Text-To-Speech Bahasa Sunda Dialek Selatan Menggunakan Metode VITS

Andre Johann Jonnius[1], Yusra*[2], Muhammad Fikry[3], Novriyanto[4], Febi Yanto[5]
[1,*2,3,4,5]Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
e-mail: [1]12150111042@students.uin-suska.ac.id, *[2]yusra@uin-suska.ac.id, [3]muhammad.fikry@uin-suska.ac.id, [4]novriyanto@uin-suska.ac.id, [5]febiyanto@uinsuska.ac.id

Abstrak

Bahasa adalah alat untuk berkomunikasi dan juga identitas suatu wilayah. Indonesia memiliki keberagaman bahasa daerah, namun penggunaannya kian menurun dikarenakan berkurangnya penutur asli. Diperlukan upaya dalam pendukungan bahasa daerah, salah satunya ialah penerapan sistem TTS pada bahasa daerah. Penelitian ini bertujuan untuk mengembangkan sistem TTS bahasa daerah khususnya bahasa Sunda dialek Selatan menggunakan metode VITS (Variational Inference with Adversarial Learning for End-to-End Text-to-Speech). Penelitian ini menggunakan 450 data latih dan 50 data uji. Tahapan penelitian meliputi pengumpulan data, praproses data suara, pelatihan model, serta evaluasi hasil output dengan metode MOS (Mean Opinion Score). Pengujian MOS dilakukan kepada lima responden yang merupakan penutur asli Sunda dialek Selatan dan mendapatkan skor rata-rata sebesar 4,328. Hasil tersebut menunjukkan bahwa sistem mampu mengubah teks menjadi suara berbahasa Sunda dialek Selatan yang terdengar alami, jelas, dan mendekati penutur aslinya. Sistem ini belum melibatkan pemodelan ekspresi, jadi output masih mengikuti karakteristik dataset. Masih terdapat beberapa kalimat yang kurang jelas dan terdengar robotik, serta ketidakjelasan pelafalan pada fonem tertentu. Penelitian ini menyimpulkan bahwa pengembangan sistem Text-to-Speech bahasa Sunda dialek Selatan menggunakan metode VITS dapat menjadi kontribusi dalam mendukung penggunaan bahasa daerah dengan pemanfaatan teknologi tersebut.

***Kata kunci**: Text-to-Speech, Bahasa Sunda, Dialek Selatan, VITS, Mean Opinion Score.*

Abstract

Language is a tool for communication and an identity of a region. Indonesia has diverse local languages, but their use is declining due to fewer native speakers. Efforts are needed to support local languages, including implementing Text-to-Speech (TTS) systems. This study aims to develop a TTS system for the Southern dialect of Sundanese using the VITS method (Variational Inference with Adversarial Learning for End-to-End Text-to-Speech). The study used 450 training samples and 50 test samples. Research stages included data collection, voice preprocessing, model training, and output evaluation using the Mean Opinion Score (MOS) method. MOS testing with 5 native speakers yielded an average score of 4.328. Results indicate that the system can convert text into Southern Sundanese speech that sounds natural, clear, and close to native pronunciation. The system has not incorporated expression modeling, so outputs still follow dataset characteristics. Some sentences remain unclear or robotic, and certain phoneme pronunciations are ambiguous. This study concludes that developing a TTS system for the Southern Sundanese dialect using VITS can contribute to supporting local language use through technology.

***Keywords:** Text-to-Speech, Sundanese Language, Southern Dialect, VITS, Mean Opinion Score.*

1. PENDAHULUAN

Bahasa mempunyai peran yang sangat penting dalam kehidupan sehari-hari sebagai alat komunikasi dan ekspresi antar manusia sehari-hari. Selain sebagai alat untuk komunikasi, bahasa juga menjadi identitas daerah yang mencerminkan kebudayaan dan karakteristik wilayah tertentu [1]. Setiap wilayah memiliki karakteristik bahasanya tersendiri. Indonesia merupakan negara yang kaya akan keberagaman budaya, salah satunya ialah bahasa daerah. Dengan kekayaan budaya yang dimiliki, penting bagi budaya di negara ini terutama bahasa daerah untuk tetap dijaga. Bahasa daerah adalah komponen budaya yang sangat penting dan juga mempengaruhi penerima serta perilaku di setiap daerah [2].

Indonesia merupakan negara yang kaya akan keberagaman, dan salah satunya adalah keberagaman bahasa daerah. Terdapat banyak bahasa daerah di Indonesia, dan salah satunya adalah Bahasa Sunda yang merupakan fokus pada penelitian ini. Bahasa Sunda digunakan oleh masyarakat yang bertempat di wilayah Jawa Barat dan sekitarnya, termasuk Banten dan sebagian DKI Jakarta [3]. Bahasa Sunda memiliki berbagai

ragam dialek yang berkembang sesuai wilayah penuturnya, seperti dialek Sunda Tengah, Sunda Timur, dan Sunda Selatan [4]. Dialek yang digunakan pada penelitian ini adalah dialek Selatan, dikarenakan dialek Selatan ini merupakan dialek yang paling banyak digunakan masyarakat Sunda dan juga mencerminkan ragam bahasa Sunda yang relatif umum dan netral dalam pelafalan, sehingga sesuai digunakan sebagai objek kajian.

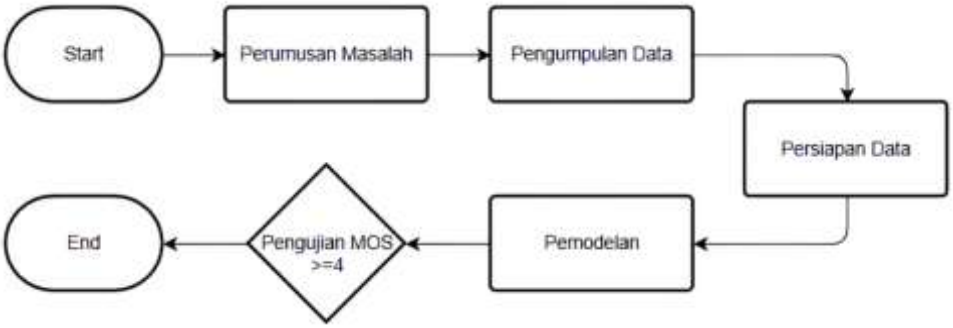
Meskipun memiliki peran penting sebagai alat komunikasi dan identitas budaya, penggunaan bahasa daerah kian menurun. Seiring perkembangan zaman bahasa daerah mengalami penurunan penggunaannya dikarenakan berkurangnya penutur asli dan meningkatnya penggunaan bahasa gaul serta bahasa media. Kondisi tersebut berpotensi mengancam keberlangsungan bahasa daerah hingga menyebabkan kepunahan [5]. Oleh karena itu, diperlukan upaya untuk mendukung penggunaan bahasa daerah. Salah satunya melalui penerapan teknologi modern seperti *Text-to-Speech* (TTS).

TTS atau *Text-to-Speech* merupakan teknologi yang dapat mengubah teks menjadi suara yang mana suara tersebut dapat menyerupai manusia [6]. Teknologi ini dapat dimanfaatkan dalam bentuk apapun seperti asisten virtual, pembaca layer untuk penyandang disabilitas, navigasi suara, dan lain sebagainya. TTS ini telah mengalami perkembangan yang signifikan. Beberapa penelitian telah dilakukan dalam pengembangan TTS, seperti *Tacotron2* [7], *FastSpeech2* [8], *VISinger* [9], serta pendekatan non-autoregressive berbasis variational inference [10]. Salah satu arsitektur TTS yang menghasilkan kualitas suara alami adalah VITS, yang menggabungkan *Variational Autoencoder* dengan *Adversarial Learning* untuk menghasilkan sintesis suara end-to-end [11]. Teknologi TTS terus dikembangkan untuk meningkatkan kejelasan dan naturalitas suara yang dihasilkan [12]. Peningkatan tersebut tercapai melalui model *end-to-end* yang lebih efisien dengan kemampuan mempelajari pola intonasi yang lebih konsisten. Kemajuan dalam TTS telah mencapai tingkat kealamian yang sangat baik, dengan kemampuan untuk menghasilkan ucapan yang mendekati bahasa manusia. Penelitian ini berfokus pada penerapan metode VITS untuk menghasilkan sistem TTS bahasa Sunda dialek Selatan, karena hingga saat ini penelitian yang mengembangkan model TTS bahasa daerah Sunda dengan metode VITS belum ditemukan secara spesifik.

Dari latar belakang tersebut, penelitian ini berfokus kepada pembuatan sistem TTS bahasa daerah Sunda dialek Selatan menggunakan metode VITS. Penelitian ini menggunakan dataset berupa 500 kalimat bahasa Sunda yang mana kalimat tersebut merupakan kalimat bahasa Sunda dialek Selatan itu sendiri yang nantinya akan dibagi menjadi 450 kalimat latih dan 50 kalimat uji. Dataset tersebut akan digunakan pada proses pelatihan untuk mendapatkan model yang nantinya akan digunakan pada tahap *inference.* Proses inference menghasilkan berupa output audio suara yang akan dievaluasi menggunakan Mean Opinion Score dengan ketentuan penilaian 1 hingga 5, yang mencerminkan tingkat kealamian dan kejelasan suara berdasarkan persepsi pendengar.

2. METODE PENELITIAN

Metode yang digunakan dalam penelitian TTS bahasa Sunda Dialek Selatan ini adalah VITS. Metode yang digunakan meliputi pemahaman tentang arsitektur model VITS, persiapan *dataset*, proses pelatihan model, tahap pengujian, serta evaluasi akhir menggunakan MOS. Bagian ini memberikan gambaran menyeluruh mengenai langkah-langkah yang akan ditempuh mulai dari persiapan awal hingga evaluasi akhir. Untuk alur pada penelitian ini dapat dilihat pada Gambar 1.

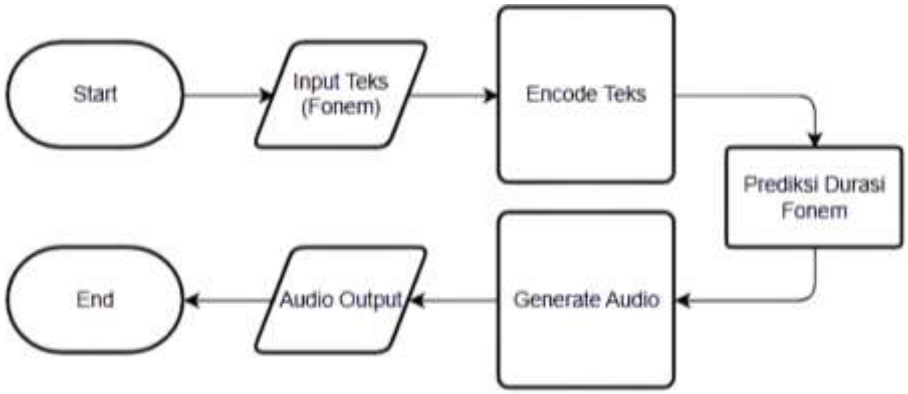


Gambar 1 Metodologi Penelitian

2.1 Arsitektur VITS (*Variational Inference with Adversarial Learning for End-to-End Text-to-Speech)*

VITS (*Variational Inference with Adversarial Learning for End-to-End Text-to-Speech*) merupakan pendekatan modern dalam pengembangan sistem TTS yang menggabungkan beberapa komponen penting

menjadi satu arsitektur yang tersusun. Tidak seperti metode lainnya yang membutuhkan model terpisah untuk *text encoder*, duration predictor, VITS menggabungkan seluruh proses tersebut dalam satu pemrosesan end-to-end. *Variational Inference with Adversarial Learning for end-to-end Text-to-Speech* (VITS) merupakan salah satu teknik yang digunakan untuk meningkatkan teknik sintesis ucapan [11]. Metode ini dapat meningkatkan kejelasan data dan membuatnya lebih efektif daripada metode TTS lainnya. VITS dapat menciptakan ucapan dari teks dengan menggabungkan audio tanpa menggunakan banyak gambar. VITS juga memiliki rentang frekuensi probabilistik yang memungkinkan sistem mempelajari distribusi kecepatan ucapan yang diharapkan dengan mengidentifikasi perbedaan signifikan dalam frekuensi dan kecepatan. Untuk alur kerja VITS dapat dilihat pada Gambar 2.



Gambar 2 Alur kerja VITS

Secara umum alur kerja model TTS berlangsung sebagai berikut:

1. Input
   Teks yang telah dikonversi menjadi fonem terlebih dahulu diproses oleh *text encoder* untuk menghasilkan representasi yang dapat diproses.
2. Mengukur durasi
   Komponen *stochastic duration predictor* menentukan estimasi panjang waktu penyebutan setiap fonem sehingga pola bicara dapat dicontoh secara lebih natural.
3. Proyeksi dan Transformasi
   Fonem yang telah direpresentasikan oleh *encoder* menghasilkan representasi laten setelah dilakukan tahap perubahan melalui tahapan projection dan flow model.
4. *Decoder*
   Representasi laten tersebut selanjutnya diterjemahkan kembali menjadi bentuk sinyal audio oleh *decoder*, menghasilkan gelombang suara yang menyerupai ucapan manusia. Dengan alur ini, model dapat mengubah teks menjadi suara yang terdengar lebih alami melalui teknik deep *learning*.

Pada penelitian ini terdapat tahap *inference*, pada dasarnya tahap ini mirip dengan tahap sebelumnya hanya saja tahap ini sedikit terbalik dengan tahap sebelumnya karena proses yang dilakukan adalah menguji model yang telah di latih. Dengan metode ini maka model yang dihasilkan dapat membuat suara yang dapat didengar lebih alami. Berikut merupakan komponen yang terdapat pada tahap *inference*:

1. *Text Encoder*: Ini merupakan salah satu komponen penting pada arsitektur VITS. Bagian ini mengubah masukan teks menjadi representasi linguistik yang dapat diproses oleh model. Representasi ini membantu model memahami struktur fonetik dan sintaksis teks. Jadi dengan itu proses ini dapat mengetahui pola dari data yang dimasukkan dan akan dipelajari di tahap selanjutnya.
2. *Posterior Encoder*: Mengambil fitur akustik dari suara asli sebagai referensi representasi laten selama pelatihan. Tahap ini sangat penting agar model memahami karakteristik suara penutur.
3. *Normalizing Flow*: Bagian ini berfungsi menyesuaikan distribusi laten agar lebih stabil sehingga proses pelatihan lebih konsisten.
4. *Decoder*: Ini adalah komponen yang bertugas untuk mengubah hasil dari proses sebelumnya ke dalam bentuk gelombang suara. Secara sederhana, proses ini adalah tahap akhir yang mengubah representasi laten menjadi bentuk suara yang utuh dan dapat digunakan.
5. Diskriminator: Ini adalah bagian yang bertugas untuk menyeleksi data yang akan dinilai, jadi pada tahap ini *discriminator* akan menilai kualitas suara yang dihasilkan. Komponen ini akan mencari perbedaan antara audio asli dan audio sintetis, dengan itu komponen ini dapat terus memperbaiki

hasil keluaran nya. Peran utama komponen ini adalah memastikan bahwa suara sintesis yang dihasilkan semakin sulit dibedakan dengan suara asli.

6. *Stochastic Duration Predictor* : Komponen ini merupakan modul yang bertugas untuk memprediksi durasi pada pengucapan atau pelafalan setiap fonem. Pemodelan durasi ini dilakukan secara stokastik sehingga dapat menghasilkan ritme yang terdengar lebih alami. Komponen ini sangat penting untuk menghasilkan suara yang dapat menyerupai pola bicara penutur aslinya.

Arsitektur VITS ini memungkinkan proses pelatihan yang lebih stabil, hasil suara yang lebih natural, serta efisiensi waktu karena seluruh proses berjalan pada satu model yang menyatu.

2.2 Persiapan Data

Pada tahap persiapan data yang perlu dilakukan adalah menyiapkan teks dan audio yang akan digunakan sebagai pasangan data pelatihan. Kualitas *dataset* sangat menentukan kualitas akhir dari model TTS yang dihasilkan. Oleh karena itu, tahap persiapan harus dilakukan secara teliti dan sistematis.

2.2.1 Pembuatan Teks

Teks pelatihan disusun dalam bahasa Sunda dialek Selatan dengan bantuan penutur asli untuk memperhatikan variasi kosakata, struktur kalimat khas daerah Selatan terpenuhi. Pembuatan teks dilakukan dengan memperhatikan keberagaman konteks agar model dapat menghasilkan intonasi alami, keterwakilan fonem yang lengkap sehingga model mudah memahami berbagai pola bunyi, kalimat tidak terlalu panjang maupun terlalu pendek untuk menjaga konsistensi perekaman, serta menghindari kata-kata yang sulit diucapkan atau jarang digunakan untuk mengurangi potensi kesalahan pelafalan.

2.2.2 Perekaman Suara

Audio direkam secara terkontrol dengan standar kualitas tertentu. Tahapan ini penting karena noise dan kualitas perekaman sangat berpengaruh pada hasil pelatihan. Proses perekaman dilakukan menggunakan mikrofon yang stabil, melakukan perekaman di ruangan dengan latar kebisingan serendah mungkin, menjaga konsistensi jarak antara mulut dan mikrofon, menjaga intonasi yang natural, jelas, dan tidak terburu-buru, menghasilkan file audio .wav dengan *sample rate* 22050 Hz, mono, dan format *PCM 16-bit* [11]. Setiap file dinamai berurutan sesuai daftar kalimat yang telah disiapkan.

2.2.3 Pelabelan dan Penyusunan Metadata

Data yang akan digunakan dimasukkan ke dalam file *metadata.csv* berfungsi menghubungkan teks dan audio. Format standar metadata pada VITS merujuk pada format LJ Speech [11] yaitu:

NamaFile|Penulisan|Pengucapan

Pada penelitian ini, kolom kedua dan ketiga tidak dibedakan karena tidak ada proses normalisasi tambahan. Jika terdapat seperti angka pada kolom kedua maka kolom ketiga dituliskan sebagaimana kata diucapkan. Metadata harus konsisten dalam format, penamaan, dan isi kolom agar proses pelatihan tidak mengalami kesalahan.

2.3 Pelatihan (*Training*)

Proses pelatihan dilakukan menggunakan Google Colab karena menyediakan GPU yang mendukung proses komputasi besar. Tahapan pelatihan mencakup pengimporan *dataset* dari Google Drive, menjalankan skrip pelatihan VITS yang telah disesuaikan, mengatur parameter seperti jumlah langkah (*steps)*, *learning* rate, dan batch size, memantau nilai *loss generator* dan *discriminator* untuk memastikan pelatihan berjalan stabil, melakukan *checkpointing* untuk menyimpan model setiap beberapa ribu langkah.

2.4 Pengujian (*Inference*)

Setelah proses pelatihan selesai, model diuji dengan memberikan input teks untuk menghasilkan suara. Pengujian ini dilakukan untuk memastikan model dapat membaca teks dengan natural, memiliki intonasi yang sesuai, tidak menghasilkan suara pecah atau robotik. Hasil inferensi kemudian disimpan sebagai audio untuk dianalisis lebih lanjut.

2.5 Evaluasi *Mean Opinion Score* (MOS)

MOS adalah skor rata-rata yang diambil dari opini beberapa subjek [13]. Setelah dilakukan tahap inference, audio yang disimpan akan dievaluasi menggunakan metode MOS untuk menilai kualitas suara berdasarkan penilaian subjektif pendengar. Tahapan MOS meliputi pengambilan sampel audio hasil inferensi, menyebarkan kuesioner kepada responden lalu responden menilai kualitas suara dari skala 1–5, di

mana 1 itu menandakan kualitas yang sangat buruk dan 5 menandakan kualitas sangat baik [14]. Nilai akhir dihitung dengan rata-rata dari seluruh penilaian. MOS digunakan sebagai acuan utama untuk menilai keberhasilan model TTS, dengan skor MOS yang digunakan pada penelitian ini lebih dari 4. Untuk keterangan penilaian MOS dapat dilihat pada Tabel 1.

Tabel 1 Skor MOS

MOS	Keterangan
5	Sangat Baik
4	Baik
3	Cukup
2	Buruk
1	Sangat Buruk

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pembuatan Teks

Pembuatan teks bahasa Sunda dialek Selatan dilakukan bersama 1 orang penutur asli untuk memastikan teks yang dibuat sesuai dengan sebagaimana teks dituliskan dan diucapkan. Jumlah teks yang akan digunakan ialah sebanyak 500 kalimat. Data yang telah dibuat akan divalidasi oleh pakar. Pakar yang akan memvalidasi ini adalah tetua adat atau budayawan Sunda. Kalimat yang telah disusun dapat dilihat pada Tabel 2.

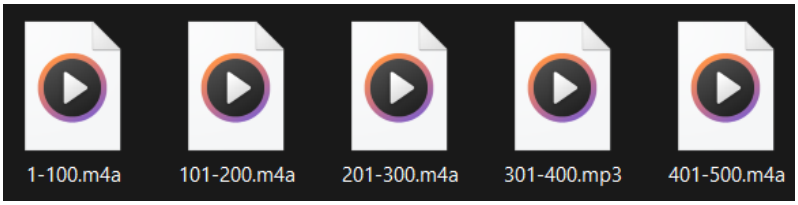
Tabel 2 Kalimat bahasa Sunda dialek Selatan

No	Kalimat bahasa Sunda	Kalimat bahasa Indonesia
1	Anjeun acan tuang ti isuk tadi	Kamu belum makan dari pagi tadi
2	Hayam eta keur ngubek taneuh	Ayam itu sedang mengais tanah
3	Kuring rek maca buku samemehna	Saya mau baca buku sebelumnya
4	Budak leutik maen di buruan	Anak kecil bermain di halaman
5	Nini keur nyulam di juru dapur	Nenek sedang menyulam di pojok dapur

Tabel 2 menampilkan 10 kalimat atau sebagian kecil dari 500 kalimat bahasa Sunda dialek Selatan yang mana kalimat ini disusun berdasarkan kalimat yang biasa diucap dan digunakan pada kehidupan sehari-hari dan juga telah disesuaikan dengan dialek Selatan itu sendiri.

3.2 Hasil Perekaman

Pada tahap kedua ini dilakukan perekaman audio yang berdurasi panjang berdasarkan teks yang telah disusun sebelumnya lalu dilakukan pemotongan per kalimatnya. Proses perekaman ini menggunakan pengisi suara wanita bernama Resti Nuragustiani berusia 21 tahun yang merupakan masyarakat asli Sukabumi yang mana daerah tersebut menggunakan bahasa daerah Sunda dialek Selatan. Narasumber akan membaca teks yang telah disusun sebelumnya dan direkam dengan total 500 teks kalimat. File audio yang telah direkam dapat dilihat pada Gambar 3.



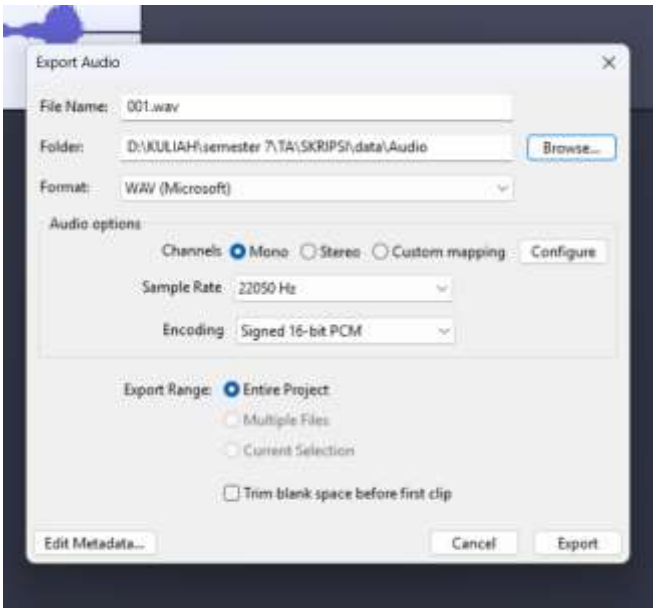
Gambar 3 *File* audio

Gambar 3 merupakan tampilan dari *file* suara yang telah direkam. Terdapat 500 kalimat yang direkam dengan 5 sesi dan setiap sesi merekam 100 kalimat. Suara yang direkam ini akan dipotong nantinya per kalimat dan dibersihkan pada tahap selanjutnya

3.3 Hasil Pembersihan dan Pemotongan Audio

Pembersihan data audio dilakukan secara manual menggunakan Audacity. Tahap ini meliputi

pemotongan jeda panjang di awal, tengah, dan akhir rekaman. Selanjutnya menghilangkan *noise* pada rekaman. Setelah melakukan pembersihan dan pemotongan, audio diekspor dengan ketentuan: format .wav, mono, 16-bit PCM, dengan *sample rate* 22050 Hz [11]. Konfigurasi ini dinilai ideal untuk kebutuhan pelatihan model TTS karena menghasilkan kualitas suara yang cukup baik tanpa ukuran file yang terlalu besar. Konfigurasi pada pembersihan audio dapat dilihat pada Gambar 4.



Gambar 4 Pembersihan audio

3.4 Hasil Persiapan Dataset Pelatihan

Tahap ini sangat penting pada penelitian TTS *bahasa Sunda Dialek Selatan* ini. Penyiapan *dataset* pada penelitian ini mengacu kepada format LJ Speech yang terdiri dari dua komponen yaitu *metadata.csv* dan audio WAV [11]. Ini bertujuan untuk memastikan data yang akan dikerjakan tetap terstruktur. Berikut merupakan tahapan pada penyiapan *dataset* untuk pelatihan:

a) Pembuatan *metadata.csv*

Metadata merupakan kumpulan informasi dari rekaman suara yang telah disiapkan, rekaman tersebut berbentuk *file* cvs yang mendeskripsikan rekaman tersebut. *Metadata.csv* tersebut memiliki 3 kolom yang setiap kolom memberikan makna yang berbeda. Kolom pertama menjelaskan tentang nama *file*, tanpa diawali dengan 'wavs' dan/atau diakhiri '.wav'. Untuk kolom kedua menjelaskan tentang teks sebagaimana tertulis, dan kolom ketiga menjelaskan teks sebagaimana dibacakan

b) Direktori WAV

Pada direktori ini terdapat file audio wav yang mana telah sesuai dengan *metadata.csv* tersebut

c) Pembagian *dataset*

Setelah *dataset* tersebut tersusun, *dataset* tersebut dibagi dua menjadi data uji dan data latih. Data latih terdiri dari 450 data dan data uji terdiri dari 50 data. Kalimat yang telah disesuaikan penulisannya dengan *metadata.csv* dapat dilihat pada Tabel 3.

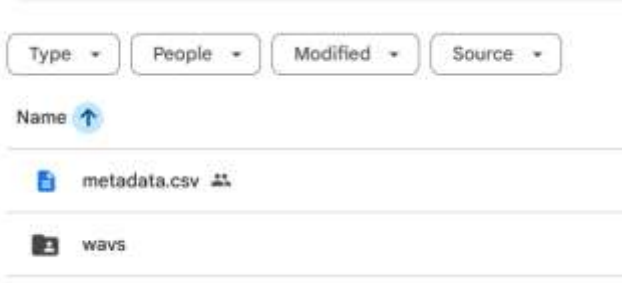
Tabel 3 *Metadata.csv*

1|Anjeun acan tuang ti isuk tadi|anjeun acan tuang ti isuk tadi
2|Hayam eta keur ngubek taneuh|hayam eta keur ngubek taneuh
3|Kuring rek maca buku samemehna|kuring rek maca buku samemehna
4|Budak leutik maen di buruan|budak leutik maen di buruan
5|Nini keur nyulam di juru dapur|nini keur nyulam di juru dapur

Tabel 3 menampilkan sebagian kecil dari total 450 kalimat yang akan menjadi *dataset* pelatihan, dan juga kalimat ini merupakan isi dari *metadata.csv* yang mana penulisan dan juga pengucapannya telah disesuaikan.

3.5 Hasil Pelatihan (*Training*)

Tahap selanjutnya ialah pelatihan, yaitu dengan melatih data untuk menghasilkan model TTS. *Dataset* yang telah dikumpulkan sebelumnya akan dimasukkan ke dalam Google Drive agar mudah diakses oleh Google Colab. *Dataset* ini diimplementasikan menggunakan bahasa pemrograman Python. Data yang digunakan pada proses pelatihan ini sebanyak 450 data latih yang akan diproses menjadi data model dan proses tersebut berlangsung selama 5 hari. Struktur *dataset* pelatihan yang digunakan pada tahap ini dapat dilihat pada Gambar 4. *Dataset* pelatihan yang akan digunakan pada tahap ini meliputi *metadata.csv* dan folder wav yang mana terdapat *file* audio di dalamnya.



Gambar 4 *Dataset* Pelatihan

Selanjutnya dapat dilihat pada Gambar 5 yang merupakan proses pelatihan model dengan menampilkan log selama *training* berlangsung. Pada gambar tersebut terlihat tampilan seperti step, eval, *loss* values, dan lainnya. Proses ini akan menghasilkan *output* berupa : *best_model.pth* dan config.json.

Gambar 5 Proses pelatihan model

3.6 Hasil *Inference*

Inference merupakan proses penggunaan model yang telah dilatih [15]. Setelah mendapatkan model maka akan dilakukan tahap selanjutnya, file model kemudian digunakan dalam proses inferensi di Google Colab untuk menghasilkan 50 *output* audio berdasarkan data uji. Audio yang dihasilkan berdasarkan 50 kalimat uji yang telah disiapkan sebelumnya. *Output* disimpan secara otomatis dalam format wav. Proses *inference* dapat dilihat pada Gambar 6.

Gambar 6 Proses *inference* model

Gambar 6 menampilkan proses *inference* atau proses dari pengubahan teks menjadi suara. Tahap ini dapat dilakukan dengan cara menginput manual seperti pada Gambar 6 dan dapat dilakukan otomatis dengan mengubah banyak kalimat dalam satu pemrosesan.

Berdasarkan hasil *inference* yang dilakukan, model TTS yang telah dilatih mampu menghasilkan dari teks bahasa Sunda dialek Selatan. Beberapa percobaan telah dilakukan pada tahap *inference* ini seperti mencoba kalimat yang tidak dimiliki pada data latih sebelumnya untuk memastikan kualitas model dan model dapat menghasilkan suara sebagai mana suara penutur asli yang digunakan pada dataset pelatihan. Meskipun model yang dihasilkan dapat meniru suara penutur aslinya namun terkadang terdapat beberapa kata yang terdengar kurang jelas dan robotik, dan juga terdapat beberapa fonem yang pengucapannya kurang jelas. Hal ini dipengaruhi oleh kualitas suara perekaman dan keterbatasan jumlah fonem yang belum optimal, sehingga model masih memiliki kekurangan pada beberapa kalimat yang mengakibatkan kalimat tersebut kurang jelas pengucapannya.

3.7 Hasil Pengujian MOS

Evaluasi dilakukan menggunakan MOS oleh lima penutur asli Sunda dialek Selatan, masing-masing menilai 50 audio hasil TTS. Skor MOS berada pada rentang 1–5, dan hasil rata-rata evaluasi adalah 4.328**.** Hasil ini menunjukkan kualitas sintesis suara yang sangat baik dan hampir menyerupai suara alami. Secara keseluruhan, sistem TTS bahasa Sunda dialek Selatan berhasil menghasilkan suara mudah dipahami. Hasil pengujian MOS yang telah dilakukan dapat dilihat pada Tabel 4.

Tabel 4 Pengujian MOS

No	Kalimat	Responden					Total	Rata-rata
		1	2	3	4	5		
1	Kuring bakal ngadamel acara ngabersihan pantai	4	3	4	5	5	21	4,2
2	Hujan ngagebray ti peuting tadi	4	4	5	5	5	23	4,6
3	Kumaha damang di dinya	4	4	4	4	3	19	3,8
4	Urang bakal ngajak ngadamel acara donor darah	5	4	4	5	5	23	4,6
5	Kuring bade ngawangun lembaga di desa	4	4	4	5	4	21	4,2
6	Urang bakal ngatur acara lomba kebersihan lingkungan	5	4	5	5	4	23	4,6
7	Kuring bakal ngadamel revitalisasi taman kota	5	4	3	3	4	19	3,8
8	Urang bakal ngadamel acara bagi pemuda	5	3	4	5	5	22	4,4
9	Kuring hoyong ngadamel aplikasi belajar bahasa sunda	3	4	4	4	5	20	4
10	Sim kuring bade ngumbara ka kulon	4	5	5	4	3	21	4,2
11	Tong lalumpatan di jero imah	5	5	5	5	5	25	5
12	Urang bakal ngadamel program peningkatan kualitas pendidikan	5	4	5	5	4	23	4,6
13	Kuring bakal ngawangun rumah sakit di daerah terpencil	5	3	3	4	4	19	3,8
14	Leungeun anjeun tiis kawas es	3	4	3	3	3	16	3,2
15	Anjeun kudu datang isuk keneh	5	4	5	4	4	22	4,4
16	Beunang lauk loba keneh	5	5	3	4	4	21	4,2

No	Kalimat	Responden					Total	Rata-rata
		1	2	3	4	5		
17	Ucing hideung ngalayang di imah	5	5	5	5	5	25	5
18	Aya pameran Sae Pisan di dinya	5	5	4	4	3	21	4,2
19	Angkat ka tempat olahraga di desa	5	5	5	5	4	24	4,8
20	Urang ngadamel acara seni tradisional	5	5	5	5	4	24	4,8
21	Mahasiswa keur diajar masak di asrama	5	5	5	5	4	24	4,8
22	Ubar na kudu diinum isuk	3	4	3	4	4	18	3,6
23	Kuring bakal ngadamel program peningkatan kualitas kesehatan	5	4	4	4	4	21	4,2
24	Urang bakal ngatur acara penggalangan dana untuk pendidikan	5	3	5	5	5	23	4,6
25	Kuring hoyong ngajak ngadamel taman untuk berolahraga	5	3	5	5	4	22	4,4
26	Urang bakal ngadamel acara diskusi budaya di kampus	5	5	5	5	5	25	5
27	Manuk bodas hiber luhur pisan	3	4	3	4	3	17	3,4
28	Urang bakal ngadamel acara sosial di lingkungan masyarakat	5	5	5	5	4	24	4,8
29	Kuring bakal ngadamel program pemberdayaan masyarakat	5	5	5	5	4	24	4,8
30	Tong hilap tutup panto imah	3	4	4	4	3	18	3,6
31	Kuring hoyong ngajak ngadamel kursus memasak di kampus	5	4	5	5	5	24	4,8
32	Para petani panen padi taun ieu	4	4	4	4	3	19	3,8
33	Kuring bade ngawangun sekolah untuk anak-anak berprestasi	5	3	4	5	4	21	4,2
34	Nabung teh akar tina kakayaan	3	5	5	5	5	23	4,6
35	Melak tangkal pikeun nyalametkeun hawa sareng lingkungan urang	4	4	4	5	5	22	4,4
36	Kaseueuran masyarakat di dieu nyaeta mahasiswa	4	4	4	4	3	19	3,8
37	Ulah miceun runtah ka walungan	4	4	4	3	4	19	3,8
38	Cobian atuh leumpang ka alun-alun kota	5	5	5	4	4	23	4,6
39	Warga desa urang salawasna ngahiji	4	4	5	5	4	22	4,4
40	Ulah miceun runtah ka walungan	4	5	4	3	4	20	4
41	Jembatan na rusak sataun katukang	5	4	5	5	5	24	4,8
42	Urang bakal ngatur pelatihan keterampilan bagi ibu rumah tangga	5	4	3	5	4	21	4,2
43	Di dinya aya tangkal mangga	4	4	4	5	4	21	4,2
44	Kuring bade ngadamel acara bazar murah di kota	5	5	5	5	5	25	5
45	Ieu sapedah dibeuli pikeun olahraga	4	4	5	5	4	22	4,4
46	Urang bakal ngatur acara karir untuk siswa	4	4	5	4	3	20	4
47	Kuring bade ngawangun tempat pelatihan untuk petani muda	5	4	5	5	4	23	4,6
48	Urang bakal ngajak ngadamel acara pertemuan antar komunitas	5	4	5	5	4	23	4,6
49	Kuring hoyong ngawangun pasar tradisional di tengah kota	4	4	4	5	4	21	4,2
50	Urang bakal ngadamel acara sosial di lingkungan sekolah	4	4	5	5	4	22	4,4
TOTAL								216,4
SKOR MOS								4,328

Nilai MOS sebesar 4,328 menunjukkan bahwa kualitas suara sangat bagus dan dapat dipahami oleh

para responden. Meskipun demikian, hasil penilaian juga menunjukkan bahwa terdapat beberapa kalimat yang mendapatkan skor rendah dibandingkan kalimat lainnya, seperti kalimat pada nomor 3 dan juga 7. Responden juga menuliskan catatan khusus terkait nomor tersebut yakni pengucapan yang kurang jelas pada fonem /ny/. Selanjutnya terdapat beberapa kalimat lainnya yang juga kurang jelas pengucapannya seperti kata "teuing" dan "hoyong". Pada kalimat yang diuji juga terdapat variasi kejelasan pada fonem /h/ dan /u/ yang di mana fonem tersebut untuk di beberapa kalimat terucap dengan jelas namun di beberapa kalimat lainnya terdengar kurang jelas. Beberapa responden juga mencatat bahwa terdapat kalimat yang pengucapannya sedikit cepat. Hasil evaluasi yang didapati ini menunjukkan bahwa meskipun kualitas suara yang secara menyeluruh itu baik namun masih terdapat beberapa variasi kejelasan fonem dan kecepatan pengucapan kalimat yang dipengaruhi oleh dataset pada pelatihan

4. KESIMPULAN

Berdasarkan hasil penelitian TTS bahasa Sunda dialek Selatan yang berbasis VITS, penelitian ini berhasil menghasilkan sistem TTS untuk bahasa Sunda dengan *dataset* yang digunakan sebanyak 450 kalimat pelatihan bahasa Sunda dan 50 kalimat uji yang direkam oleh penutur asli wanita dewasa yang merupakan masyarakat asli Sunda. Model yang dihasilkan mampu menghasilkan suara sintetis yang menyerupai suara penutur aslinya. Evaluasi dilakukan menggunakan metode *Mean Opinion Score* (MOS) dengan melibatkan lima responden yang merupakan penutur asli Sunda dan memperoleh nilai rata-rata 4,328 yang menunjukan bahwa sistem ini sukses menghasilkan suara yang jelas dan natural dan mencapai target pada penelitian ini yang mana penelitian ini telah menetapkan skor MOS yaitu di atas 4. Keterbatasan penelitian ini terletak pada *dataset* yang hanya menggunakan suara dari satu orang penutur, yaitu wanita dewasa penutur asli Sunda dialek Selatan, sehingga model yang dihasilkan hanya menampilkan suara dari satu jenis penutur saja. Oleh karena itu, pada penelitian selanjutnya diharapkan untuk menambah jumlah penutur serta variasi pada data pelatihan agar dapat menghasilkan sistem TTS dengan kualitas yang lebih baik.

DAFTAR PUSTAKA

[1] N. Anggini, N. Y. Afifah, and E. Syaputra, "Pengaruh bahasa gaul (slang) terhadap bahasa Indonesia pada generasi muda," *Concept J. Soc. Humanit. Educ.*, vol. 1, no. 4, pp. 39–48, 2022.

[2] T. F. Pandaleke, V. I. . F. Koagouw, and G. J. Waleleng, "the Role of Community Social Communication in Preserving thePasan Regional Languages in Rasi Village Ratahan Sub-DistrictSoutheast Minahasa Regency," *Acta Diurna Komun.*, vol. 2, 2020.

[3] W. L. Rachma, "Bahasa Sunda Sebagai Identitas Budaya Mahasiswa Etnis Sunda di Universitas Serang Raya," *Pros. Semin. Nas. Komunikasi, Adm. Negara dan Huk.*, vol. 1, no. 1, pp. 283–292, 2023, doi: 10.30656/senaskah.v1i1.178.

[4] P. Asteka, "Ragam Dialek Sunda Majalengka Dalam Interaksi Komunikasi Pada Mahasiswa Bahasa Dan Sastra Indonesia Universitas Majalengka," *Konf. Nas. Bhs. Dan Sastra V*, vol. 5, no. 1, pp. 209–215, 2019.

[5] M. S. Hidayat and M. F. Hibban, "Analysis of Mudjia Rahardjo's Views: Language Philosophy and Efforts to Preserve Regional Languages," *J. Penelit. Medan Agama*, vol. 15, no. 1, p. 1, 2024, doi: 10.58836/jpma.v15i1.17521.

[6] M. Chen *et al.*, "MultiSpeech: Multi-speaker text to speech with transformer," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 4024–4028, 2020, doi: 10.21437/Interspeech.2020-3139.

[7] Y. Li, D. H. Qin, and J. B. Zhang, "Speech Synthesis Method Based on Tacotron2," in *2021 13th International Conference on Advanced Computational Intelligence, ICACI 2021*, 2021, pp. 94–99. doi: 10.1109/ICACI52617.2021.9435882.

[8] Y. Ren *et al.*, "Fastspeech 2: Fast and High-Quality End-To-End Text To Speech," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, pp. 1–15, 2021.

[9] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational Inference With Adversarial Learning for End-To-End Singing Voice Synthesis," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 7237–7241, 2022, doi: 10.1109/ICASSP43922.2022.9747664.

[10] Y. Lee, J. Shin, and K. Jung, "Bidirectional Variational Inference for Non-Autoregressive Text-To-Speech," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, pp. 1–19, 2021.

[11] 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.
[12] W. Zhao and Z. Yang, "An Emotion Speech Synthesis Method Based on VITS," *Appl. Sci.*, vol. 13, no. 4, pp. 1–12, 2023, doi: 10.3390/app13042225.
[13] Y. Gao, X. Min, Y. Zhu, J. Li, X. P. Zhang, and G. Zhai, "Image Quality Assessment: From Mean Opinion Score to Opinion Score Distribution," *MM 2022 - Proc. 30th ACM Int. Conf. Multimed.*, no. October 2022, pp. 997–1005, 2022, doi: 10.1145/3503161.3547872.
[14] Y. Wang *et al.*, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, pp. 4242–4246, 2022, doi: 10.21437/Interspeech.2022-48.
[15] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 9483–9492, 2019.