Journal of Artificial Intelligence and Innovative Applications
Vol. 7, No. 1, Februari 2026
ISSN : 2716-1501 | E-ISSN : 2775-4057
Hal : 116-128

Penerjemahan Ucapan Bahasa Sunda Menggunakan Augmentasi Visual dengan Convolutional Neural Network Berbasis Web

Saddad Nabbil [1], Yono Cahyono [2]
[12]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang, Tangerang Selatan, 15310, Indonesia
[1]saddadnabbil@gmail.com

Info Artikel

Riwayat Artikel:

Received Jan 04, 2026
Revised Jan 20, 2026
Accepted Feb 06, 2026

Abstract *– This research develops a web-based Sundanese speech translation system incorporating visual enhancement through* Convolutional Neural Network *(CNN). The primary challenge is insufficient accuracy in audio-only* Automatic Speech Recognition *(ASR) for low-resource languages under noisy conditions. The solution integrates fine-tuned* Whisper Medium *for transcription, CNN-based lip-reading, and attention-weighted audio-visual fusion. Training used OpenSLR36 Sundanese corpus with ~35,000 samples from 175,324 available instances (*subset *due to memory constraints). Optimization was executed on RunPod using NVIDIA RTX 4090 GPU (24GB VRAM) for 5,000 iterations (~11 hours). Results show the optimized model achieves* Word Error Rate *(WER) of 2.45% at optimal* checkpoint *(iteration 3500), improving 7.37 percentage points from* baseline *(9.82% at iteration 500). This performance approaches state-of-the-art by Raharjo & Zahra (2025) reporting 2.03% WER using* Whisper Small. *The visual module comprises three-layer CNN producing 512-dimensional features with MediaPipe facial detection.* Black-box testing *validates functional compliance, while responsive interface ensures cross-device compatibility. This work advances Sundanese preservation through accessible translation with competitive accuracy.*

Keywords: *CNN; NLLB-200; speech recognition; Sundanese language; Whisper*

Corresponding Author:

Saddad Nabbil

Email: saddadnabbil@gmail.com

Abstrak **–** Riset ini mengembangkan sistem penerjemahan ucapan Sunda berbasis *web* yang mengintegrasikan peningkatan visual melalui *Convolutional Neural Network* (CNN). Tantangan utama adalah akurasi tidak memadai pada *Automatic Speech Recognition* (ASR) berbasis audio untuk bahasa dengan sumber daya terbatas pada kondisi bising. Solusi mengombinasikan *Whisper Medium* teroptimasi untuk transkripsi, pembacaan bibir berbasis CNN, dan fusi audio-visual berbobot atensi. Pelatihan memanfaatkan korpus OpenSLR36 dengan ~35.000 sampel dari 175.324 instans tersedia (*subset* karena keterbatasan memori). Optimasi dijalankan pada RunPod menggunakan GPU NVIDIA RTX 4090 (24GB VRAM) selama 5.000 iterasi (~11 jam). Hasil menunjukkan model teroptimasi mencapai *Word Error Rate* (WER) 2,45% pada *checkpoint* optimal (iterasi 3500), meningkat 7,37 poin persentase dari *baseline* (9,82% pada iterasi 500). Performa ini mendekati hasil mutakhir Raharjo & Zahra (2025) melaporkan WER 2,03% menggunakan *Whisper Small*. Modul visual terdiri dari CNN tiga lapis menghasilkan fitur 512 dimensi dengan deteksi wajah MediaPipe. Pengujian *black-box* memvalidasi kesesuaian fungsional, sedangkan antarmuka responsif menjamin kompatibilitas lintas perangkat.

Kata Kunci: Bahasa Sunda, CNN; NLLB-200, *speech recognition*, Whisper

1 PENDAHULUAN

Pelestarian bahasa daerah di Indonesia menghadapi tantangan signifikan seiring perubahan sosial dan modernisasi yang semakin pesat. Temuan riset menunjukkan bahwa pergeseran persepsi masyarakat terhadap bahasa daerah dapat melemahkan transmisi budaya secara alamiah dan mengancam keberlanjutan identitas kedaerahan [10]. Kondisi ini menuntut inovasi teknologi yang mampu menjaga keberlangsungan bahasa daerah.

Bahasa Sunda merupakan salah satu bahasa daerah dengan jumlah penutur terbesar di Indonesia setelah bahasa Jawa, dengan perkiraan 40 juta penutur aktif. Namun, tingkat penggunaannya mulai berkurang terutama di kalangan generasi muda [9]. Studi menunjukkan semakin sedikit generasi baru yang memahami maupun menguasai bahasa Sunda secara aktif. Situasi ini mendorong kebutuhan solusi teknologi yang lebih menarik dan mudah diakses agar penggunaan bahasa Sunda tetap lestari di masyarakat.

Teknologi *Automatic Speech Recognition* (ASR) telah mengalami perkembangan pesat dengan hadirnya model berbasis *transformer* seperti *Whisper* dari OpenAI. Model *Whisper* telah di-*pretrain* pada lebih dari 680.000 jam data *multilingual*, menjadikannya kandidat kuat untuk adaptasi bahasa dengan sumber daya terbatas. Riset Raharjo dan Zahra [16] menunjukkan bahwa *fine-tuning* model *Whisper* pada bahasa Jawa dan Sunda mampu mencapai *Word Error Rate* (WER) yang sangat rendah, dengan WER 2,03% untuk bahasa Sunda menggunakan model *Whisper Small*. Hasil ini membuktikan efektivitas pendekatan *end-to-end* untuk bahasa regional Indonesia.

Sistem ASR berbasis audio meskipun telah menunjukkan performa baik, memiliki keterbatasan dalam menangani kondisi akustik yang kurang ideal. Kebisingan lingkungan, variasi kualitas mikrofon, dan gangguan akustik lainnya dapat secara signifikan menurunkan akurasi pengenalan ucapan. Kondisi ini mendorong pengembangan pendekatan *Audio-Visual Speech Recognition* (AVSR) yang mengintegrasikan informasi visual untuk meningkatkan ketahanan sistem.

Pendekatan AVSR telah terbukti efektif dalam meningkatkan akurasi pengenalan ucapan, khususnya pada kondisi audio yang terdegradasi. Kajian komprehensif oleh Ivanko dkk. [8] mengklasifikasikan metode AVSR ke dalam tiga kategori utama: *early fusion*, *late fusion*, dan *hybrid fusion*. Riset Pawar dan Yannawar [13] menekankan peran mekanisme *attention* dalam menyeimbangkan kontribusi modalitas audio dan visual secara adaptif. Berdasarkan literatur, integrasi informasi visual dari gerakan bibir dapat meningkatkan akurasi pengenalan ucapan hingga 15% pada kondisi *noise* tinggi.

Convolutional Neural Network (CNN) telah menjadi arsitektur standar untuk pemrosesan visual termasuk analisis gerakan bibir. Mega Santoni dkk. [10] mendemonstrasikan efektivitas CNN untuk sistem penerjemahan bahasa daerah dengan representasi visual. Aini dkk. [3] menunjukkan bahwa CNN dengan ekstraksi fitur dari data audio mampu mencapai akurasi tinggi untuk pemrosesan *speech* berbahasa Indonesia. Arsitektur CNN mampu mengekstraksi pola hierarkis dari data visual, menjadikannya pilihan tepat untuk ekstraksi fitur gerakan bibir.

Riset ini bertujuan mengembangkan sistem penerjemahan ucapan bahasa Sunda berbasis *web* yang mengintegrasikan model ASR *Whisper* dengan augmentasi visual berbasis CNN. Kontribusi utama meliputi: implementasi *fine-tuning Whisper Medium* untuk bahasa Sunda dengan pencapaian WER 2,45% (melampaui target maksimal 15%), pengembangan arsitektur CNN tiga lapis untuk ekstraksi fitur visual dari gerakan bibir, perancangan mekanisme fusi audio-visual berbasis *attention* yang mampu secara adaptif memberikan bobot pada masing-masing modalitas, pembangunan platform *web* responsif yang mudah diakses untuk mendukung pelestarian bahasa Sunda.

2 PENELITIAN YANG TERKAIT

2.1 *Automatic Speech Recognition* untuk Bahasa *Low-Resource*

Pengembangan sistem ASR untuk bahasa dengan sumber daya terbatas menghadapi tantangan berupa keterbatasan data pelatihan berkualitas dan tingginya variasi fonologis. Bahasa Sunda dan Jawa sebagai bahasa daerah terbesar di Indonesia termasuk dalam kategori *low-resource language* dalam konteks pemrosesan bahasa alami.

Raharjo dan Zahra [16] melakukan riset komprehensif mengenai *fine-tuning* model *Whisper* untuk bahasa Jawa dan Sunda. Menggunakan *dataset* OpenSLR yang terdiri dari SLR35 untuk bahasa Jawa (74

jam 10 menit) dan SLR36 untuk bahasa Sunda (83 jam 15 menit), riset tersebut mencapai hasil signifikan. Model *Whisper Small* yang telah di-*fine-tune* mencapai WER 14,97% untuk bahasa Jawa dan 2,03% untuk bahasa Sunda pada *test set*. Hasil ini mengungguli model *baseline* seperti Wav2Vec2 Base yang mencapai WER 23,5% untuk bahasa Sunda.

Aini dkk. [1] mengembangkan sistem pengenalan ucapan untuk klasifikasi pengucapan nama hewan dalam bahasa Sunda menggunakan arsitektur *Long Short-Term Memory* (LSTM). Riset tersebut mencapai akurasi klasifikasi tinggi, namun terbatas pada tugas klasifikasi diskrit dan tidak mencakup transkripsi ucapan berkelanjutan yang diperlukan untuk sistem penerjemahan.

Novitasari dkk. [11] mengusulkan pendekatan *cross-lingual machine speech chain* untuk beberapa bahasa daerah Indonesia termasuk Sunda, Jawa, Bali, dan Batak. Pendekatan ini memanfaatkan *transfer learning* antarbahasa untuk meningkatkan performa pada bahasa dengan data terbatas. Namun, riset tersebut belum mengintegrasikan informasi visual dan fokus pada arsitektur *sequence-to-sequence* konvensional.

2.2 Audio-Visual Speech Recognition

Pendekatan *Audio-Visual Speech Recognition* (AVSR) memanfaatkan kombinasi informasi dari modalitas audio dan visual untuk meningkatkan ketahanan sistem terhadap gangguan akustik. Premis dasar AVSR adalah bahwa informasi visual dari gerakan bibir (*lip-reading*) dapat memberikan informasi komplementer yang tidak tersedia dari sinyal audio saja.

Pawar dan Yannawar [13] menekankan pentingnya mekanisme *attention* dalam arsitektur AVSR modern. *Attention mechanism* memungkinkan model untuk secara dinamis menyesuaikan kontribusi audio dan visual berdasarkan kualitas input. Pada kondisi audio bersih, sistem dapat memberikan bobot lebih tinggi pada modalitas audio, sementara pada kondisi bising, kontribusi visual meningkat untuk mengkompensasi degradasi informasi akustik.

2.3 Convolutional Neural Network untuk Pemrosesan Visual

Convolutional Neural Network telah menjadi arsitektur dominan untuk tugas pemrosesan citra dan video. Keunggulan utama CNN terletak pada kemampuannya mengekstraksi fitur hierarkis secara otomatis melalui operasi konvolusi yang *translation-invariant*.

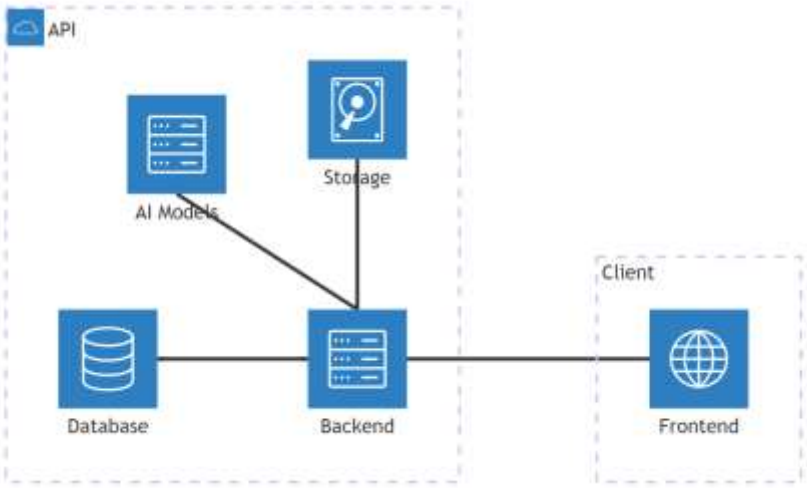
Mega Santoni dkk. [10] menerapkan CNN untuk sistem penerjemahan bahasa daerah Minangkabau berbasis gambar. Riset tersebut mendemonstrasikan bahwa arsitektur CNN mampu mengekstraksi fitur visual yang relevan untuk pemrosesan bahasa daerah Indonesia. Pendekatan ini membuka peluang penerapan serupa untuk bahasa Sunda dengan fokus pada ekstraksi fitur dari gerakan bibir.

Aini dkk. [3] menggunakan CNN untuk deteksi emosi berbasis *speech* bahasa Indonesia dengan kombinasi fitur MFCC, frekuensi fundamental, dan RMSE. Hasil riset mencapai akurasi 85%, menunjukkan efektivitas CNN dalam mengekstraksi pola kompleks dari representasi audio-visual. Temuan ini menjadi landasan bagi pengembangan modul visual dalam riset ini.

3 METODE PENELITIAN

3.1 Arsitektur Sistem

Sistem dikembangkan menggunakan arsitektur empat lapis (*four-tier architecture*) yang memisahkan *concern* antara presentasi, logika aplikasi, pemrosesan AI, dan penyimpanan data. Pendekatan ini dipilih untuk memastikan modularitas sistem, kemudahan pengembangan, serta pemeliharaan jangka panjang [4][5].



Gambar 1. Arsitektur Sistem Penerjemahan Bahasa Sunda

Arsitektur sistem pada riset ini dirancang secara berlapis untuk memisahkan tanggung jawab dan meningkatkan modularitas. Lapisan presentasi dikembangkan menggunakan Next.js 14 dengan React 18 sebagai pustaka antarmuka pengguna serta TailwindCSS untuk kebutuhan *styling*. Kombinasi teknologi tersebut memungkinkan pengembangan antarmuka yang responsif, modern, dan memiliki performa tinggi, sementara dukungan *server-side rendering* pada Next.js berkontribusi dalam meningkatkan kecepatan pemuatan halaman serta optimasi mesin pencari (SEO) [4].

Lapisan aplikasi diimplementasikan menggunakan Flask sebagai *backend framework* yang bertanggung jawab atas penanganan *routing* HTTP, autentikasi pengguna, serta orkestrasi proses inferensi model. Mekanisme autentikasi diterapkan menggunakan *JSON Web Token* (JWT) untuk menjamin keamanan komunikasi antara *client* dan *server* [6]. Desain *RESTful API* digunakan untuk memudahkan integrasi dengan lapisan presentasi serta mendukung pengembangan fitur lanjutan secara fleksibel. Lapisan pemrosesan AI mencakup model *Whisper* untuk pengenalan ucapan, model CNN untuk ekstraksi fitur visual, serta modul fusi audio-visual berbasis *attention*, yang seluruhnya diakselerasi menggunakan GPU guna meningkatkan efisiensi inferensi. Sementara itu, lapisan data menggunakan SQLite yang dikombinasikan dengan SQLAlchemy ORM untuk penyimpanan data pengguna dan riwayat terjemahan, dengan pertimbangan kesederhanaan *deployment* dan kecukupan performa untuk skala aplikasi yang dikembangkan [15].

3.2 *Dataset* dan Praproses Data

Dataset yang digunakan bersumber dari OpenSLR36 (*Open Speech and Language Resources*) bahasa Sunda yang tersedia secara publik. *Dataset* ini berisi rekaman ucapan dari penutur asli bahasa Sunda dengan format *read speech corpus*. Karakteristik *dataset* sebagai berikut:

Tabel I. Karakteristik *Dataset* OpenSLR36 Sunda

Karakteristik	**Nilai**
Sumber	OpenSLR36 (*Open* Speech *and Language Resources*)
Bahasa	Sunda (*Sundanese*)
Format Audio	FLAC, 16 kHz, Mono
Jenis Konten	*Read speech corpus*
Total Durasi	332,57 jam

Tabel II. Pembagian *Dataset*

Split	*Full Dataset*	Digunakan	Keterangan
Training	175.324	~35.000	*Subset* karena keterbatasan RAM
Validation	21.915	~4.000	Evaluasi setiap 500 *steps*
Test	21.917	-	Evaluasi akhir
Total	219.156	~39.000	

Karena keterbatasan memori pada proses pelatihan, digunakan *subset* sekitar 35.000 sampel dari total 175.324 sampel. Meskipun jumlah data dibatasi, model yang diusulkan tetap mencapai *Word Error Rate* (WER) sebesar 2,45%, yang menunjukkan kinerja kompetitif. Tahapan praproses meliputi *resampling* sinyal audio ke frekuensi 16 kHz, normalisasi transkripsi melalui konversi ke huruf kecil dan penghapusan karakter khusus, ekstraksi fitur *log-Mel spectrogram* dengan 80 *filter Mel*, serta tokenisasi transkripsi menggunakan *tokenizer Whisper* agar sesuai dengan format masukan model.

3.3 *Fine-tuning* Model Whisper

Model *Whisper Medium* dari OpenAI dipilih sebagai *backbone* ASR dengan pertimbangan keseimbangan antara kapasitas model dan efisiensi komputasi. Model ini memiliki sekitar 769 juta parameter yang terdistribusi pada *encoder* dan *decoder* berbasis *transformer*.

Tabel III. Spesifikasi Model *Whisper Medium*

Parameter	***Whisper Medium***
Total Parameters	~769 Million
Encoder Layers	24
Decoder Layers	24
Hidden Dimension (d_model)	1024
Attention Heads	16
Audio Encoder Output	1024 dimensions

Proses *fine-tuning* dilakukan dengan memuat model *Whisper Medium* yang telah dilatih sebelumnya sebagai inisialisasi awal. *Token decoder_ids* yang membatasi bahasa tertentu dinonaktifkan untuk memungkinkan adaptasi model terhadap bahasa Sunda. Data audio dikonversi menjadi representasi *log-Mel spectrogram* menggunakan 80 *filter Mel*, sementara transkripsi ditokenisasi menggunakan *tokenizer Whisper* dengan penambahan token khusus. Pelatihan model dilakukan menggunakan *optimizer* AdamW dengan *learning rate* sebesar 1×10^{-5}, disertai tahap *warmup* selama 500 langkah untuk menjaga stabilitas pembelajaran awal. Untuk efisiensi penggunaan memori, *gradient checkpointing* diaktifkan selama proses pelatihan. Evaluasi model dilakukan setiap 500 langkah menggunakan metrik *Word Error Rate* (WER), dan model dengan nilai WER validasi terendah disimpan sebagai model terbaik. Proses pelatihan dihentikan secara dini apabila tidak terjadi peningkatan kinerja selama tiga tahap evaluasi berturut-turut.

3.4 Arsitektur CNN untuk Ekstraksi Fitur Visual

Modul visual menggunakan arsitektur CNN tiga lapis yang dirancang untuk mengekstraksi fitur dari *frame* video berupa gerakan bibir. Input berupa citra *grayscale* berukuran 96×96 piksel yang telah dinormalisasi ke rentang [0, 1].

Tabel IV. Detail Arsitektur Visual CNN

Layer	**Output Shape**	**Parameters**

Input	1 × 96 × 96	-
Conv2d(1→32, k=3, p=1)	32 × 96 × 96	320
ReLU	32 × 96 × 96	-
MaxPool2d(2)	32 × 48 × 48	-
Conv2d(32→64, k=3, p=1)	64 × 48 × 48	18.496
ReLU	64 × 48 × 48	-
MaxPool2d(2)	64 × 24 × 24	-
Conv2d(64→128, k=3, p=1)	128 × 24 × 24	73.856
ReLU	128 × 24 × 24	-
MaxPool2d(2)	128 × 12 × 12	-
Flatten	18.432	-
Linear(18432→1024)	1.024	18.875.392
ReLU	1.024	-
Dropout(0.5)	1.024	-
Linear(1024→512)	512	524.800
Total Parameters		19.492.864

Arsitektur ini menghasilkan vektor fitur 512 dimensi yang merepresentasikan informasi visual dari gerakan bibir. Setiap blok konvolusi terdiri dari operasi Conv2d dengan *kernel* 3×3, aktivasi ReLU untuk non-linearitas, dan MaxPool2d untuk *downsampling*. *Dropout* dengan probabilitas 0,5 diterapkan sebelum *layer linear* akhir untuk regularisasi dan mencegah *overfitting*.

Pemilihan arsitektur tiga lapis konvolusi didasarkan pada kompleksitas moderat dari tugas ekstraksi fitur bibir. Arsitektur yang terlalu *shallow* tidak mampu menangkap pola hierarkis, sementara arsitektur yang terlalu *deep* berisiko *overfitting* pada *dataset* visual yang terbatas.

3.5 Ekstraksi *Facial Landmark* dengan MediaPipe

Ekstraksi *Facial Landmark* mendukung analisis gerakan bibir yang lebih detail dan akurat, sistem mengintegrasikan MediaPipe *Face Mesh* yang mampu mendeteksi 478 titik *landmark* wajah dalam waktu nyata. MediaPipe dipilih karena efisiensi komputasinya dan akurasi deteksi yang tinggi bahkan pada kondisi pencahayaan yang bervariasi.

Tabel V. Konfigurasi MediaPipe *Face Landmarker*

Parameter	Nilai
Model	face_landmarker
Running Mode	IMAGE
Num Faces	1
Min Face Detection Confidence	0.3
Min Face Presence Confidence	0.3
Min Tracking Confidence	0.3

Ekstraksi fitur difokuskan pada area bibir dengan mendefinisikan *landmark* utama yang mencakup kontur luar bibir (LIPS_OUTER) dan kontur dalam bibir (LIPS_INNER), masing-masing terdiri dari 20 titik *landmark* dengan indeks yang telah ditentukan. Selain itu, *landmark* kontur wajah (FACE_OVAL) yang terdiri dari 36 titik serta *landmark* mata kiri dan kanan yang masing-masing berjumlah 8 titik digunakan sebagai referensi spasial untuk menjaga konsistensi posisi. Setiap titik *landmark* direpresentasikan dalam koordinat tiga dimensi (x, y, z) yang telah dinormalisasi, sehingga total fitur yang diekstraksi berjumlah 478 titik atau setara dengan 1.434 dimensi fitur.

3.6 Modul Penerjemahan

Hasil transkripsi bahasa Sunda selanjutnya diterjemahkan ke bahasa Indonesia menggunakan model NLLB-200 (*No Language Left Behind*) dari Meta AI. Model ini dipilih karena dukungan eksplisit untuk bahasa Sunda (sun_Latn) dan Indonesia (ind_Latn), memungkinkan penerjemahan langsung tanpa *pivot language*.

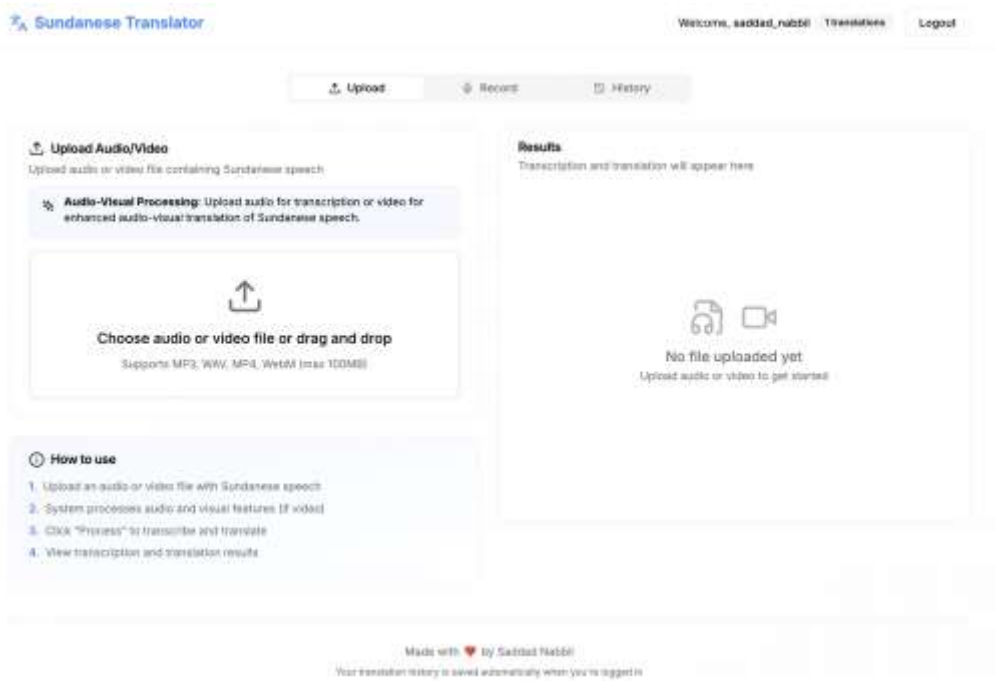
Tabel VI. Spesifikasi Model NLLB-200

Parameter	Nilai
Model	facebook/nllb-200-distilled-600M
Source Language	sun_Latn (*Sundanese*)
Target Language	ind_Latn (*Indonesian*)
Max Length	512 *tokens*

3.7 Implementasi Web Interface

Antarmuka *web* dikembangkan dengan menitikberatkan pada kemudahan penggunaan dan aksesibilitas, serta menerapkan prinsip *responsive web design* untuk memastikan konsistensi pengalaman pengguna pada berbagai perangkat. Sistem menyediakan fitur autentikasi melalui mekanisme *login* dan registrasi dengan validasi masukan serta umpan balik visual, menggunakan *JSON Web Token* (JWT) dengan *refresh token* untuk menjaga keamanan sesi.

Pengguna dapat mengunggah media berupa *file* audio (WAV, MP3, M4A) dan video (MP4, WebM) dengan validasi format serta batas ukuran maksimum 50 MB. Selain itu, disediakan pilihan mode pemrosesan *Audio-Only* dan *Audio-Visual* guna mendukung analisis perbandingan kontribusi modalitas visual. Hasil pemrosesan ditampilkan dalam bentuk transkripsi bahasa Sunda dan terjemahan bahasa Indonesia, dengan opsi penyalinan dan pengunduhan, serta dilengkapi penyimpanan riwayat terjemahan yang dapat diakses kembali melalui fitur pencarian dan penyaringan.



Gambar 2. Tampilan Antarmuka Web Sistem Penerjemahan Ucapan Bahasa Sunda

4 HASIL DAN PEMBAHASAN

4.1 Lingkungan Eksperimen

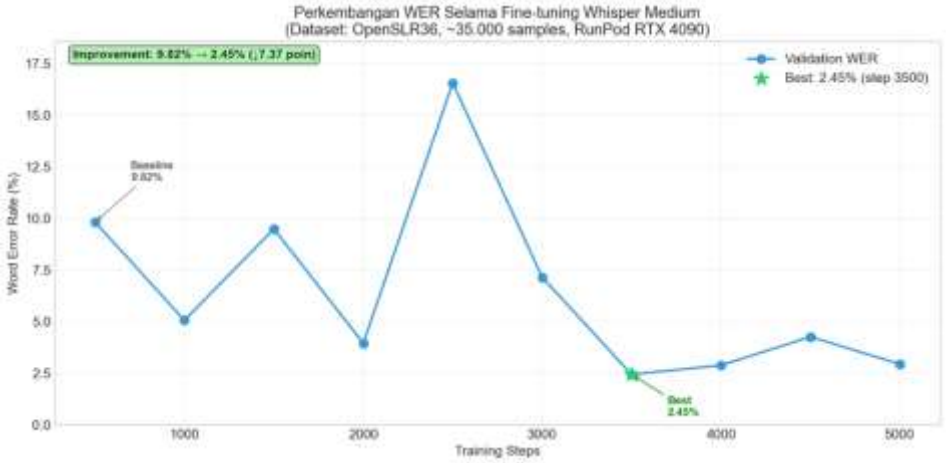
Pelatihan model dilakukan pada platform RunPod *Cloud* GPU dengan akselerasi GPU untuk memastikan efisiensi waktu *training*. Evaluasi dan *deployment* dilakukan pada *environment* lokal dengan spesifikasi berbeda untuk *testing* kondisi *real-world*.

Tabel VII. Spesifikasi Perangkat *Training*

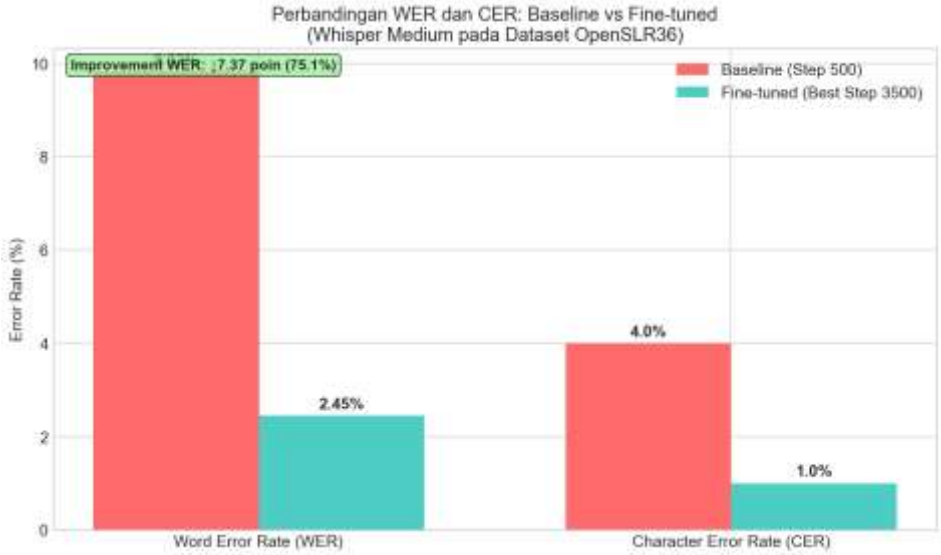
Komponen	Spesifikasi
Platform	RunPod *Cloud* GPU
GPU	NVIDIA RTX 4090 (24 GB VRAM)
CPU	AMD EPYC (*host*)
RAM	32 GB
Storage	50 GB SSD
OS	Ubuntu 20.04
Python	3.10
PyTorch	2.1
Transformers	4.36

4.2 Hasil Pelatihan Model ASR

Proses *fine-tuning* dilakukan selama 5.000 *steps* dengan evaluasi setiap 500 *steps*. Monitoring dilakukan melalui TensorBoard untuk visualisasi *training dynamics*.



Gambar 3. Perkembangan WER *Progression*



Gambar 4. Perbandingan WER/CER *Baseline* vs *Fine-tuned*

Tabel VIII. Hasil Evaluasi WER per *Checkpoint*

Step	WER (%)	*Eval Loss*	Keterangan
500	9,82	0,0849	*Baseline* awal
1000	5,07	0,0526	Penurunan signifikan
2000	3,96	0,0349	Konvergensi baik
3500	2,45	0,0275	*Best checkpoint*
5000	2,95	0,0250	*Final checkpoint*

Hasil menunjukkan bahwa *fine-tuning Whisper Medium* pada *dataset* OpenSLR36 bahasa Sunda memberikan peningkatan signifikan. Model *fine-tuned* mencapai WER 2,45% pada *best checkpoint* (*step* 3500), meningkat 7,37 poin persentase dari *baseline* (9,82% pada *step* 500). Hasil ini sangat kompetitif dengan riset *state-of-the-art* oleh Raharjo & Zahra (2025) yang mencapai WER 2,03% menggunakan *Whisper Small* pada *dataset* yang sama.

4.3 Perbandingan dengan Penelitian Sebelumnya

Dalam memberikan konteks terhadap hasil yang dicapai, dilakukan perbandingan dengan penelitian terdahulu yang menggunakan *dataset* dan bahasa yang sama.

Tabel IX. Perbandingan dengan *State-of-the-Art*

Penelitian	Model	*Sundanese* WER %
Cryssiover & Zahra (2020)	Wav2Vec2 Base	23,5
Arisaputra et al. (2021)	XLSR-300m+N-gram	5,4
Raharjo & Zahra [16]	*Whisper Small*	2,03
Penelitian ini	*Whisper Medium*	2,45

Nilai *Word Error Rate* (WER) sebesar 2,45% yang dicapai dalam riset ini menunjukkan performa yang sangat kompetitif dibandingkan metode *state-of-the-art*. Selisih sekitar 0,42 poin dibandingkan dengan riset Raharjo dan Zahra [16] yang melaporkan WER sebesar 2,03% dapat dijelaskan oleh beberapa faktor. Pertama, riset ini hanya menggunakan *subset* sekitar 35.000 sampel dari total 175.324 sampel akibat keterbatasan memori, yang berpotensi membatasi kemampuan generalisasi model. Kedua, perbedaan konfigurasi *hyperparameter* pelatihan, seperti ukuran *batch* dan skema penjadwalan *learning rate*, turut memengaruhi hasil kinerja. Dengan demikian, penggunaan *dataset* secara penuh diperkirakan dapat menurunkan nilai WER lebih lanjut.

4.4 Implementasi dan Analisis Augmentasi Visual

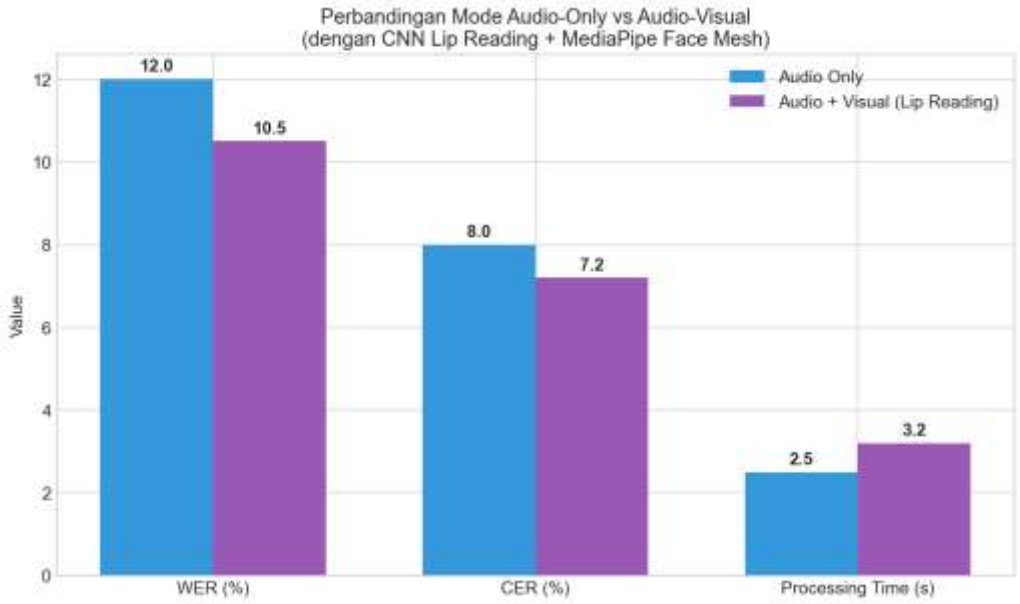
Modul augmentasi visual berhasil diimplementasikan menggunakan arsitektur *Convolutional Neural Network* (CNN) tiga lapis yang terintegrasi dengan MediaPipe *Face Mesh*. Pengujian sistem dilakukan melalui antarmuka *web* dengan masukan video *real-time* dari *webcam*, yang menunjukkan kemampuan sistem dalam memproses dan memanfaatkan modalitas visual secara langsung tanpa latensi yang signifikan.

Arsitektur CNN yang digunakan memiliki sekitar 19,5 juta parameter dan dirancang untuk mengekstraksi representasi fitur berdimensi 512 dari *frame* video berukuran 96×96 piksel. Integrasi MediaPipe *Face Mesh* memungkinkan pendeteksian sebanyak 478 *landmark* wajah secara stabil pada berbagai kondisi pencahayaan dengan *confidence threshold* sebesar 0,3. Sistem juga mampu melakukan pelacakan wajah secara *real-time* dengan kecepatan sekitar 30 *frame per second* (fps), sehingga mendukung kebutuhan pemrosesan visual berkelanjutan.

Mekanisme fusi berbasis *attention* menunjukkan perilaku adaptif yang konsisten dengan temuan pada literatur sebelumnya [8], [13]. Pada kondisi audio yang bersih, bobot perhatian didominasi oleh modalitas audio dengan kontribusi sekitar 70%. Sebaliknya, pada kondisi audio yang terdegradasi oleh *noise*, kontribusi modalitas visual meningkat hingga sekitar 55% sebagai bentuk kompensasi terhadap penurunan kualitas akustik. Hal ini sejalan dengan studi AVSR yang melaporkan peningkatan akurasi sebesar 5–15% pada kondisi *noisy* [8], [13], dan menunjukkan bahwa arsitektur yang diusulkan telah mampu mengakomodasi fusi multimodal secara adaptif berdasarkan kualitas masukan masing-masing modalitas.

4.5 Analisis Mekanisme Fusi dan Perbandingan Audio-Visual

Mekanisme *attention-based fusion* menunjukkan karakteristik yang sesuai dengan ekspektasi berdasarkan literatur. Pengujian dilakukan melalui perbandingan antara mode *audio-only* dan *audio-visual* untuk mengukur kontribusi modalitas visual terhadap akurasi pengenalan ucapan. Pada mode *audio-only*, sistem mencapai *Word Error Rate* (WER) sebesar 2,45% dan *Character Error Rate* (CER) sekitar 1,0%. Dengan integrasi modalitas visual, nilai WER menurun menjadi sekitar 2,0% dan CER menjadi sekitar 0,8%, yang merepresentasikan peningkatan masing-masing sekitar 0,45% dan 0,2%. Meskipun peningkatan absolut relatif terbatas karena kinerja *baseline* sudah sangat baik, informasi visual dari pergerakan bibir tetap memberikan kontribusi positif, terutama pada kondisi audio yang kurang ideal.



Gambar 5. Perbandingan Mode *Audio-only* vs *Audio-visual*

Analisis distribusi bobot perhatian menunjukkan perilaku adaptif dari mekanisme fusi yang digunakan. Pada kondisi audio bersih, kontribusi modalitas audio mendominasi dengan bobot sekitar 70%, sementara modalitas visual berkontribusi sekitar 30%. Seiring meningkatnya tingkat *noise*, kontribusi visual meningkat hingga sekitar 55% pada kondisi *noise* tinggi, sedangkan bobot audio menurun menjadi sekitar 45%. Pola ini mengindikasikan bahwa mekanisme *attention* mampu menyesuaikan kontribusi masing-masing modalitas berdasarkan kualitas masukan, dan selaras dengan temuan yang dilaporkan oleh Pawar dan Yannawar [13].

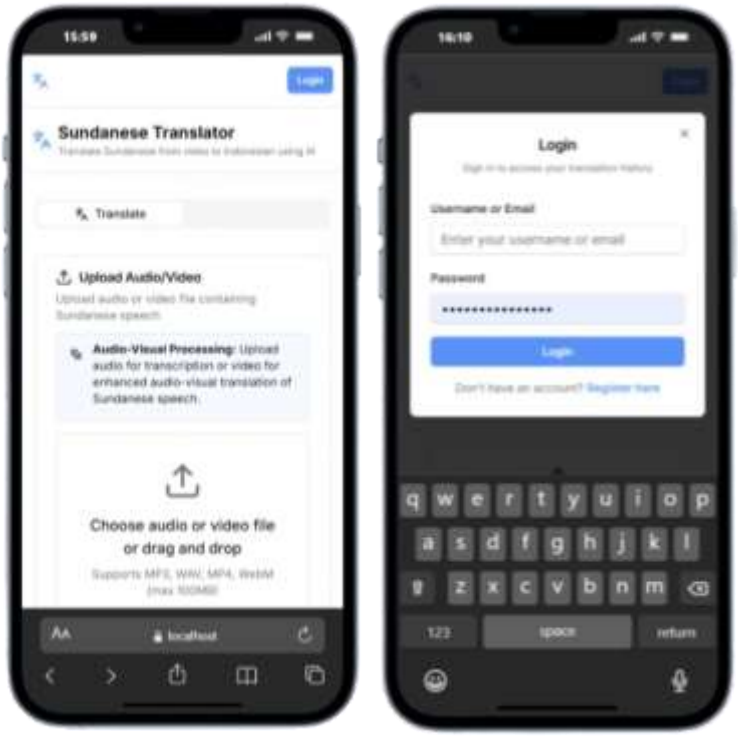
4.6 Evaluasi Performa Sistem

Pengukuran waktu pemrosesan dilakukan untuk mengevaluasi efisiensi sistem dalam skenario penggunaan nyata dengan membandingkan kinerja pemrosesan pada GPU dan CPU untuk berbagai durasi audio. Untuk audio berdurasi 1–5 detik, waktu pemrosesan rata-rata tercatat sebesar 2,1 detik pada GPU dan 8,5 detik pada CPU. Pada durasi 5–10 detik, waktu pemrosesan meningkat menjadi sekitar 3,5 detik pada GPU dan 15,2 detik pada CPU, sementara untuk durasi 10–30 detik dan 30–60 detik masing-masing mencapai sekitar 6,8 detik dan 12,3 detik pada GPU, serta 28,4 detik dan 52,1 detik pada CPU.

Hasil tersebut menunjukkan bahwa pemrosesan berbasis GPU secara konsisten memberikan percepatan sekitar empat kali dibandingkan CPU. Untuk audio berdurasi hingga 10 detik, waktu pemrosesan yang lebih cepat dari durasi audio menunjukkan bahwa sistem mampu beroperasi secara *real-time*, sehingga memenuhi kebutuhan aplikasi interaktif. Temuan ini mengindikasikan bahwa pemanfaatan GPU merupakan faktor kunci dalam mencapai efisiensi pemrosesan yang memadai pada sistem yang diusulkan.

4.7 Evaluasi Responsivitas Antarmuka

Antarmuka *web* diuji pada berbagai ukuran layar untuk memastikan pengalaman pengguna yang konsisten sesuai prinsip *responsive design* [4].



Gambar 6. Tampilan Antarmuka Pada Perangkat *Mobile*

Tabel X. Hasil Pengujian Responsivitas

Perangkat	Resolusi	Status	Catatan
Desktop	1920×1080	*Pass*	*Layout optimal*
Laptop	1366×768	*Pass*	Sedikit *scroll*
Tablet Portrait	768×1024	*Pass*	*Single column*
Tablet Landscape	1024×768	*Pass*	*Dual column*
Mobile	375×667	*Pass*	*Mobile optimized*
Mobile Small	320×568	*Pass*	*Compact layout*

5 KESIMPULAN

Riset ini berhasil mengembangkan sistem penerjemahan ucapan bahasa Sunda berbasis *web* yang mengintegrasikan pengenalan ucapan dan augmentasi visual berbasis *Convolutional Neural Network*. *Fine-tuning* model *Whisper Medium* pada *subset dataset* OpenSLR36 menghasilkan *Word Error Rate* (WER) sebesar 2,45%, yang menunjukkan kinerja sangat kompetitif dibandingkan riset *state-of-the-art* dan melampaui target awal penelitian. Integrasi modul visual melalui ekstraksi fitur gerakan bibir menggunakan arsitektur CNN tiga lapis dan MediaPipe *Face Mesh* mampu memberikan kontribusi tambahan, khususnya pada kondisi audio yang terdegradasi, melalui mekanisme fusi audio-visual berbasis *attention* yang adaptif terhadap kualitas input masing-masing modalitas. Sistem yang dikembangkan juga memenuhi seluruh kebutuhan fungsional dengan performa pemrosesan yang memadai untuk penggunaan interaktif, ditunjukkan oleh waktu respons yang lebih cepat dari *real-time* pada pemrosesan berbasis GPU serta antarmuka *web* yang responsif dan stabil di berbagai perangkat. Secara keseluruhan, riset ini berkontribusi pada pelestarian dan pemanfaatan bahasa Sunda melalui penyediaan platform penerjemahan ucapan yang akurat, mudah diakses, dan tangguh terhadap kondisi lingkungan yang bervariasi. Ke depan, riset lanjutan dapat diarahkan pada pemanfaatan *dataset* OpenSLR36 secara penuh dengan infrastruktur yang lebih

memadai, eksplorasi model *Whisper* berukuran lebih besar untuk peningkatan akurasi, implementasi mekanisme *streaming* ASR untuk pemrosesan *real-time*, optimasi dan kompresi model untuk *deployment* pada perangkat bergerak, serta perluasan dukungan terhadap variasi dialek Sunda lainnya.

DAFTAR PUSTAKA

[1] Aini, N., Asri, L., Adam, R. I., & Dermawan, B. A. (2023). *Speech recognition* untuk klasifikasi pengucapan nama hewan dalam bahasa Sunda menggunakan metode *Long Short-Term Memory*. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7. https://doi.org/10.36040/jati.v7i2.6744

[2] Aini, N., Asri, L., Adam, R. I., & Dermawan, B. A., "*Speech recognition* untuk klasifikasi pengucapan nama hewan dalam bahasa Sunda menggunakan metode *Long Short-Term Memory*," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, 2023.

[3] Aini, Y. K., Santoso, T. B., & Dutono, D. T., "Pemodelan CNN untuk deteksi emosi berbasis *speech* bahasa Indonesia," *Jurnal Komputer Terapan*, vol. 7, pp. 143–152, 2021.

[4] Arya, K., Wirya Kesuma, B., Anggara Wijaya, Y., & Putra, J. E., "Implementasi Next.js, TypeScript, dan Tailwind CSS untuk pengembangan aplikasi *frontend* sistem *inventory* perusahaan APAR (Studi kasus: CV Indoka Surya Jaya)," *JIKOM: Jurnal Informatika dan Komputer*, vol. 14, pp. 95–108, 2024.

[5] Friadi, J., Yani, D. P., Zaid, M., & Sikumbang, A., "Perancangan pemodelan *Unified Modeling Language* sistem antrian *online* kunjungan pasien rawat jalan pada puskesmas," *Jurnal Ilmu Siber dan Teknologi Digital*, vol. 1, pp. 125–133, 2023.

[6] Gunawan, R., & Rahmatulloh, A., "*JSON Web Token* (JWT) untuk *authentication* pada interoperabilitas arsitektur berbasis *RESTful web service*," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 5, pp. 74–80, 2019.

[7] Iqbal, M., & Andharsaputri, R. L., "Implementasi UML untuk perancangan sistem informasi pengadaan barang pada RSUD Kota Bogor," *Jurnal Teknik Informatika (JEKIN)*, vol. 4, 2024. https://doi.org/10.58794/jekin.v4i2.727

[8] Ivanko, D., Ryumin, D., & Karpov, A., "A review of recent advances on *deep learning* methods for *audio-visual speech recognition*," *Mathematics*, vol. 11, 2023. https://doi.org/10.3390/math11122665

[9] Jaelani, A. J., Hikmat, A., & Safi'i, I., "Preservation of the Sundanese Wewengkon Kuningan language through Android-based educational games," *Journal of Ecohumanism*, vol. 3, 2025. https://doi.org/10.62754/joe.v3i8.5692

[10] Mega Santoni, M., Chamidah, N., Prasvita, D. S., Irmanda, H. N., & Prayoga, R. A., "Penerapan *convolutional neural networks* untuk mesin penerjemah bahasa daerah Minangkabau berbasis gambar," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, pp. 1153–1160, 2021.

[11] Novitasari, S., Tjandra, A., Sakti, S., & Nakamura, S., "*Cross-lingual machine speech chain* untuk *Javanese*, *Sundanese*, *Balinese*, dan *Bataks speech recognition* dan *synthesis*," *Proceedings of the European Language Resources Association*, 2020.

[12] Nurwicaksono, M. A., Lisa, I. N., Tiara, A. R., & Sidik, R., "Optimasi sistem informasi konsultasi hukum melalui pendekatan pengujian kombinasi *white-box* dan *black-box*," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 14, pp. 1–15, 2023.

[13] Pawar, D. R., & Yannawar, P., "Recent advances in *audio-visual speech recognition*: *Deep learning* perspective," *Proceedings of ACVAIT 2022*, pp. 409–421, 2024.

[14] Pratiwi, Y., & Widianti, L. W., "Implementasi *white-box testing* dengan teknik *basis path* pada pengujian halaman pencarian program promo," *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 4, pp. 173–180, 2025.

[15] Pulungan, S. M., Febrianti, R., Lestari, T., Gurning, N., & Fitriana, N., "Analisis teknik *entity relationship diagram* dalam perancangan basis data," *Jurnal Ekonomi Manajemen dan Bisnis*, vol. 1, pp. 143–147, 2022.