalam bahasa yang digunakan [9]. Teknologi ini bekerja secara otomatis menghasilkan ucapan dari teks dengan melalui proses transkripsi grafem ke fonem, menggunakan sampel suara yang mirip dengan suara asli [10]. Tujuan utama TTS adalah menghasilkan ucapan yang jelas, alami, dan fasih, meniru cara pengucapan manusia [11]. Salah satu pendekatan terkini yang digunakan adalah pengembangan dengan menggunakan metode VITS.

Metode *Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech* (VITS) merupakan teknik pengembangan TTS yang menggabungkan inferensi variasional, aliran normalisasi, dan pelatihan adversarial [12]. Beberapa model sebelumnya seperti *FastSpeech* dan *EATS* telah mencoba menghubungkan penyandi teks dan suara secara langsung, namun kualitas hasilnya masih belum baik. VITS mengatasi keterbatasan ini melalui integrasi inferensi variasional dan aliran normalisasi, sehingga mampu menghasilkan sintesis suara dengan kualitas yang lebih unggul [17]. Keunggulan VITS juga terlihat dalam

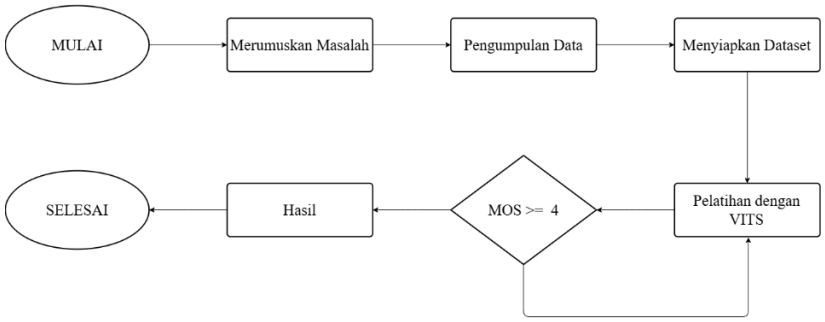
kemampuannya menghadapi kebisingan. Bahkan pada lingkungan dengan gangguan *Additive White Gaussian Noise* (AWGN), VITS tetap menghasilkan ucapan yang dapat dimengerti [13]. Maka, Pemilihan VITS dalam penelitian ini didasarkan pada keunggulannya dalam menghasilkan suara yang lebih alami, stabil, dan mendekati pelafalan manusia, sehingga dianggap paling relevan untuk merepresentasikan kekayaan fonetik dialek Solo–Jogja secara lebih akurat. Dalam konteks bahasa Jawa, metode ini diharapkan dapat menangkap nuansa intonasi dan pola pengucapan khas yang menjadi ciri unik dialek Solo-Jogja.

Dengan demikian, penelitian ini menjadi penting karena saat ini belum tersedia model TTS yang dikembangkan khusus untuk dialek Solo–Jogja, padahal dialek ini digunakan secara luas dan menjadi standar dalam banyak buku ajar serta komunikasi budaya. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan aksebilitas bahasa daerah dan kemudahan berkomunikasi masyarakat Indonesia melalui implementasi teknologi TTS yang mendukung bahasa-bahasa daerah di Indonesia, khususnya bahasa Jawa. Selain sebagai upaya pelestarian bahasa, implementasinya juga diharapkan dapat digunakan dalam memudahkan akses informasi bagi penutur bahasa daerah, mendukung upaya pembelajaran bahasa lokal di sekolah, serta dapat digunakan dalam industri kreatif yang memerlukan konten berbasis suara yang khas dan autentik. Hal ini juga bertujuan untuk mendorong pengembangan teknologi berbasis budaya lokal, yang tidak hanya berfokus pada bahasa nasional, tetapi juga mengakomodasi keragaman bahasa dan budaya di Indonesia.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui beberapa tahapan utama yang dimulai dari pengumpulan data, pembuatan *dataset*, pelatihan model, hingga tahap pengujian. Tahapan awal dimulai dengan studi literatur, di mana peneliti melakukan penelusuran dan pembelajaran dari berbagai referensi seperti jurnal, artikel, dan buku yang berkaitan dengan studi kasus yang diteliti. Tahapan kedua adalah pembuatan data teks, yaitu penyusunan 500 kalimat dalam bahasa Jawa yang terdiri dari 4–9 kata per kalimat. Kalimat-kalimat tersebut disusun berdasarkan hasil diskusi dan wawancara dengan penutur asli bahasa Jawa dialek Solo-Jogja untuk memastikan keotentikan dan kelayakan penggunaan kalimat tersebut dalam proses perekaman. Tahap ketiga adalah perekaman suara, dilakukan menggunakan perangkat telepon genggam dengan narasumber merupakan penutur asli bahasa Jawa dialek Solo-Jogja. Diagram Alir disajikan pada Gambar 1.



Gambar 1 *Flowchart* Metodologi Penelitian

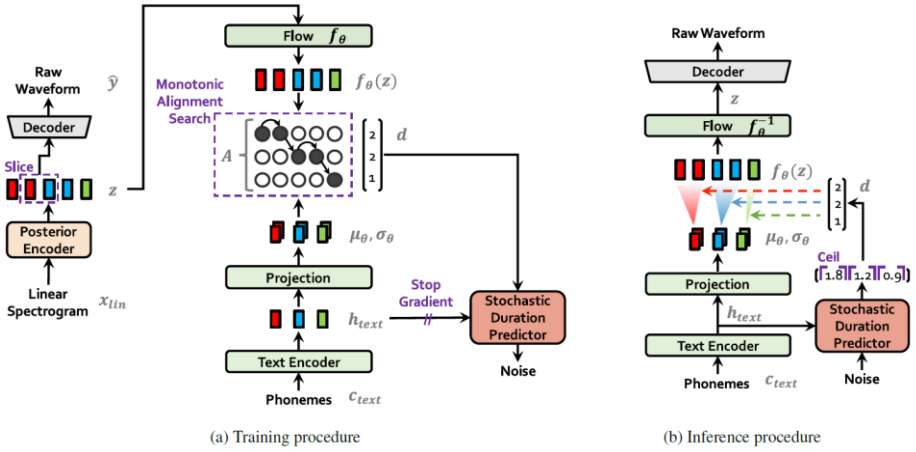
2.2 Pembuatan Dataset

Setelah data suara berhasil direkam, tahapan selanjutnya adalah pembuatan *dataset*. Dataset ini disusun melalui evaluasi kualitas *audio* hasil rekaman yang dilakukan dengan pemutaran ulang (*playback*). Jika kualitas tidak memenuhi standar, dilakukan perekaman ulang. Rekaman suara yang telah lolos evaluasi dilakukan *cleaning* audio lalu diubah formatnya dari .mp3 menjadi .wav dengan durasi tiap file antara 1–5 detik. Seluruh file disimpan dalam folder

bertajuk "wavs", dan dilengkapi dengan file CSV yang dimuat dengan format *LJSpeech*. *Dataset* ini kemudian dibagi menjadi dua bagian, yakni 450 data latih yang digunakan untuk proses pelatihan model, dan 50 data uji yang digunakan untuk tahap pengujian.

2.3 Pemodelan

Proses pemodelan dilakukan menggunakan pendekatan VITS (*Variational Inference with adversarial learning for end-to-end Text-to-Speech*) seperti tersaji pada Gambar 2 yang memiliki tahapan sebagaimana dijelaskan oleh [12] yaitu terdiri dari tahap training dan inferensi.



Gambar 2 Model VITS
Sumber: [12]

Proses pemodelan ini memiliki beberapa komponen. Pertama, *text encoder* mengubah input berupa teks fonem menjadi representasi numerik. Kemudian hasilnya diproses pada tahap *projection*, di mana representasi tersebut diubah menjadi bentuk statistik yang dapat dihubungkan dengan representasi audio. Selanjutnya ada proses *monotonic alignment search* (MAS) yang hanya terdapat pada proses *training* yang digunakan untuk mencocokkan urutan fonem dan durasi suara yang sesuai. Kemudian tahap *flow* digunakan untuk memodelkan distribusi sinyal suara secara fleksibel. Tahap *stochastic duration predictor* untuk memperkirakan durasi fonem yang akan memengaruhi panjang sinyal audio, pada proses training tahapan ini bekerja menggunakan output dari MAS sedangkan pada proses inferensi tahapan ini berperan melalui best model dari hasil *training*. Pada tahap *posterior encoder*, sinyal audio mentah diolah menjadi *encoding* yang lebih kecil dan efisien, sebelum akhirnya dikonversi kembali menjadi sinyal audio melalui proses *decoder*.

2.4 Pengujian MOS

Penelitian ini membutuhkan metode penilaian untuk mengevaluasi suara yang dihasilkan dari pengujian *Text to Speech*, yaitu menggunakan *Mean Opinion Score* (MOS) yang merupakan metode dalam mengukur kinerja dan kualitas dari aplikasi yang dibangun dari sisi pengguna [15]. Pada tahap ini, sebanyak 50 data teks dalam bentuk kalimat diujikan ke dalam model, dan hasil suaranya diperdengarkan kepada tiga orang penguji yang merupakan penutur asli atau memiliki pemahaman yang baik terhadap bahasa Jawa. Para penguji memberikan penilaian pada skala 1–5 berdasarkan kesesuaian dan kealamian hasil output suara seperti pada Tabel 1. Apabila skor rata-rata dari hasil pengujian masih di bawah angka 4, maka dilakukan pemodelan ulang hingga diperoleh suara yang dianggap cukup baik, alami, dan sesuai dengan pelafalan aslinya. MOS (*Mean Opinion Score*) sendiri berfungsi untuk menentukan tingkat kualitas kepuasaan pengguna terhadap layanan yang digunakan [14]. Pada pengujian MOS terdapat skala penilaian dari 1-5, dengan nilai 1 merupakan sangat buruk dan nilai 5 adalah sangat baik:

Tabel 1 Skala Penilaian MOS

Skala	Kualitas	Kriteria
1	Buruk	Ucapan tidak dapat dipahami, perangkaian ucapan sangat tidak jelas
2	Kurang Baik	Ucapan tidak dapat dipahami, perangkaian ucapan kurang jelas
3	Cukup Baik	Ucapan dapat dipahami, perangkaian ucapan cukup jelas
4	Baik	Ucapan dapat dipahami, perangkaian ucapan jelas
5	Sangat Baik	Ucapan dapat dipahami, perangkaian ucapan sangat jelas

Pengujian MOS memiliki rumus (1) yaitu:

$$MOS = \frac{\sum R}{N \, x \, M} \quad (1)$$

Pengujian MOS untuk mendapatkan kesimpulan hasil berdasarkan total skor rata-rata pada seluruh atribut dimana:

MOS = total skor rata-rata pengujian
R = Total Rating yang diberikan pendengar
N = Jumlah pendengar
M = Jumlah kalimat pegujian

Tujuan MOS sebagai sarana untuk mengukur, mengevaluasi dan meningkatkan kualitas suara yang dihasilkan oleh model TTS agar sesuai dengan kebutuhan dan harapan pengguna.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Pengumpulan Data

1. Pengumpulan Data Teks

Proses penyusunan data teks dalam penelitian ini dimulai dari hasil wawancara dengan penutur asli bahasa Jawa dialek Solo-Jogja, yaitu Andhika Adiputra, yang lahir, tumbuh dan bersekolah di Yogyakarta. Pemilihan penutur ini dilakukan secara sengaja karena ia memenuhi kriteria geografis, etnis, dan kebahasaan sebagai representasi otentik penutur dialek Solo–Jogja. Dari aspek usia, penutur berada dalam kelompok usia muda, yang secara sosiolinguistik dinilai paling konsisten dalam mempertahankan ciri khas fonologis dan leksikal suatu dialek karena keterlibatan aktif dalam praktik berbahasa sehari-hari [18]. Selain data wawancara, teks juga dihimpun dari literatur buku berjudul *Mardika Ngleluri Basa Jawa 3*, yang diterbitkan oleh Kementerian Pendidikan dan Kebudayaan Provinsi Jawa Tengah. Buku ini dipilih karena memuat bentuk bahasa Jawa formal berbasis pendidikan lokal. Validasi terhadap data yang dikumpulkan telah divalidasi oleh Alfu Hidayatullah, perwakilan dari Dewan Pimpinan Pusat Perhimpunan Pemuda dan Mahasiswa Jawa Riau (DPP PMJR) yang juga merupakan penutur aktif bahasa Jawa. Total data terdiri atas 500 kalimat berbahasa Jawa dialek Solo-Jogja beserta padanan terjemahannya dalam Bahasa Indonesia. Meskipun jumlah ini terbatas, data yang dipilih mencakup variasi panjang kalimat, struktur gramatikal, dan pola intonasi yang umum digunakan dalam dialek Solo–Jogja, sehingga dinilai cukup representatif untuk membangun model dasar TTS berbasis VITS.

2. Perekaman Suara

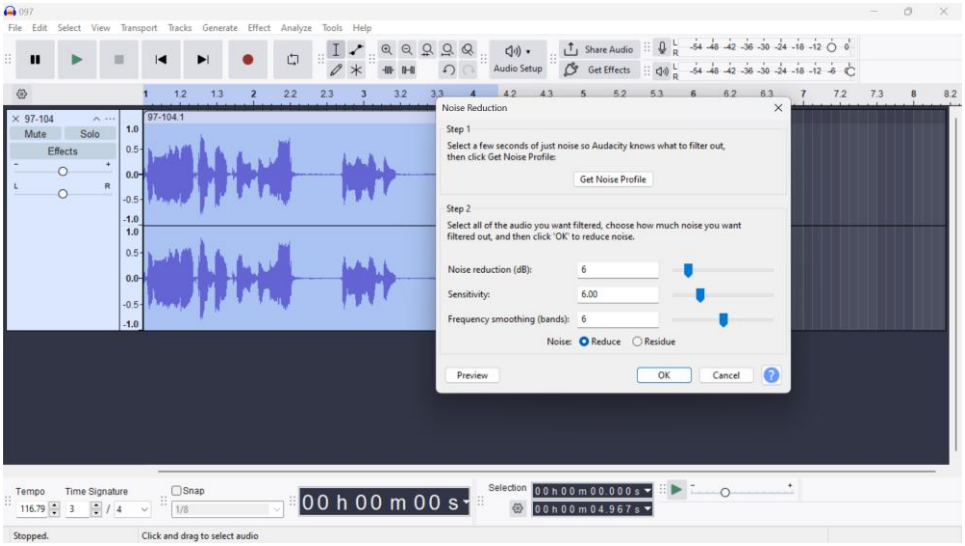
Seluruh data teks yang telah dikompilasi kemudian dibacakan oleh penutur asli laki-laki berusia 21 tahun bernama Muhammad Arif Aditya, yang memiliki kompetensi dalam pengucapan bahasa Jawa dialek Solo-Jogja. Proses perekaman dilakukan menggunakan perangkat telepon

genggam dalam kondisi lingkungan yang tenang, serta memperhatikan kejernihan, ketelitian, dan durasi rekaman agar sesuai dengan kebutuhan pemodelan suara.

3.2. Persiapan Dataset

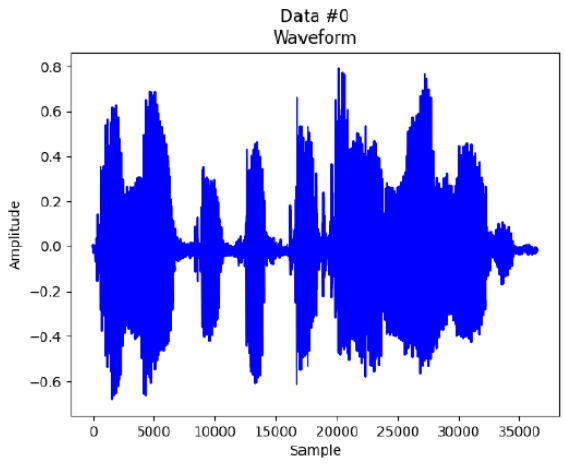
1. Pembersihan Data Audio

Proses *cleaning* data audio pada penelitian menggunakan *Software Audacity*. Pada tahap ini dilakukan pemotongan hening panjang yang terdapat di depan, di tengah atau di akhir audio. Kemudian dilakukan juga pembersihan rekaman jika terdapat *noise* kecil pada rekaman menggunakan fitur *Noise Reduction*. Gambar 3 memperlihatkan proses pembersihan *noise* dan pemotongan jeda pada sample rekaman.

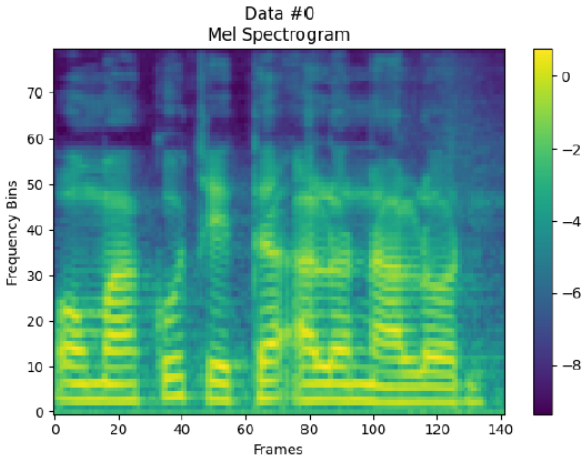


Gambar 3 *Cleaning* Data Audio dengan *Audacity*

Hasil audio kemudian disimpan dalam format .wav dengan spesifikasi teknis: *sample rate* 22050 Hz, encoding *Signed 16-bit PCM*, dan *mono channel*. Untuk memastikan kualitas sinyal audio dan kesesuaian format dengan kebutuhan model TTS berbasis VITS, dilakukan visualisasi terhadap hasil perekaman. Gambar 4 menunjukkan *waveform* dari salah satu rekaman setelah dilakukan pembersihan, sedangkan Gambar 5 memperlihatkan *mel-spectrogram* yang merepresentasikan fitur akustik yang digunakan dalam pelatihan model.



Gambar 4 *Waveform* Data



Gambar 5 Visualisasi *Mel-spectogram*

Visualisasi ini digunakan sebagai validasi awal terhadap struktur fonetik dan durasi suara, sekaligus memastikan bahwa sinyal audio siap diproses pada tahap ekstraksi fitur oleh model VITS.

2. Konversi Data Teks ke Format CSV

Teks berjumlah 500 kalimat dikonversi ke dalam format metadata.csv sesuai dengan standar format *LJSpeech*, yang menjadi acuan dalam pemuatan data model TTS. Format ini mengharuskan representasi teks eksplisit, tanpa singkatan, serta penulisan angka dalam bentuk kata. File CSV dibagi menjadi tiga kolom: nama file audio, teks tertulis sebagaimana ada, dan representasi pelafalan sesuai pembacaan. Karakter pemisah antar kolom menggunakan simbol "|". Misalnya *001 | Iki kedadeyan naliko aku kelas 1 | Iki kedadeyan naliko aku kelas siji.*

3. Pengunggahan *Dataset* dan Pemodelan TTS

Dataset akhir terdiri dari folder wavs/ yang berisi rekaman audio yang sudah dibersihkan dari *noise* dan memiliki format yang sudah ditentukan. Lalu file metadata.csv yang dijadikan sebagai media penyimpanan data 450 data teks untuk pelatihan dan inferensi serta datamos.csv yang digunakan untuk pengujian diunggah ke direktori *Google Drive*. Tahapan ini bertujuan untuk memudahkan integrasi selama proses pelatihan dan pengujian agar menghindari kekeliruan dalam pemanggilan data.

3.3. Hasil Pemodelan

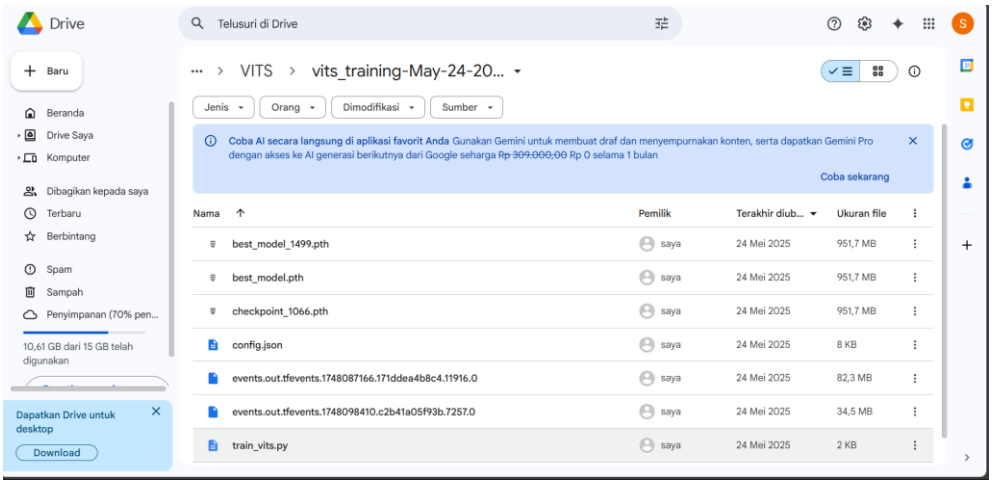
Pada tahap ini, dilakukan pemodelan TTS untuk Bahasa Jawa dengan dialek Solo-Jogja menggunakan pendekatan VITS. Pemodelan dalam penelitian ini menggunakan *Google Colab* dengan *Python 3.11.12* dan *library Coqui TTS*, serta fonemisasi *eSpeak*. Proses pelatihan dilakukan pada *hosted runtime* GPU dari *Google Colab* yaitu A100 GPU. Pemodelan ini dibagi ke dalam tiga bagian utama, yaitu: Konfigurasi data, pelatihan, dan inferensi model.

1. Konfigurasi Data

Sebelum memulai pelatihan, peneliti menyiapkan *dataset* dan parameter pelatihan, diantaranya, *dataset* yang ditetapkan 450 data. Dari jumlah ini, 5% digunakan untuk evaluasi sejumlah 22 data dan sisanya untuk pelatihan sebanyak 428 data. Pada penelitian ini peneliti mengatur *batch size* sebesar 32, sehingga untuk 428 data akan terbentuk 14 *batch* dengan masing-masing *batch* memiliki 32 data kecuali *batch* terakhir yang hanya memiliki 12 data. *Epoch* ditetapkan sebanyak 20.000, yang artinya model akan membaca seluruh dataset sebanyak 20.000 kali dengan 14 batch, maka hal ini berarti model memiliki total step sebanyak 280.000 langkah.

2. Pelatihan

Pada masa pelatihan dengan pendekatan VITS dari *Coqui TTS* menggunakan file *train_vits.py*, model melewati proses *text encoder*, *posterior encoder*, *normalizing flow*, dan *duration predictor* sesuai dengan tahapan VITS sebagaimana dijelaskan oleh [12]. Setelah proses pelatihan selesai, akan dihasilkan dua file utama. Pertama adalah *best_model.pth* yang merupakan parameter terbaik model selama proses pelatihan, kedua adalah file *config.json* yang merupakan konfigurasi model yang sesuai. Kedua file ini akan otomatis berada dalam folder penelitian seperti pada Gambar 6.



Gambar 6 Folder Data Penelitian

3. Inferensi Model

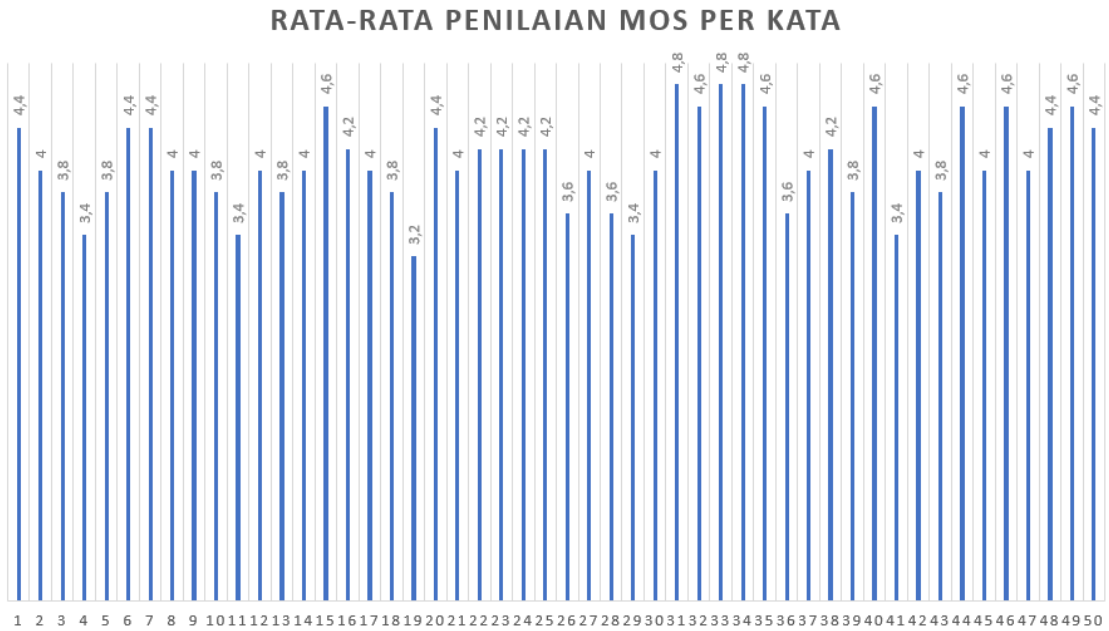
Setelah model selesai dilatih, proses inferensi dilakukan untuk menghasilkan suara dari input teks. Proses ini menggunakan file best_model.pth dan config.json yang dihasilkan dari proses pelatihan sebelumnya. Menggunakan inisialisasi:

```
SENTENCE = " Iki kedadeyan naliko aku kelas siji"
MODEL_PATH =
"/content/drive/MyDrive/TtsKira/VITS/best_model.pth"
CONFIG_PATH = "/content/drive/MyDrive/TtsKIra/VITS/config.json"
OUTPUT_PATH =
"/content/drive/MyDrive/TtsKira/dataset/output/456.wav"
!tts --text "{SENTENCE}" --model_path "{MODEL_PATH}" --
config_path "{CONFIG_PATH}" --out_path "{OUTPUT_PATH}"
```

Parameter *--text* adalah input yang akan diubah menjadi suara menggunakan *--model path* yaitu direktori file best_model.pth hasil pengujian dan *--config_path* yang merupakan direktori file *config.json* lalu outputnya akan disimpan di direktori *--out_path*. Hal ini dilakukan berulang untuk 50 data yang digunakan dalam pengujian. File inilah yang disimpan dan menjadi hasil akhir dari proses inferensi yang merepresentasikan suara sintesis dari input teks. Dalam penerapannya seluruh file output dikumpulkan dalam satu folder direktori untuk diperdengarkan kepada penguji dalam pengujian menggunakan MOS

3.4. Hasil Uji Mean Opinion Score (MOS)

Hasil pengujian kualitas TTS dilakukan melalui uji *Mean Opinion Score* (MOS) terhadap 5 responden penutur asli. Untuk menggambarkan distribusi skor penilaian secara lebih rinci, Gambar 7 menyajikan grafik persebaran nilai MOS dari 50 kalimat yang diujikan.



Gambar 7 Grafik Skor MOS Per-kata

Dengan penyebaran yang relatif merata di antara para penguji. Hal ini mengindikasikan bahwa suara yang dihasilkan oleh sistem TTS umumnya dinilai baik dan sesuai dengan pelafalan yang diharapkan dalam dialek Solo–Jogja. Model memperoleh skor rata-rata keseluruhan sebesar 4,088 dari skala 1 sampai 5. Skor ini menunjukkan bahwa suara sintetis yang dihasilkan memiliki kualitas yang baik dan cukup alami, meskipun belum sepenuhnya menyerupai suara manusia.

Analisis terhadap hasil uji dan tanggapan responden mengungkapkan beberapa kendala. Pertama, artikulasi beberapa kata terdengar terlalu cepat sehingga kejelasannya menurun, seperti pada kata *"seseg-seseg"* dan *"gaweo"*. Kedua, terjadi kesalahan segmentasi pengucapan fonem, contohnya "wacaan" yang diucapkan menjadi *"wacan"*. Ketiga, jeda antarkata tidak memadai, sehingga kata-kata seperti *"sida njupuk"* terdengar menyatu menjadi *"sidanjupuk"*. Keempat, terdapat ketidaksesuaian intonasi pada vokal tertentu, seperti tekanan vokal yang berlebihan pada kata *"kemekelen"*. Terakhir, pelafalan suara dalam beberapa kasus masih terdengar kaku dan monoton, mencerminkan keterbatasan model dalam meniru prosodi dan ekspresi alami penutur manusia. Temuan-temuan tersebut memberikan masukan penting untuk pengembangan lanjutan sistem TTS bahasa Jawa, terutama dalam aspek prosodi, kejelasan fonetik, dan naturalitas pelafalan, guna mencapai sintesis suara yang lebih menyerupai penutur asli.

4. KESIMPULAN

Berdasarkan hasil penelitian mengenai model *Text-to-Speech* (TTS) untuk Bahasa Jawa dengan dialek Solo-Jogja menggunakan metode VITS, dapat disimpulkan sebagai berikut.

1. Model berhasil dibangun dengan baik. Hal ini dibuktikan melalui skor MOS sebesar 4,088 dari lima penutur asli yang menunjukkan bahwa model telah memenuhi standar.
2. Sebanyak 450 data dikumpulkan dan dibersihkan lalu disusun dalam format *LJSpeech* untuk dilatih menggunakan model VITS di *Google Colab* dan menghasilkan model terbaik pada dengan *epoch* 2000.
3. Model yang dikembangkan menghasilkan suara yang cukup alami dan mudah dipahami, meskipun masih terdapat catatan terkait kejelasan fonem, jeda antar kata, dan kealamian suara.

5. SARAN

Berdasarkan hasil penelitian, peneliti menyarankan agar model TTS yang sudah dibangun ini dikembangkan lebih jauh menjadi aplikasi-aplikasi praktis berbasis *web* atau *mobile*, seperti kamus digital, *platform* terjemahan interaktif, atau asisten suara berbahasa Jawa. Pengembangan ini diharapkan bisa mendukung proses pembelajaran, melestarikan bahasa daerah, sekaligus menjadi teknologi bantu yang berguna di lingkungan lokal. Dengan pemanfaatan yang lebih luas, model ini tidak hanya akan menjadi hasil akademik semata, tetapi juga dapat memberikan manfaat nyata bagi kehidupan sehari-hari masyarakat.

DAFTAR PUSTAKA

[1] Khairul, Haryati, S., & Yusman, Y. (2018). Aplikasi Kamus Bahasa Jawa Indonesia Dengan Algoritma. Jurnal Teknologi Informasi & Pendidikan Vol. 11 No. 1, 1-6.

[2] Nacikit, J., & Susiati, S. (2020). Pentingnya Melestarikan Bahasa Daerah, Law Archive (Jun.1, 2020), Osf.Io/Preprints/Lawarchive/Wk8xm. Jurnal Sendika: Seminar Nasional Pendidikan Fakultas Keguruan Dan Ilmu Pendidikan Universitas Ahmad Dahlan Vol. 4, No. 2, 1-7.

[3] Kodir, N. H., Arafik, M., & Taufiq, A. (2021). Pengembangan Buku Cerita Rakyat Berbahasa Jawa Untuk Siswa Sd Kelas Iv Berkultur Madura. Jurnal Pembelajaran, Bimbingan, Dan Pengelolaan Pendidikan, 544–551.

[4] Yahya, M. (2023). Kajian Kontrastif Fonologi Bahasa Jawa Dialek Wonosobo Dengan Dialek Solo-Yogyakarta. Sutasoma: Jurnal Sastra Jawa, 54-64.

[5] Prabowo, D. S. (2022). Perbedaan Pelafalan Bahasa Jawa Baku Pada Khotbah Jum'at Di Cilacap Dan Jogjasolo. Jurnal Diwangkara, 65-76.

[6] Sunardi, L., Irawan, D., & Pratami, I. (2023). Pengenalan Bahasa Daerah Lubuklinggau Dilengkapi Dengan Text To Speech Berbasis Android. Jsai : Journal Scientific And Applied Informatics Vol. 06, No. 02, 177-183.

[7] Wiryany, D., Natasha, S., & Kurniawan, R. (2022). Perkembangan Teknologi Informasi Dan Komunikasi Terhadap Perubahan Sistem Komunikasi Indonesia. Jurnal Nomosleca, 242-252.

[8] Ernawati, S., Wati, R., & Maulana, I. (2021). Penerapan Model Fountain Untuk Pengembangan Aplikasi Text Recognition Dan Text To Speech Berbasis Android Menggunakan Flutter. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (Snast) 2021, 178-186.

[9] Fitriawati, L. S., Negara, A. B., & Nyoto, R. D. (2020). Implementasi Text To Speech Pada Website Menggunakan Metode Shallow Parsing. Jurnal Nasional Informatika Dan Teknologi Jaringan Vol. 5 No.1, 61-69.

[10] Manu, G. A., & Masan, P. L. (2020). Aplikasi Text To Speech Untuk Meningkatkan Pembelajaran Bahasa Inggris Bagi Siswa Disabilitas. Jurnal Pendidikan Teknologi Informasi (Jukanti) Vol.3 No.2, 17-26.

[11] Melangi, S. (2018). Text To Speech Bahasa Indonesia Menggunakan Synthesizer Concatenation Berbasis Fonem Synthesizer Concatenation Berbasis Fonem. Jurnal Teknik Elektro Cosphi. Vol. 2. No. 2, 31- 36.

[12] Kim, J., Kong, J., & Son, J. (2021). Conditional Variational Autoencoder With Adversarial Learning For End-To-End Text-To-Speech. International Conference On Machine Learning, 5530-5540.

[13] Yang, J. (2023). The Evaluation Of Performance Related To Noise Robustness Of Vits For Speech Synthesis. Highlights In Science, Engineering And Technology Volume 57, 62-68.

[14] Irwan, R. W., Hamza, S., & Hamid, M. (2023). Analisis Perbandingan Mean Opinion Score Aplikasi Voip Facebook Messenger, Whatsapp Dan Zoom Metting Menggunakan Metode E-Model Pada Jaringan Lte. Jurnal Biosainstek. Vol. 5 No. 2, 32-38.

[15] Nofiyati, Nugroho, A. K., & Wijayanto, B. (2022). Evaluation of The Quality of Academic Information System Unsoed Using Iso 9126 and Mean Opinion Score (MOS). Jurnal Teknik Informatika (Jutif) Vol. 3, No. 3, 771-779.
[16] Sarif, S., & Amran. (2024). Efektivitas Artificial Intelligence Text To Speech Dalam Meningkatkan Maharatul Qiraah. Naskhi Jurnal Kajian Pendidikan Dan Bahasa Arab Volume 6, No. 1, 1-8.
[17] Yoon, H., Um, S., Kim, C., & Kang, H.-G. (2023). Adversarial Learning Of Intermediate Acoustic Feature For End-To-End Lightweight Text-To-Speech. Interspeech, 3023-3027
[18] Rukmana, s. U., & subiyantoro. (2022). Variasi leksikal dan inovasi fonologis diaspora india: kajian sosiolinguistik pada generasi muda di kota medan. Jurnal bahasa dan sastra, 15–26