Text-To-Speech Bahasa Ocu Dialek Siak Hulu Menggunakan Metode VITS

Putri Juniarti[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]12150121672@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@uin-suska.ac.id

Abstrak

Bahasa Ocu adalah bahasa daerah yang digunakan oleh masyarakat Kabupaten Kampar, Provinsi Riau. Meskipun Bahasa Indonesia telah menjadi bahasa nasional, keberadaan bahasa daerah tetap dihargai di Indonesia. Penelitian ini bertujuan untuk mengimplementasikan teknologi Text-to-Speech (TTS) dalam Bahasa Ocu dialek Desa Pangkalan Baru, Kecamatan Siak Hulu, menggunakan metode Variational Inference Text to Speech (VITS). Penelitian ini menggunakan dataset yang terdiri dari rekaman 500 kalimat Bahasa Ocu dan teks 500 kalimat Bahasa Ocu. Evaluasi kinerja model dilakukan dengan Mean Opinion Score (MOS). MOS dilakukan dengan meminta 5 orang penutur asli memberikan skor 1 sampai 5 kepada setiap file rekaman yang telah dihasilkan model. Seluruh skor dijumlahkan dan dicari rata-ratanya untuk mendapatkan skor akhir. Hasil dari implementasi TTS menunjukkan performa yang sangat baik, dengan skor akhir MOS sebesar 4,508, yang menandakan kualitas suara sangat mirip dengan pengucapan manusia. Terdapat beberapa catatan dari evaluator seperti huruf yang tertukar atau huruf tidak terdengar jelas serta suara yang dihasilkan terdengar kaku. Penelitian ini diharapkan dapat menjadi referensi untuk pengembangan teknologi TTS dalam bahasa daerah lainnya, serta membantu dalam pengajaran Bahasa Ocu untuk generasi mendatang.

Kata kunci: *Bahasa Ocu, Text-to-Speech, Variational Inference Text-to-Speech*

Abstract

Ocu Language is a regional language spoken by the community of Kampar Regency, Riau Province. Although Indonesian has been established as the national language, the existence of regional languages continues to be respected in Indonesia. This study aims to implement Text-to-Speech (TTS) technology for the Ocu language, specifically the Pangkalan Baru Village dialect in Siak Hulu District, using the Variational Inference Text-to-Speech (VITS) method. This research utilizes a dataset consisting of recordings of 500 Ocu language sentences and an additional 500 Ocu language sentences text. The performance of the model is evaluated using the Mean Opinion Score (MOS). The MOS evaluation is conducted by asking five native speakers to assign scores ranging from 1 to 5 to each audio file generated by the model. All scores are then summed and averaged to obtain the final MOS score. The results of the TTS implementation demonstrate very good performance, with a final MOS score of 4.508, indicating that the synthesized speech is highly similar to human pronunciation. However, several issues were noted by the evaluators, such as swapped letters, unclear pronunciation of certain sounds, and slightly rigid speech output. This study is expected to serve as a reference for the development of TTS technology for other regional languages and to support the teaching and preservation of the Ocu language for future generations.

Keywords: *Ocu Language, Text-to-Speech, Variational Inference Text-to-speech*

1. PENDAHULUAN

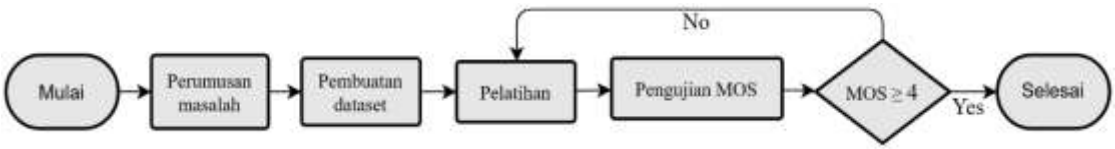
Bahasa merupakan sistem lambang bunyi yang bersifat arbitrer dan digunakan oleh anggota masyarakat untuk berkomunikasi, bekerja sama, serta mengidentifikasi diri dalam kehidupan sosial. Bahasa berperan sebagai alat komunikasi utama yang memungkinkan manusia berinteraksi secara efektif dalam kehidupan bermasyarakat[1]. Bahasa sangat terikat dengan manusia karena segala kegiatan manusia tidak pernah lepas dari bahasa[2]. Berdasarkan fungsinya, bahasa dapat diklasifikasikan menjadi bahasa internasional, bahasa nasional, bahasa daerah, dan bahasa asing dalam komunikasi antar negara di dunia[3]. Keberadaan bahasa sangat penting karena menjadi fondasi utama dalam membangun hubungan sosial dan mempertahankan identitas suatu kelompok masyarakat.

Bahasa daerah merupakan salah satu bentuk bahasa yang digunakan oleh komunitas tertentu dalam wilayah geografis yang terbatas dan sering berfungsi sebagai bahasa ibu. Meskipun Bahasa Indonesia telah ditetapkan sebagai bahasa nasional, bahasa daerah tetap memiliki peran penting sebagai penanda identitas budaya dan sosial masyarakat. Indonesia memiliki keragaman bahasa dan budaya yang sangat kaya. Terdapat lebih dari 700 bahasa daerah yang ada di Indonesia[4]. Salah satunya adalah Bahasa Ocu yang digunakan oleh masyarakat Kabupaten Kampar, Provinsi Riau. Bahasa ini kerap dipandang sebagai turunan dari Bahasa Melayu maupun Bahasa Minangkabau, yang dipengaruhi oleh proses akulturasi budaya serta mobilitas penduduk dari berbagai daerah[5]. Di wilayah Kecamatan Siak Hulu, khususnya Desa Pangkalan Baru, Bahasa Ocu yang digunakan memiliki kosakata yang mendekati Bahasa Minangkabau. Hal ini terlihat dari kosakata tertentu yang berbeda dengan varian Bahasa Ocu di wilayah lain, seperti penggunaan kata “balik” untuk menyatakan makna “pulang”. Perbedaan tersebut dipengaruhi oleh faktor geografis seperti pemisahan antar desa oleh sungai dan kawasan hutan, serta keterbatasan akses transportasi pada masa lalu[6]. Hingga saat ini, penelitian ilmiah yang membahas Bahasa Ocu dialek Siak Hulu masih sangat terbatas. Sebagian besar penelitian bahasa daerah lebih banyak berfokus pada dialek atau bahasa daerah yang telah memiliki dokumentasi linguistik yang lebih lengkap sehingga dialek Desa Pangkalan Baru Kecamatan Siak Hulu belum banyak mendapatkan perhatian dalam pengembangan maupun penerapan teknologi berbasis bahasa, khususnya pada bidang sintesis suara dan pengolahan bahasa alami. Kondisi ini menyebabkan keterbatasan sumber data serta referensi ilmiah yang dapat digunakan sebagai acuan[7] dalam pengembangan sistem berbasis Bahasa Ocu dialek Desa Pangkalan Baru Kecamatan Siak Hulu.

Perkembangan teknologi pemrosesan bahasa alami memungkinkan bahasa lisan direpresentasikan dalam bentuk digital melalui sistem Text-to-Speech (TTS), yaitu teknologi yang mengubah teks menjadi suara yang menyerupai ucapan manusia[8]. TTS umumnya bekerja melalui tahapan pembentukan representasi laten, pembuatan mel-spectrogram, dan konversi menjadi sinyal audio. Teknologi ini telah banyak dimanfaatkan dalam berbagai bidang, seperti asisten virtual, fitur aksesibilitas bagi penyandang disabilitas, dan aplikasi penerjemahan otomatis[9]. Penelitian mengenai sistem Text-to-Speech (TTS) telah banyak dilakukan dengan berbagai pendekatan dan metode, baik untuk bahasa internasional maupun bahasa daerah. Beberapa penelitian sebelumnya menunjukkan bahwa teknologi TTS mampu menghasilkan suara sintetis yang semakin natural seiring dengan perkembangan model berbasis *deep learning*. Berbagai arsitektur telah diterapkan, seperti pada penelitian *Adaspeech* dan *Tacotron*[10][11]. Meskipun demikian, sebagian besar penelitian TTS masih berfokus pada bahasa dengan sumber data yang melimpah dan dokumentasi linguistik yang lengkap, sementara penerapan TTS pada bahasa atau dialek daerah dengan sumber data terbatas masih relatif sedikit dilakukan. Salah satu model dalam pengembangan TTS adalah Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS), yang mampu menghasilkan suara lebih alami karena menggabungkan beberapa tahapan pemodelan ke dalam satu arsitektur end-to-end. VITS juga memanfaatkan Variational Autoencoder (VAE) sehingga mampu menangkap variasi intonasi dan karakteristik dialek secara lebih akurat[12].

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada pengembangan TTS Bahasa Ocu, khususnya dialek Desa Pangkalan Baru, Kecamatan Siak Hulu, Kabupaten Kampar, menggunakan VITS. Dialek ini dipilih karena masih minimnya penelitian ilmiah yang secara khusus membahas karakteristik dan penerapan dialek tersebut, sehingga penelitian ini diharapkan dapat memberikan kontribusi awal dalam upaya dokumentasi serta pengembangan teknologi berbasis Bahasa Ocu dialek Desa Pangkalan Baru Kecamatan Siak Hulu. Penelitian ini menggunakan 500 kalimat sebagai data latih yang direkam dari penutur asli wanita berusia 61 tahun. Tujuan penelitian ini adalah untuk mengembangkan sistem TTS yang mampu menghasilkan suara Bahasa Ocu yang alami dan akurat, serta menguji performa VITS dalam merepresentasikan karakteristik dialek lokal. Diharapkan penelitian ini dapat menjadi kontribusi bagi pengembangan teknologi bahasa daerah serta menjadi bahan pembelajaran bahasa daerah**.**

2. METODE PENELITIAN



Gambar 1 Flowchart Metode Penelitian

Metodologi penelitian ini disusun untuk membangun dan mengevaluasi sistem TTS Bahasa Ocu dialek Siak Hulu secara sistematis. Metodologi ini mencakup beberapa tahapan utama, yaitu pembuatan *dataset*, pelatihan model, serta evaluasi kualitas suara yang dihasilkan. Pada tahap pembuatan *dataset*, dilakukan penyusunan data teks dan audio sebagai bahan pelatihan model. Selanjutnya, *dataset* tersebut digunakan dalam proses pelatihan model TTS berbasis VITS untuk menghasilkan model yang mampu merepresentasikan karakteristik dialek yang diteliti. Tahap akhir penelitian adalah evaluasi menggunakan metode MOS untuk menilai kualitas keluaran suara berdasarkan penilaian penutur asli. Uraian lebih rinci mengenai setiap tahapan penelitian akan dijelaskan pada bagian-bagian berikutnya.

2.1. Pembuatan *dataset*

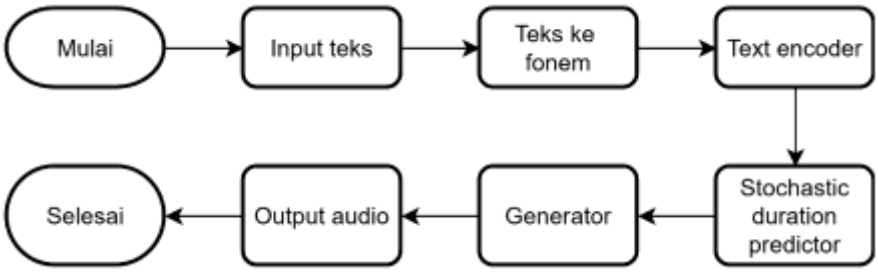
Langkah pertama dalam membuat *dataset* adalah membuat 500 kalimat dalam Bahasa Ocu dialek Siak Hulu. Pembuatan kalimat dilakukan bersama penutur asli yang kemudian akan dilakukan validasi data oleh pemuka adat setempat. 500 kalimat tersebut akan dibagi dalam 2 *file* yaitu metadata dan testdata. *Dataset* terdiri dari 2 *file* CSV dan 1 *folder audio*. *File* .csv yaitu metadata.csv dan testdata.csv masing-masing berisi 450 dan 50 kalimat Bahasa Ocu dialek Siak Hulu yang telah dibuat sebelumnya. *File* metadata.csv akan menjadi data latih dan testdata.csv menjadi data uji. *Folder audio* terdiri dari 500 *file audio* yang telah direkam oleh penutur asli. *Folder audio* diberi nama wavs. Terdapat beberapa hal yang perlu diperhatikan dalam membuat *dataset* yaitu:

1. Kosakata yang digunakan bervariasi.
2. Setiap kalimat terdiri dari 4 sampai 8 kata.
3. Metadata dan testdata masing-masing memiliki 3 kolom yang dipisahkan menggunakan karakter |. Kolom pertama berisi nama *file audio* dari kalimat tersebut, kolom kedua berisi kalimat sebagaimana ditulis, dan kolom ketiga berisi kalimat sebagaimana dibacakan. Jika tidak ada angka pada kalimat, maka kolom kedua dan ketiga ditulis sama persis. *Format* ini mengikuti *format* LJ Speech.
4. Perekaman dilakukan menggunakan peralatan yang sama untuk menjaga konsistensi suara yang dihasilkan.
5. Setiap kalimat dibacakan dengan jelas.
6. *File audio* menggunakan *format* wav 16-bit PCM dan *sample rate* 22050 Hz sesuai dengan format LJ Speech[12].

2.2. Pelatihan

Penelitian ini menggunakan model VITS sebagai pendekatan utama dalam proses pelatihan sistem TTS. Tahap pelatihan dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan platform Google Colab sebagai lingkungan komputasi. *Dataset* yang digunakan terdiri dari 450 data latih berupa pasangan teks dan audio, yang disimpan pada Google Drive untuk memudahkan akses selama proses pelatihan berlangsung. *Dataset* tersebut digunakan untuk melatih model TTS berbasis VITS hingga diperoleh model yang mampu menghasilkan suara Bahasa Ocu sesuai dengan karakteristik dialek Siak Hulu yang diteliti.

Setelah proses pelatihan selesai, model yang dihasilkan selanjutnya diuji pada tahap inferensi. Inferensi adalah proses di mana sistem TTS mengubah teks menjadi suara yang dapat didengar[13]. Tahap inferensi bertujuan untuk menghasilkan keluaran berupa *file* audio dari 50 kalimat data uji yang terdapat dalam *file testdata*. Proses inferensi dilakukan sesuai dengan alur pada flowchart inferensi VITS, dimulai dari penerimaan input berupa teks. Teks masukan kemudian melalui tahap normalisasi dan konversi ke dalam bentuk fonem sebagai representasi bunyi bahasa. Representasi fonem tersebut diproses oleh *text encoder* untuk menghasilkan representasi laten linguistik, kemudian dilanjutkan dengan *stochastic duration predictor* untuk menentukan durasi setiap fonem agar irama suara terdengar lebih natural. Selanjutnya, *decoder* atau *generator* mengubah representasi laten tersebut menjadi sinyal suara. Hasil akhir dari proses inferensi berupa *file* audio dalam format *.wav* yang dihasilkan dari setiap kalimat pada data uji. Gambar 2 memperlihatkan alur kerja utama VITS dalam menghasilkan suara.



Gambar 2 Alur VITS

Arsitektur VITS terdiri dari beberapa komponen utama yaitu:

1. Teks input**,** yang kemudian diubah menjadi fonem sebagai representasi bunyi bahasa.
2. *Text encoder***,** yang memproses urutan fonem untuk menghasilkan representasi laten yang menggambarkan informasi linguistik.
3. *Posterior encoder***,** yang mempelajari karakteristik suara dari data *audio* selama proses pelatihan.
4. *Stochastic duration predictor***,** yang memprediksi durasi setiap fonem secara probabilistik sehingga durasi dan ritme suara lebih natural.
5. *Decoder* atau *generator***,** yang mengubah representasi laten menjadi sinyal suara.
6. *Discriminator***,** yang digunakan dalam proses pelatihan untuk membedakan suara asli dan suara hasil sintesis guna meningkatkan kualitas suara yang dihasilkan.

2.3. Evaluasi MOS

Evaluasi performa model dilakukan menggunakan metode *Mean Opinion Score* (MOS) yang berfungsi untuk menilai kualitas suara yang dihasilkan oleh model. MOS telah digunakan dalam banyak penelitian TTS sebagai metode evaluasi subjektif[14]. Sebanyak 5 penutur asli Bahasa Ocu dari Desa Pangkalan Baru dilibatkan sebagai responden. Responden diminta untuk mendengarkan 50 *file output audio* model TTS. MOS menggunakan skala 1-5 untuk penilaiannya di mana 5 adalah nilai tertinggi dan 1 adalah nilai terendah[14]. Tabel 1 menunjukkan skor beserta keterangannya.

Tabel 1 Skor Penilaian MOS

Skor	Keterangan
1	Buruk
2	Kurang
3	Cukup
4	Baik
5	Sangat Baik

Skor yang diberikan oleh responden untuk masing-masing *file* rekaman dijumlahkan kemudian dibagi 5 sesuai jumlah responden. Rata-rata dari setiap *file* dijumlahkan kemudian dibagi 50 sesuai jumlah *file* untuk mendapatkan skor akhir dari evaluasi MOS. Target keberhasilan dalam penelitian ini ditetapkan pada skor akhir MOS minimal 4. Tabel 2 menggambarkan rentang skor akhir MOS. Skor 4 dianggap sebagai ambang batas kualitas yang menunjukkan bahwa suara yang dihasilkan model terdengar alami, mudah dipahami, dan sesuai dengan karakteristik Bahasa Ocu dialek Siak Hulu. Proses evaluasi diharapkan mampu memberikan gambaran objektif mengenai performa model TTS yang dikembangkan.

Tabel 2 Rentang Skor Akhir MOS

Rentang MOS	Keterangan
1 - 1.99	Buruk
2 - 2.99	Cukup
3 - 3.99	Baik
4 - 5	Sangat Baik

3. HASIL DAN PEMBAHASAN

3.1. Pembuatan *Dataset*

Dataset terdiri dari teks dan audio. Teks Adalah 500 kalimat Bahasa Ocu dialek Siak Hulu yang dibuat bersama penutur asli. Sedangkan *dataset* audio adalah rekaman dari 500 teks yang telah dibuat sebelumnya. Pembuatan *dataset* melalui langkah-langkah seperti membuat teks, merekam audio, cleaning audio, dan membagi data latih dan data uji yang digunakan pada proses pelatihan model. Langkah-langkah tersebut dirincikan di bawah ini.

3.1.1 Pembuatan metadata dan testdata

Penulis bersama penutur asli membuat 500 kalimat Bahasa Ocu dialek Siak Hulu. Setiap kalimat terdiri dari 4-8 kata. Seluruh kalimat tersebut kemudian akan dilakukan validasi oleh pemuka adat di Siak Hulu untuk memastikan kosakata sesuai dengan dialek yang diteliti. Setelah seluruh kalimat divalidasi, 500 kalimat tersebut dibagi menjadi 2 *file* dengan *format* .csv. *File* metadata .csv akan menjadi data latih yang berisi 450 kalimat dan *file* testdata.csv akan menjadi data uji yang berisi 50 kalimat. Kedua *file* berisi 3 kolom yaitu nama *file audio*, kalimat sebagaimana ditulis dan kalimat sebagaimana dibacakan. Kolom kedua dan ketiga ditulis sama persis jika tidak ada angka dalam kalimat. Tebel 3 menunjukkan contoh 5 dari 450 data dalam *file* metadata.csv.

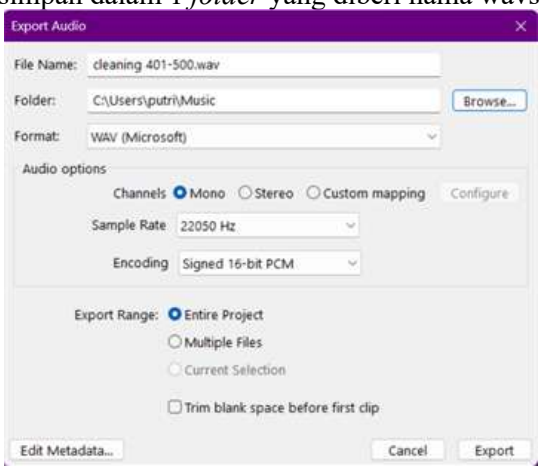
Tabel 3 *File* metadata.csv

1|masi di jakarta kalian bisuk|masi di jakarta kalian bisuk
2|muko den podih dek luko bedaghah|muko den podih dek luko bedaghah
3|sepatu kau digotok kucing tadi|sepatu kau digotok kucing tadi
4|giliran den le menyipak bola du ke gawang|giliran den le menyipak bola du ke gawang
5|sepatu kau dilataan balik suda main diluou|sepatu kau dilataan balik suda main diluou

3.1.2 Pembuatan audio

500 kalimat yang telah dibuat kemudian dilakukan perekaman. Penutur pada penelitian ini adalah wanita berusia 61 tahun. *Microphone* yang digunakan untuk merekam *audio* adalah *microphone* pada *headset* Logitech H111. *Software* yang digunakan untuk merekam adalah Sound Recorder pada Windows 11. Perekaman dilakukan menggunakan peralatan yang sama dari kalimat nomor 1 sampai 500 sehingga kualitas suara yang dihasilkan konsisten. Pada proses perekaman, penutur diminta untuk membacakan setiap kalimat dengan jelas dan tidak terburu-buru. Perekaman dilakukan pada malam hari di dalam ruangan tertutup untuk meminimalisir *noise* yang ikut terekam.

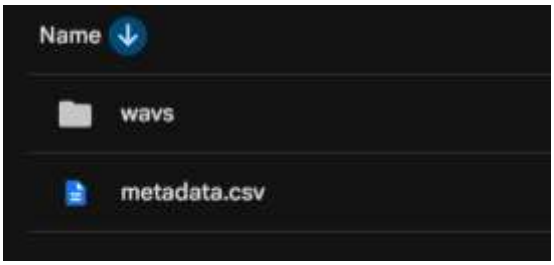
500 *file audio* yang didapatkan dari proses perekaman selanjutnya melalui proses *cleaning*. *Cleaning* dilakukan dengan *software* Audacity. Proses *cleaning* meliputi memotong jeda di awal, tengah, atau akhir kalimat, mengurangi *noise*, serta normalisasi volume agar setiap rekaman memiliki volume yang sama. Rekaman yang telah melalui proses *cleaning* kemudian dilakukan *export*. Pada proses *export*, ditetapkan *format file* wav, *channel* mono, 16-Bit PCM dengan *sample rate* 22050 Hz. Gambar 2 menunjukkan proses *export* serta *format* yang dipilih. Seluruh *audio* disimpan dalam 1 *folder* yang diberi nama wavs.



Gambar 3 Jendela *export* audacity

3.1.3 Pembuatan folder *dataset*

Folder dataset berisi *folder* wavs dan metadata.csv yang akan digunakan dalam proses pelatihan. Folder ini diunggah ke Google Drive agar bisa diakses di Google Colab untuk melakukan proses pelatihan. Gambar 3 menampilkan isi dari *folder dataset*.

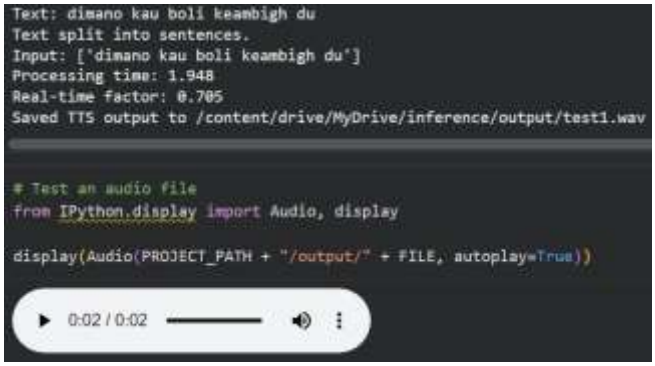


Gambar 4 Folder *dataset*

3.2. Pelatihan

Dataset yang telah dibuat selanjutnya digunakan dalam proses pelatihan model Text-to-Speech (TTS). *Dataset* yang digunakan terdiri dari dua komponen utama, yaitu *file metadata.csv* yang berisi 450 kalimat sebagai pasangan teks dan audio, serta folder *wavs* yang berisi 500 *file* audio hasil perekaman. Data tersebut digunakan sebagai bahan pembelajaran model agar mampu memetakan teks menjadi bentuk gelombang suara yang sesuai. Proses pelatihan model dilakukan menggunakan bahasa pemrograman Python dengan memanfaatkan platform Google Colab sebagai lingkungan komputasi. Selama proses pelatihan yang berlangsung selama kurang lebih lima hari, model secara bertahap mempelajari pola pelafalan, intonasi, serta karakteristik suara dari *dataset* yang digunakan.

Hasil dari proses pelatihan berupa *file best_model.pth* yang menyimpan bobot model terbaik, serta *file config.json* yang berisi konfigurasi model. Kedua *file* ini selanjutnya digunakan pada tahap inferensi untuk menghasilkan suara sintetis berdasarkan teks masukan. Pada proses inferensi, model yang telah dilatih digunakan untuk menghasilkan 50 *file* audio dari data uji yang terdapat dalam *file testdata.csv*. Audio hasil inferensi secara otomatis disimpan dalam format *wav* sehingga dapat langsung digunakan untuk proses evaluasi maupun pengujian lanjutan. Proses inferensi dapat dilakukan secara satu per satu untuk setiap kalimat uji, atau dilakukan secara langsung dengan memproses seluruh isi *file testdata.csv* sehingga menghasilkan 50 *file* audio sekaligus. Gambar 4 menunjukkan contoh hasil inferensi untuk satu kalimat sebagai ilustrasi keluaran suara yang dihasilkan oleh model TTS.



Gambar 5 Inferensi 1 Kalimat

3.3. Evaluasi MOS

Evaluasi terhadap model Text-to-Speech (TTS) dilakukan menggunakan metode Mean Opinion Score (MOS) sebagai pendekatan penilaian subjektif kualitas suara. Proses penilaian ini melibatkan lima orang responden yang merupakan penutur asli, sehingga diharapkan mampu memberikan penilaian yang lebih akurat terhadap kejelasan pengucapan dan kesesuaian dialek yang dihasilkan. Setiap responden diminta untuk mendengarkan dan memberikan penilaian terhadap 50 data audio yang dihasilkan oleh model TTS. Penilaian dilakukan secara individual terhadap setiap *file* audio dengan memberikan skor dalam rentang 1 hingga 5, di mana skor 1 merepresentasikan kualitas suara terburuk dan skor 5 menunjukkan kualitas suara terbaik

sebagaimana ditunjukkan pada Tabel 2[14].

Skor yang diberikan oleh responden didasarkan pada beberapa aspek penilaian utama, yaitu kejelasan suara, tingkat kealamian suara, serta keakuratan dialek yang dihasilkan oleh sistem TTS. Aspek kejelasan suara menilai sejauh mana pengucapan kata dapat dipahami dengan baik, sedangkan kealamian suara berkaitan dengan kemiripan intonasi dan ritme ucapan dengan suara manusia asli. Sementara itu, keakuratan dialek menilai kesesuaian pengucapan dengan karakteristik dialek yang menjadi fokus penelitian.

Skor yang diberikan oleh para responden terhadap setiap *file* audio selanjutnya dijumlahkan dan dibagi dengan jumlah responden, yaitu lima orang, untuk memperoleh nilai rata-rata skor dari masing-masing *file*. Nilai rata-rata dari seluruh *file* audio tersebut kemudian dijumlahkan dan dibagi dengan jumlah total *file*, yaitu 50 *file* audio, sehingga diperoleh skor akhir MOS yang merepresentasikan performa keseluruhan dari model TTS yang dikembangkan. Skor akhir MOS inilah yang digunakan sebagai acuan dalam menentukan hasil evaluasi kualitas suara dari model TTS. Tabel 4 menampilkan rincian skor yang diberikan oleh masing-masing responden terhadap setiap *file* audio yang diuji.

Tabel 4 Hasil evaluasi MOS

No	Kalimat	Responden					Total	Rata-rata
		1	2	3	4	5		
1	pelangi nampak di langik suda hujan	4	2	4	5	4	19	3.8
2	warnanyo ancak botul dipandang mato	5	5	5	5	5	25	5
3	ai pane membuek kulik jadi itam	5	5	5	5	5	25	5
4	pakai payung bilo nak ke luou	4	3	4	5	4	20	4
5	minum aigh puti banyak banyak	5	4	5	5	5	24	4.8
6	supoyo badan ndak koing	4	2	4	4	5	19	3.8
7	musim kemaghau panjang aigh sumu koing	5	5	4	5	4	23	4.6
8	sumu di uma awak la mulai koing	4	2	5	5	5	21	4.2
9	imat imat menggunon ai berosi	5	5	5	5	4	24	4.8
10	gunung merapi du tenggi nampak doi jauh	4	5	5	5	5	24	4.8
11	asok puti keluou dai puncaknyo	4	2	5	5	3	19	3.8
12	pemandangan dai ate puncak ancak botul	5	5	5	5	5	25	5
13	lauik awan nampak tebontang dai ate puncak	3	5	4	5	5	22	4.4
14	mendaki gunung perolu badan nan kuat	5	5	4	5	4	23	4.6
15	Awak harus be5siap sebolun beangkek	5	4	5	5	4	23	4.6
16	jan cecubo membuang sampah di ate gunung	5	1	5	4	5	20	4
17	alam nan ancak go kuaso tuhan	4	5	5	5	5	24	4.8
18	ombak di pantai bagulung gulung godang	4	3	5	5	5	22	4.4
19	kosik puti tebontang sepanjang pantai	5	3	5	5	4	22	4.4
20	budak budak membuek gambar istana doi kosik	5	5	5	5	5	25	5
21	mandi di lauik harus behati hati	4	5	4	4	5	22	4.4
22	jan bonang jauh botul ke tonga	5	4	4	5	5	23	4.6
23	matoai tengolam bola ke barat	5	4	5	5	5	24	4.8
24	warnanyo megha membuek langik jadi ancak	3	5	5	5	5	23	4.6
25	menengok langik sonjo di topi pantai sonang	4	1	5	5	5	20	4
26	uwang mencai ikan banyak membaok ikan balik	5	5	4	5	5	24	4.8
27	ikannyo dijual di pasagh isuk pagi	4	1	5	2	5	17	3.4
28	hasilnyo untuk memboli belanjo dapu	5	5	5	5	5	25	5
29	pencai ikan boek kojonyo	3	4	5	5	4	21	4.2
30	begantung pado keadaan lauik	4	2	5	4	5	20	4
31	kapal ketek diayun ombak godang	4	5	5	5	5	24	4.8
32	bedoa supayo selamat sampai ke tompek nan dituju	5	5	4	5	5	24	4.8
33	api unggun dibuek malam ai	5	5	5	5	5	25	5

34	bekumpul mengeliling sambil menyanyi besamo	5	4	5	5	5	24	4.8
35	memanggang jagung samo ikan	4	1	5	5	5	20	4
36	malam nan ancak di bawah bintang	5	5	5	5	5	25	5
37	beajagh dai kesalahan nan ola lalu	4	3	4	2	5	18	3.6
38	jan diulang kesalahan nan samo	5	5	5	5	5	25	5
39	setiap uwang pasti pona membuek salah	5	5	5	5	5	25	5
40	pebaiki dii untuk maso dopan	4	4	5	5	5	23	4.6
41	maso dopan masi panjang go	4	5	5	5	5	24	4.8
42	rencanaan segalonyo dai kini elok elok	5	5	5	5	5	25	5
43	bekojo koe supayo tecapai angan angan	5	3	5	2	5	20	4
44	kegagalan jadikan pelajaran untuk maso nan kan datang	4	5	4	5	5	23	4.6
45	jan pona menyorah pado keadaan	3	5	5	5	5	23	4.6
46	sukuri apo nan dapek dek awak	5	4	5	5	5	24	4.8
47	nikmati setiap keadaan dalam iduik go	4	3	5	5	5	22	4.4
48	mencai bahagia du sonang sebotulnyo	4	2	4	5	5	20	4
49	iduik besamo dalam damai du lomak aso	5	5	5	5	5	25	5
50	makanan angek langsung dimakan lomak asonyo	3	2	5	5	5	20	4
TOTAL								225.4
RATA-RATA								4.508

Tabel 4 menunjukkan bahwa skor akhir Mean Opinion Score (MOS) yang diperoleh sebesar 4,508. Nilai ini mengindikasikan bahwa kualitas suara hasil sintesis berada pada kategori baik hingga sangat baik berdasarkan penilaian subjektif responden. Meskipun demikian, masih terdapat beberapa catatan yang diberikan oleh responden terkait kualitas audio yang dihasilkan. Beberapa responden menyatakan bahwa bunyi konsonan ’gh’ belum terdengar dengan jelas pada beberapa kata tertentu. Selain itu, masih ditemukan sejumlah kalimat yang diucapkan dengan intonasi yang terkesan kaku atau robotik, sehingga kurang menyerupai intonasi alami manusia. Responden juga mencatat adanya kesalahan pelafalan berupa tertukarnya pengucapan huruf ’g’ dan ’k’ pada beberapa bagian kalimat. Hal-hal tersebut terjadi dikarenakan jumlah data yang terbatas dalam penelitian sehingga mengurangi kualitas suara yang dihasilkan[15].

Walaupun terdapat kekurangan tersebut, secara keseluruhan model TTS yang dihasilkan mampu menghasilkan suara yang mendekati suara manusia asli. Suara yang dihasilkan dapat dipahami dengan baik oleh pendengar, baik dari segi kejelasan pengucapan maupun kelancaran kalimat. Hasil ini menunjukkan bahwa model TTS yang dikembangkan telah memiliki performa yang cukup baik dan berpotensi untuk dikembangkan lebih lanjut guna meningkatkan kejelasan fonem serta naturalitas intonasi suara yang dihasilkan.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem TTS untuk Bahasa Ocu dialek Desa Pangkalan Baru, Kecamatan Siak Hulu, Kabupaten Kampar menggunakan VITS. Proses pengembangan dilakukan melalui pembuatan *dataset* berupa 500 kalimat Bahasa Ocu dialek Siak Hulu, pelatihan model menggunakan 450 data latih, dan pengujian menggunakan 50 data uji. Hasil pengujian menggunakan metode MOS mendapatkan nilai 4.508. Nilai ini termasuk dalam rentang sangat baik berdasarkan tabel 2 yang menjelaskan keterangan dari setiap skor. Nilai ini menunjukkan bahwa sistem mampu menghasilkan suara yang terdengar natural, jelas, dan dapat dipahami oleh penutur asli. Penelitian ini diharapkan dapat berkontribusi untuk pengembangan TTS bahasa daerah lainnya.

DAFTAR PUSTAKA

[1] O. Mailani, I. Nuraeni, S. A. Syakila, J. Lazuardi, and P. I. Komunikasi, “Bahasa Sebagai Alat Komunikasi Dalam Kehidupan Manusia,” *KAMPRET*, vol. 1, no. 2, pp. 1–10, 2021, doi:

https://doi.org/10.35335/kampret.v1i1.8.

[2] A. D. Izzanti, R. M. Nasution, A. H. Wasik, I. M. Juanda, and S. Nasution, "Hakikat Bahasa dalam Objek Kajian Linguistik," *Semantik : Jurnal Riset Ilmu Pendidikan, Bahasa dan Budaya*, vol. 3, no. 1, pp. 188–194, Jan. 2025, doi: https://doi.org/10.61132/semantik.v3i1.1394.

[3] B. Aritonang, "Penggunaan Bahasa Daerah Generasi Muda Provinsi Maluku Utara dan Papua Barat," *Ranah: Jurnal Kajian Bahasa*, vol. 9, no. 2, pp. 160–177, Dec. 2020, doi: https://doi.org/10.26499/rnh.v9i2.2936.

[4] R. Prina Br Sembiring and F. Ayu Lestari, "Revitalisasi Bahasa Daerah Dalam Era Globalisasi Antara Pelestarian Dan Modernisasi," *BASAYA*, vol. 1, no. 1, pp. 24–29, 2024.

[5] Zikri Ahmad and Fadlilah Afi, "Pemertahanan Bahasa Ocu pada Interaksi Masyarakat di Kawasan Wisata Sungai Gelombang (Kajian Sosiolinguistik)," *Logat*, vol. 9, no. 1, pp. 42–51, May 2022, doi: https://doi.org/10.36706/logat.v9i1.248.

[6] S. Dahlan, "Hubungan Bahasa & Dialek Melayu Kab Kampar Bagian Timur Dengan Bahasa Di Daerah Bekas Kerajaan Siak (1983)," *Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan*, 1983.

[7] J. Xu *et al.*, "LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition," *KDD*, vol. 11, no. 20, 2020, doi: https://doi.org/10.1145/3394486.

[8] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-End Adversarial Text-to-Speech," in *ICLR*, Mar. 2021. doi: https://doi.org/10.48550/arXiv.2006.03575.

[9] R. Ardiansyah, B. Bastiar, A. Adzikirani, D. Marya, and A. Novianti, "Optimasi Penerjemahan Bahasa Asing Dengan Teknologi IoT Pada Kelas Internasional Politeknik Negeri Malang," *JURNAL ELTEK*, vol. 23, no. 1, pp. 1–8, Apr. 2025, doi: https://doi.org/10.33795/eltek.v23i1.6951.

[10] M. Chen *et al.*, "AdaSpeech: Adaptive Text to Speech for Custom Voice," in *ICLR*, Mar. 2021.

[11] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto: IEEE, Feb. 2021. doi: https://doi.org/10.1109/ICASSP39728.2021.9413851.

[12] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *ICML*, 2021. doi: https://doi.org/10.48550/arXiv.2106.06103.

[13] F. A. Martin, M. Malfaz, Á. Castro-González, J. C. Castillo, and M. Á. Salichs, "Four-Features Evaluation of Text to Speech Systems for Three Social Robots," *Electronics (Basel)*, vol. 9, no. 2, pp. 1–23, Feb. 2020, doi: 10.3390/electronics9020267.

[14] Y. Kowalczuk and J. Holub, "Evaluation of Digital Watermarking on Subjective Speech Quality," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: https://doi.org/10.1038/s41598-021-99811-x.

[15] B. T. Vecino *et al.*, "Lightweight End-to-end Text-to-speech Synthesis for low resource on-device applications," in *12th ISCA Speech Synthesis Workshop*, May 2025. doi: https://doi.org/10.21437/SSW.2023-35.