

IMPLEMENTASI *DEEP LEARNING HYBRID* RESNET50V2-BIGRU PADA APLIKASI WEB *VISUAL SPEECH RECOGNITION* BAGI KOMUNITAS TULI

Salma Nurfauziah^{1*}, Asep Budiman Kusdinar², Fathia Frazna Az-Zahra³

^{1,2,3}Program Studi Teknik Informatika, Universitas Muhammadiyah Sukabumi; Jl. R. Syamsudin No. 50, Kota Sukabumi, Jawa Barat

Keywords:

Bi-GRU; *Deep Learning*; ResNet50V2; *Visual Speech Recognition*; Web.

Correspondent Email:

salmanurfauziah100@ummi.ac.id

Abstrak. Visual Speech Recognition (VSR) merupakan teknologi komunikasi krusial bagi komunitas Tuli, terutama di lingkungan bising di mana sistem berbasis audio gagal berfungsi. Pengembangan VSR untuk bahasa Indonesia menghadapi tantangan kelangkaan dataset audio-visual dan tingginya ambiguitas kata homofon. Selain itu, tingginya beban komputasi data video seringkali menghambat implementasi pada platform web. Oleh karena itu, penelitian ini bertujuan mengembangkan sistem VSR bahasa Indonesia yang efisien menggunakan arsitektur hibrida untuk aplikasi web responsif. Metodologi yang digunakan adalah CRISP-ML(Q). Ekstraksi fitur visual spasial dilakukan menggunakan jaringan pre-trained ResNet50V2, dikombinasikan dengan Bidirectional Gated Recurrent Unit (Bi-GRU) untuk memodelkan dependensi sekuensial temporal. Model dilatih dengan dataset publik LUMINA dan dievaluasi menggunakan 250 sampel video dataset privat. Hasil eksperimen menunjukkan akurasi validasi mencapai 95,49%. Pada pengujian data baru, sistem mencatat Word Error Rate (WER) sebesar 11,18% dan Character Error Rate (CER) 8,74%. Implementasi berbasis web menggunakan Streamlit mencapai waktu respons inferensi 2-4 detik. Kesimpulannya, model ini sukses menerjemahkan gerakan bibir menjadi teks secara akurat dan menyediakan alat komunikasi responsif bagi penyandang disabilitas rungu.



Copyright © 2026 The Author(s), Published by Jurnal Informatika dan Teknik Elektro Terapan. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract. *Visual Speech Recognition (VSR) is a crucial communication technology for the Deaf community, especially in noisy environments where audio-based systems fail to function. The development of VSR for the Indonesian language faces challenges due to the scarcity of audio-visual datasets and the high ambiguity of homophones. Additionally, the high computational load of video data often hinders implementation on web platforms. Therefore, this research aims to develop an efficient Indonesian VSR system using a hybrid architecture for responsive web applications. The methodology employed is CRISP-ML(Q). Spatial visual feature extraction is performed using a pre-trained ResNet50V2 network, combined with a Bidirectional Gated Recurrent Unit (Bi-GRU) to model temporal sequential dependencies. The model was trained on the public LUMINA dataset and evaluated using 250 video samples from a private dataset. Experimental results show a validation accuracy of 95.49%. On new data, the system achieved a Word Error Rate (WER) of 11.18% and a Character Error Rate (CER) of 8.74%. The web-based implementation using Streamlit achieved an inference response time of 2–4 seconds. In conclusion, this model successfully translates lip movements into text with high accuracy and provides a responsive communication tool for people with hearing impairments.*

1. PENDAHULUAN

Isu gangguan pendengaran dan ketulian di Indonesia memiliki urgensi yang tinggi. Kementerian Kesehatan memperkirakan sekitar 10% penduduk Indonesia merupakan penyandang disabilitas, dengan kelompok tunarungu menempati posisi terbanyak kedua [1], [2]. Dalam interaksi sosial sehari-hari, penyandang tunarungu sangat mengandalkan teknik membaca gerak bibir (*lip-reading*) [2]. Namun, metode manual ini sering memicu kesalahpahaman fatal di tempat umum akibat tingginya tingkat kebisingan [3]. Salah satu solusi teknologi yang umum digunakan untuk membantu komunikasi adalah *Automatic Speech Recognition* (ASR). Sayangnya, performa ASR menurun drastis di lingkungan dengan kebisingan tinggi (*high-noise environment*) karena sistem kesulitan membedakan antara suara penutur dan derau lingkungan [4], [5]. Keterbatasan teknologi berbasis suara inilah yang mendorong urgensi pengembangan *Visual Speech Recognition* (VSR), yang murni mengandalkan isyarat visual dari gerakan artikulator seperti bibir tanpa membutuhkan input audio sama sekali [6].

Sebagai bagian dari *state of the art*, sejumlah penelitian terdahulu telah mengeksplorasi arsitektur *Deep Learning* untuk membangun sistem VSR yang andal. Pendekatan hibrida yang menggabungkan *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN), seperti fusi 3D-CNN dan LSTM, telah berhasil diimplementasikan menjadi purwarupa asisten visual berbasis web pada dataset berbahasa Inggris [7]. Untuk mengoptimalkan ekstraksi fitur visual secara presisi pada kondisi lingkungan bising, pemodelan menggunakan modifikasi ResNet dan *Bidirectional GRU* (Bi-GRU) juga terbukti efektif [8]. Dalam konteks pengenalan gerak bibir bahasa Indonesia, inovasi telah mencapai kemajuan melalui penggunaan arsitektur kompleks seperti penggabungan *Channel Spatial Attention* (CSA) ResNet-34, Bi-GRU, dan *Mamba block*. Model tersebut berhasil mencapai akurasi *state-of-the-art* sebesar 60,51% pada klasifikasi data level kata dari himpunan data *Indonesian Lipreading in the Wild* (IDLRW) [9]. Meskipun teknologi VSR berkembang pesat, masih

terdapat kesenjangan (*gap analysis*) yang signifikan untuk implementasi praktisnya pada bahasa Indonesia. Pertama, kelangkaan dataset VSR bahasa Indonesia yang bersifat *visual-only* menyulitkan sistem dalam menghadapi kompleksitas morfologis, khususnya tingginya tingkat ambiguitas kata homofon (kata dengan gerakan bibir identik namun makna berbeda) [10]. Kedua, penelitian model mutakhir bahasa Indonesia sebelumnya (seperti model dengan *Mamba block*) memiliki beban komputasi yang sangat masif hingga mencapai 72,35 juta parameter [9]. Sifat pemrosesan data video yang *spatiotemporal* menuntut daya komputasi tinggi, sehingga *deployment* sistem pada platform web yang dapat diakses secara luas sering kali terkendala masalah latensi dan responsivitas. Oleh karena itu, diperlukan sebuah arsitektur yang mampu menyeimbangkan antara akurasi tinggi untuk memecahkan konteks homofon dengan efisiensi waktu komputasi (*inference time*) [11].

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk mengembangkan sistem VSR bahasa Indonesia yang efisien menggunakan arsitektur hibrida ResNet50V2 dan Bi-GRU. Pemilihan ResNet50V2 didasarkan pada mekanisme koneksi sisa (*skip connections*) yang terbukti tangguh mencegah masalah degradasi gradien saat mendeteksi fitur morfologi bibir yang sangat halus [12]. Sementara itu, komponen Bi-GRU bertugas memodelkan ketergantungan temporal untuk memecahkan ambiguitas kata homofon berdasarkan pemahaman konteks kalimat secara keseluruhan [13]. Tujuan akhir dari penelitian ini adalah mengimplementasikan dan memvalidasi kelayakan model hibrida tersebut ke dalam sebuah purwarupa aplikasi berbasis web menggunakan kerangka kerja Streamlit, guna menyediakan alat komunikasi visual yang responsif bagi komunitas Tuli di lingkungan terbatas.

2. TINJAUAN PUSTAKA

2.1. *Visual Speech Recognition (VSR) dan Viseme*

Visual Speech Recognition (VSR) atau *lip-reading* otomatis merupakan bidang yang mengekstraksi informasi linguistik dari sinyal visual pergerakan artikulator seperti bibir tanpa

bergantung pada sinyal audio [6]. Unit dasar dalam domain visual ini disebut *viseme* (*visual phoneme*). Pemetaan antara fonem dan *viseme* umumnya bersifat *many-to-one mapping*, di mana beberapa fonem berbeda (seperti /p/, /b/, dan /m/) menghasilkan bentuk bibir yang identic [14]. Hal ini menyebabkan tingginya fenomena ambiguitas kata homofon yang menyulitkan pengenalan wicara visual [15].

Tabel 1. Contoh Pengelompokan Fonem ke dalam Kelas Viseme (Mapping Phoneme-to-Viseme) [16].

Viseme Type	Phonemes Set
Consonant	p, t, l, g, j, z, m, sy, ng, w, v, b, k, d, c, f, h, kh, n, r, y, ny, silent
Single Vowelst	a, e, E, i, o, u
Double Vowels	ao, au, ai, oi

2.2. Ekstraksi Region of Interest (ROI) Menggunakan MediaPipe

Pra-pemrosesan data VSR membutuhkan lokalisasi area bibir yang presisi. Algoritma berbasis grafik pada MediaPipe Face Mesh digunakan karena mampu mengestimasi 468 titik *landmark* wajah 3D secara asinkron dengan akurasi tinggi [17]. Teknologi ini memberikan resolusi fitur spasial yang tinggi sekaligus meminimalkan latensi, menjadikannya sangat ideal untuk memotong (*cropping*) Region of Interest (ROI) area bibir pada aplikasi VSR responsive.

2.3. Arsitektur Hybrid ResNet50V2-BiGRU

Pemodelan VSR pada penelitian ini mengintegrasikan *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN). Jaringan ResNet50V2 digunakan sebagai *frontend* pengekstraksi fitur spasial karena struktur *identity mappings* (koneksi sisa) di dalamnya terbukti efektif mencegah masalah degradasi gradien pada jaringan dalam [12]. Fitur spasial tersebut kemudian diteruskan ke komponen *Bidirectional Gated Recurrent Unit* (Bi-GRU). Bi-GRU memproses urutan fitur secara dua arah (maju dan mundur), memungkinkan sistem menangkap konteks linguistik masa lalu dan masa depan secara komprehensif untuk memecahkan ambiguitas kata homofon [8], [13].

2.4. Metrik Evaluasi Kinerja

Performa pengenalan sistem VSR diukur menggunakan dua metrik standar komparasi

secara kuantitatif, yaitu *Word Error Rate* (WER) dan *Character Error Rate* (CER). WER digunakan untuk mengukur persentase kesalahan transkripsi pada tingkat kata menggunakan perhitungan jarak edit (*Levenshtein distance*) dengan persamaan [18]:

$$WER = \frac{S + D + I}{N} \quad (1)$$

Di mana S adalah Substitution (kata yang salah dikenali), D adalah Deletion (kata tidak muncul di prediksi), I adalah Insertion (kata prediksi berlebih), dan N adalah total kata referensi. Sementara itu, CER mengukur tingkat kesalahan pengejaan pada level karakter dengan persamaan [19]:

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad (2)$$

Di mana S_c , D_c , dan I_c masing-masing adalah jumlah kesalahan substitusi, penghapusan, dan penyisipan pada level karakter, serta N_c adalah total karakter pada referensi.

3. METODE PENELITIAN

Penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Machine Learning with Quality assurance* (CRISP-ML(Q)) [20] yang bertujuan menstandarisasi siklus hidup *Machine Learning* untuk menjamin ketangguhan sistem di lingkungan produksi. Tahapan yang dilakukan meliputi *business* dan *data understanding*, *data engineering*, *model engineering*, evaluasi, hingga *deployment* aplikasi.

3.1. Business dan Data Understanding

Penelitian ini menggunakan dua jenis himpunan data audio-visual berbahasa Indonesia. Data sekunder (pelatihan) bersumber dari korpus publik LUMINA yang terdiri dari 13.808 sampel video *crop* bibir dari 14 penutur [10]. Untuk pengujian (*testing*) dan validasi di kondisi nyata, dikumpulkan data primer (privat) sebanyak 250 video rekaman *frontal* dari 6 partisipan yang mengucapkan 112 kelas kalimat secara natural.

3.2. Data Engineering

Setiap video berformat .mp4 diekstraksi menjadi *frame* citra. Algoritma MediaPipe Face Mesh digunakan untuk melokalisasi dan memotong (*cropping*) Region of Interest (ROI)

bibir agar sistem dapat berfokus murni pada fitur artikulasi [17]. Khusus pada data latih, diterapkan augmentasi berupa rotasi spasial (-8 hingga 8 derajat), penyesuaian kecerahan, dan pembalikan horizontal. Seluruh citra diubah ukurannya (*resize*) menjadi standar 100×50 piksel dan dikonversi dari warna RGB ke format *Grayscale*. Dimensi temporal diseragamkan menjadi tepat 150 *frame* menggunakan teknik *zero-padding* untuk video berdurasi pendek dan *uniform sampling* untuk video berdurasi panjang. Hasil akhir dari tahapan ini berupa *array* tensor berdimensi masukan (150, 50, 100, 1) per sampel video.

3.3. Arsitektur Pemodelan Hybrid

Sistem mengadopsi arsitektur *Deep Learning* gabungan antara ResNet50V2 dan *Bidirectional Gated Recurrent Unit* (Bi-GRU). Lapisan pertama menggunakan jaringan pra-latih (*pre-trained*) ResNet50V2 dengan *TimeDistributed wrapper* untuk mengekstraksi fitur spasial secara independen dari ke-150 *frame* [12]. Untuk menjaga efisiensi, bobot konvolusinya dibekukan (*frozen*). Peta fitur diubah menjadi vektor 1D berukuran 2.048 melalui *Global Average Pooling 2D*. Urutan fitur tersebut kemudian diproses oleh *layer* Bi-GRU (512 unit) untuk mempelajari ketergantungan temporal gerakan bibir secara dua arah (maju dan mundur) [8], [13]. Lapisan *Dropout* (30%) ditambahkan sebelum diintegrasikan ke *Dense Layer* dengan aktivasi *Softmax* untuk memprediksi kelas kalimat.

3.4. Evaluasi dan Deployment

Model dilatih menggunakan optimasi *Adam* dan metrik *Categorical Cross-Entropy*, dilengkapi dengan mekanisme *Early Stopping* untuk mencegah *overfitting*. Kinerja prediksi dievaluasi secara kuantitatif menggunakan jarak edit *Levenshtein*, yaitu metrik *Word Error Rate* (WER) [18] dan *Character Error Rate* (CER) [19]. Model terbaik kemudian diimplementasikan ke dalam aplikasi web interaktif menggunakan kerangka kerja Streamlit [21], [22], sebuah pendekatan *deployment* yang efisien untuk mendeploy pemodelan bahasa dan visual berbasis *machine learning* ke pengguna akhir.

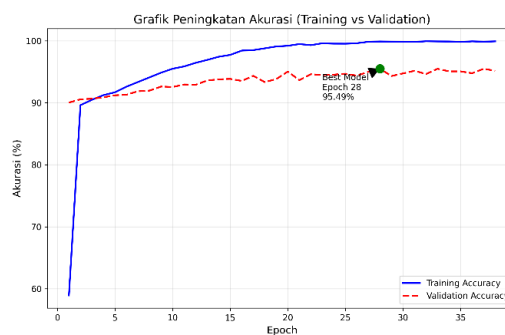
4. HASIL DAN PEMBAHASAN

Bagian ini menguraikan hasil pengujian dari arsitektur *hybrid* ResNet50V2-BiGRU yang diusulkan, interpretasi metrik evaluasi,

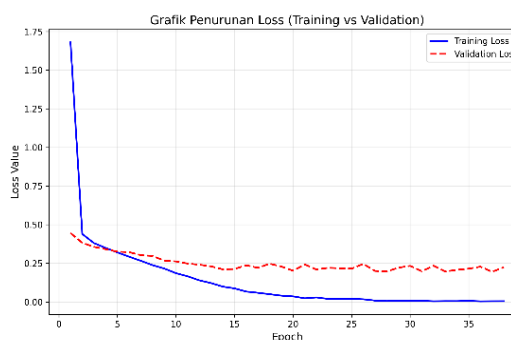
pemecahan masalah ambiguitas homofon, serta implikasi langsung dari implementasi model pada platform web.

4.1. Kinerja Konvergensi Model pada Fase Pelatihan

Pelatihan model dilakukan menggunakan data latih (LUMINA) selama maksimal 100 *epoch* dengan mekanisme *Early Stopping*. Berdasarkan hasil eksperimen, proses pelatihan berhenti secara otomatis pada *epoch* ke-38. Model menunjukkan konvergensi yang sangat baik di mana akurasi pelatihan meningkat tajam dan *loss* menurun signifikan pada 10 *epoch* pertama. Kinerja optimal (model terbaik) dicapai pada *epoch* ke-28 dengan capaian *Training Accuracy* sebesar 99,90% dan *Validation Accuracy* sebesar 95,49%.



Gambar 1. Grafik Akurasi Pelatihan dan Validasi



Gambar 2. Grafik Loss Pelatihan dan Validasi

Jarak (*gap*) yang stabil antara metrik pelatihan dan validasi membuktikan bahwa model terhindar dari *overfitting*. Hal ini terjadi karena mekanisme *identity mappings* pada ResNet50V2 sangat efektif mempertahankan aliran gradien pada jaringan dalam saat mengekstraksi morfologi bibir yang halus [12].

4.2. Evaluasi Word Error Rate (WER) dan Character Error Rate (CER)

Evaluasi ketangguhan model (*robustness*) diuji menggunakan 250 sampel

data video privat dengan kondisi natural yang belum pernah dilihat model sebelumnya.

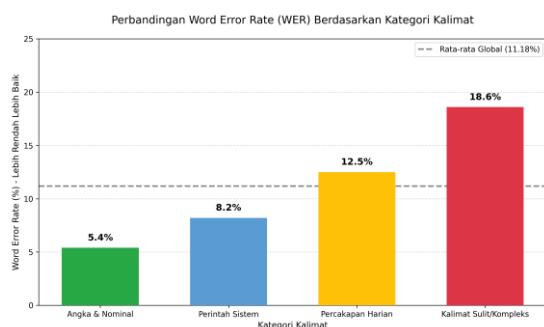
Tabel 2. Hasil Evaluasi WER dan CER

Metrik Evaluasi	Nilai Rata-rata (%)	Interpretasi
<i>Word Error Rate</i> (WER)	11.18%	Tingkat akurasi pengenalan kata (<i>Word Accuracy</i>) mencapai $\approx 88.82\%$
<i>Character Error Rate</i> (CER)	8.74%	Tingkat kesalahan ejaan per karakter sangat rendah, menunjukkan model mampu menangkap fitur viseme dengan presisi tinggi

Berdasarkan Kategori Kalimat Secara global, sistem mencatat rata-rata *Word Error Rate* (WER) sebesar 11,18% (akurasi pengenalan kata $\approx 88,82\%$) dan *Character Error Rate* (CER) yang sangat rendah sebesar 8,74%. Temuan bahwa CER lebih rendah dibandingkan WER sangat relevan dengan karakteristik bahasa aglutinatif [19]. CER berfungsi mendiagnosis presisi fitur *viseme* lokal, di mana model seringkali hanya salah memprediksi 1 huruf (satu *viseme*), namun kesalahan kecil ini memberikan penalti yang besar pada metrik WER (level kata).

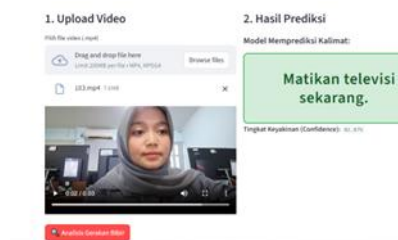
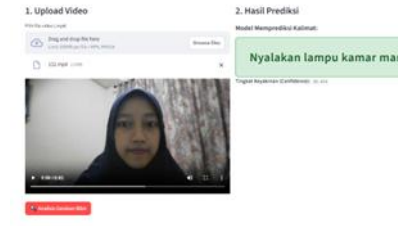
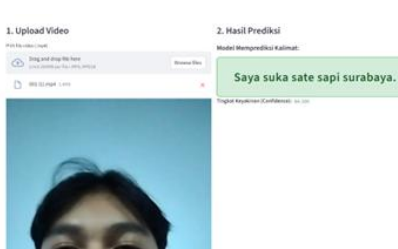
4.3. Pemecahan Ambiguitas Homofon dan Anomali Prediksi

Tantangan terbesar dalam membaca gerak bibir adalah fenomena kata homofon, seperti fonem bilabial /p/, /b/, dan /m/ yang memiliki penutupan bibir identik. Pengujian komparatif membuktikan ketajaman model dalam mengatasi masalah ini secara semantik.



Gambar 3. Perbandingan Word Error Rate

Tabel 3. Perbandingan Prediksi Model vs Ground Truth

Prediksi Model	Label Asli	WER	CER	Confidence	Hasil
matikantelevis isekarang	Matikan televisi sekarang	0.00	0.00	92.87%	
nyalakanlamp ukamarmandi	Nyalakan lampu kamar mandi	0.00	0.00	36.41%	
sayasukasatesa pisurabaya	Buka kunci pintu depan	1.25	90%	84.16%	

Tabel 4. Analisis Homofon pada Sampel Bilabial

Kalimat Asli	Prediksi Model	Status	Analisis Viseme
Papa pulang bawa pepaya pipih	Papa pulang bawa pepaya pipih	Berhasil	Model sukses membedakan /p/ dan /b/ meskipun bentuk bibir identik.

Pada pengujian kalimat aliterasi seperti "*Papa pulang bawa pepaya pipih*", model secara sempurna mampu membedakan kata "*Papa*" (subjek) dan "*Bawa*" (predikat). Hal ini membuktikan bahwa Bi-GRU yang memproses urutan fitur secara maju dan mundur berhasil mengidentifikasi konteks struktur kalimat secara global (*past to future*), melampaui kemampuan model *Pure CNN* konvensional yang kehilangan dinamika temporal [23].

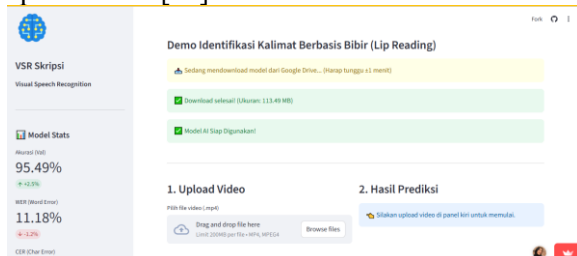
Meskipun demikian, ditemukan anomali berupa "halusinasi prediksi" di mana model memprediksi kalimat "*Buka kunci pintu depan*" menjadi "*sayasukasatesapisurabaya*" dengan tingkat keyakinan (*confidence*) yang sangat tinggi (84,16%). Anomali ini terjadi akibat keterbatasan skema *Closed-Set Sentence Classification*, di mana model memaksakan

keyakinan pada kelas kalimat terdekat saat kehilangan jejak konteks visual utama pada *frame* awal.

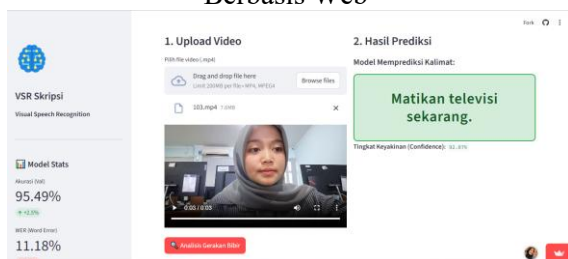
Temuan ini mengonfirmasi penelitian sebelumnya bahwa sistem pengenalan ujaran memerlukan data yang sangat masif dan integrasi *Language Model* (LM) untuk menyempurnakan prediksi *out-of-context* [10]. Namun, model hibrida ini memiliki keunggulan telak pada efisiensi komputasi jika dibandingkan arsitektur *state-of-the-art* (SOTA) seperti integrasi *Mamba block* yang membutuhkan hingga 72,35 juta parameter [9].

4.4. Implikasi Teoritis dan Implementasi Sistem

Dari sisi implikasi implementasi, model dieksekusi menggunakan kerangka kerja *Streamlit Cloud* untuk menciptakan prototipe aplikasi web [22].



Gambar 4. Tampilan Antarmuka Sistem VSR Berbasis Web



Gambar 5. Tampilan Hasil Prediksi Kalimat

Pengujian fungsionalitas mencatat bahwa arsitektur hibrida ini memiliki waktu respons pemrosesan keseluruhan hanya 2 hingga 4 detik per sampel video (dengan proses komputasi inferensi *Deep Learning* murni hanya memakan waktu 0,5 detik). Kecepatan ini membuktikan bahwa arsitektur ResNet50V2-BiGRU mencapai "titik keseimbangan" (*sweet spot*) yang ideal. Implikasi teknisnya, penerapan sistem VSR ini sangat *lightweight* (ringan) dan tidak menuntut perangkat keras (*Hardware Accelerator* GPU) mahal pada sisi klien [24]. Sebagai alat bantu disabilitas, sistem ini siap menyediakan sarana komunikasi visual yang interaktif dan *user-friendly* bagi komunitas Tuli di sektor layanan publik yang bisling [21].

5. KESIMPULAN

- Model terbukti sangat efektif mengklasifikasi kalimat bahasa Indonesia berbasis visual dengan akurasi validasi mencapai 95,49% serta tingkat kesalahan yang sangat rendah (WER 11,18% dan CER 8,74%). Sistem juga sukses diimplementasikan menjadi aplikasi web <https://vsr-skripsi-demo.streamlit.app/> yang *lightweight*, responsif (inferensi 0,5

detik), dan siap digunakan tanpa instalasi tambahan..

- Keterbatasan sistem saat ini meliputi penggunaan skema klasifikasi kalimat tertutup (hanya mengenali 112 kelas), belum mendukung deteksi *real-time* langsung dari kamera (harus unggah video), beban komputasi *backbone* ResNet50V2 yang relatif berat untuk perangkat seluler *offline*, serta variasi data latih yang masih bergantung pada kondisi terkontrol.
- Untuk menyempurnakan sistem, disarankan memperkaya variasi data latih agar model lebih tangguh (*robust*), mengintegrasikan WebRTC untuk deteksi *real-time*, mengeksplorasi arsitektur ringan (seperti MobileNetV3) untuk penggunaan *mobile offline*, serta beralih ke pendekatan *Character-based/Phoneme-based decoding* menggunakan *CTC Loss* agar mampu mengeja kosakata baru di luar data latih.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan penuh dalam penyelesaian penelitian ini, khususnya kepada dosen pembimbing yang telah memberikan arahan, pimpinan fakultas dan program studi Universitas Muhammadiyah Sukabumi atas fasilitas yang diberikan, keluarga tercinta, teman-teman seperjuangan, serta seluruh relawan yang telah bersedia meluangkan waktu untuk berpartisipasi dalam pengambilan data video penelitian ini.

DAFTAR PUSTAKA

- [1] A. Muhawarman, "Kemenkes Ajak Masyarakat Peduli Kesehatan Pendengaran," Jakarta, Mar. 03, 2025. Accessed: Feb. 10, 2026. [Online]. Available: <https://kemkes.go.id/eng/kemenkes-ajak-masyarakat-peduli-kesehatan-pendengaran>
- [2] H. L. Handojo and B. A. Sukada, "Implementasi Design Sarana Terapi Bermain untuk Pengembangan Kemampuan Wicara dan Bahasa Anak Penyandang Tunarungu," *Jurnal Sains, Teknologi, Urban, Perancangan, Arsitektur (Stupa)*, vol. 6, no. 1, pp. 603–612, Apr. 2024, doi: 10.24912/stupa.v6i1.27502.
- [3] M. Kim, H.-I. Kim, and Y. M. Ro, "Prompt Tuning of Deep Neural Networks for Speaker-adaptive Visual Speech Recognition," Oct.

- 2024, [Online]. Available: <http://arxiv.org/abs/2302.08102>
- [4] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," Jun. 2023, doi: 10.1109/ICASSP49357.2023.10096889.
 - [5] D. Ivanko, D. Ryumin, and A. Karpov, "A Review of Recent Advances on Deep Learning Methods for Audio-Visual Speech Recognition," Jun. 01, 2023, *MDPI*. doi: 10.3390/math11122665.
 - [6] R. Inamdar, K. sundarr, D. Khandelwal, and A. KB, "Lips Reading Using 3D Convolution and LSTM," Dec. 13, 2023. doi: 10.20944/preprints202312.0928.v1.
 - [7] Pradiptatmaka Feliciano Akwila, Suciati Nanik, and Navastara Adni Dini, "Analisis Pembacaan Gerakan Bibir Menggunakan Gabungan Arsitektur Convolutional Neural Network dan Recurrent Neural Network," 2021.
 - [8] G. M. Rahmatullah, S.-J. Ruan, and L. P.-H. Li, "Recognizing Indonesian words based on visual cues of lip movement using deep learning," *Measurement*, vol. 250, p. 116968, Jun. 2025, doi: <https://doi.org/10.1016/j.measurement.2025.116968>.
 - [9] E. R. Setyaningsih, A. N. Handayani, W. S. G. Irianto, Y. Kristian, and C. T. S. L. Chen, "LUMINA: Linguistic unified multimodal Indonesian natural audio-visual dataset," *Data Brief*, vol. 54, p. 110279, Jun. 2024, doi: 10.1016/j.dib.2024.110279.
 - [10] H. Laux, E. Mededovic, A. Hallawa, L. Martin, A. Peine, and A. Schmeink, "LiteVSR: Efficient Visual Speech Recognition by Learning from Speech Representations of Unlabeled Data," Dec. 2023, [Online]. Available: <http://arxiv.org/abs/2312.09727>
 - [11] N. Duklan, S. Kumar, H. Maheshwari, R. Singh, S. D. Sharma, and S. Swami, "CNN Architectures for Image Classification: A Comparative Study Using ResNet50V2, ResNet152V2, InceptionV3, Xception, and MobileNetV2," *SSRG International Journal of Electronics and Communication Engineering*, vol. 11, no. 9, pp. 11–21, Sep. 2024, doi: 10.14445/23488549/IJECE-V11I9P102.
 - [12] D. Xiao, F. Zhu, J. Jiang, and X. Niu, "Leveraging natural cognitive systems in conjunction with ResNet50-BiGRU model and attention mechanism for enhanced medical image analysis and sports injury prediction," *Front. Neurosci.*, vol. 17, 2023, doi: 10.3389/fnins.2023.1273931.
 - [13] K. T. Bibish Kumar, S. John, K. M. Muraleedharan, and R. K. Sunil Kumar, "Linguistically involved data-driven approach for Malayalam phoneme-to-viseme mapping," in *Applied Speech Processing*, Elsevier, 2021, pp. 117–145. doi: 10.1016/B978-0-12-823898-1.00003-5.
 - [14] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 2814–2822. doi: 10.1109/CVPRW50498.2020.00338.
 - [15] M. Ardiana, "Viseme Morphing and Text-to-Speech Integration for Indonesian News Broadcasting," *The Indonesian Journal of Computer Science*, vol. 14, no. 3, Jun. 2025, doi: 10.33022/ijcs.v14i3.4841.
 - [16] Sougatamoy Biswas, Asim Kumar Naskar, and Anup Nandy, "MediaPipe with LSTM Architecture for Real-Time Hand Gesture Recognition ☆," 2023.
 - [17] J. Stephen and J. MK, "A Survey on Multilingual End-to-End Automatic Continuous Speech Recognition Using Deep Neural Network," 2023. [Online]. Available: www.ijrti.org
 - [18] M. Maskur and A. Zahra, "The Impact of Data Augmentation Techniques on Improving Speech Recognition Performance for English in Indonesian Children Based on Wav2Vec 2.0," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 19, no. 2, pp. 165–176, 2025, doi: 10.22146/ijccs.104646.
 - [19] S. Studer *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, Jun. 2021, doi: 10.3390/make3020020.
 - [20] A. Wangikar, A. Khavale, S. Sharma, U. Gadekar, and U. Sinha, "Ai-Powered Lip Reading Assistant For Individuals With Hearing Impairment," 2025, [Online]. Available: www.ijrt.org
 - [21] N. Sarangpure, V. Dhamde, A. Roge, J. Doye, S. Patle, and S. Tamboli, "Automating the Machine Learning Process using PyCaret and Streamlit," in *2023 2nd International Conference for Innovation in Technology (INOCON)*, IEEE, Mar. 2023, pp. 1–5. doi: 10.1109/INOCON57975.2023.10101357.
 - [22] L. Poomhiran, P. Meesad, and S. Nuanmeesri, "Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique," 2021. [Online]. Available: www.etasr.com
 - [23] I. I. Panagos, G. Sfikas, and C. Nikou, "Lightweight Operations for Visual Speech

Recognition,” Feb. 2025, [Online]. Available:
<http://arxiv.org/abs/2502.04834>