Pembacaan Gerak Bibir Menggunakan CNN, Bi-LSTM dan CTC *Loss Function* pada Dataset Bahasa Inggris

Mahesa Tirta Panjalu[1] dan Lu'lu Mawaddah Wisudawati[2]

[1]Teknik Informatika, Universitas Gunadarma
[2]Teknologi Industri, Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
Email : mahesat8@gmail.com, lulu_mawadah@staff.gunadarma.ac.id*)

Abstrak

Tunarungu adalah orang yang mengalami gangguan pendengaran. Dampak utama dari kondisi ini adalah hambatan dalam komunikasi verbal atau lisan, sehingga menyulitkan komunikasi dengan orang yang mendengar. Bagian bibir adalah bagian yang biasa digunakan untuk berbicara atau berkomunikasi. Gerakan bibir saat berkomunikasi akan menghasilkan gerakan yang berbeda-beda setiap kata atau huruf yang diucapkan. Bibir dapat digunakan untuk memprediksi kata dari gerak bibir yang akan terdeteksi saat berbicara. Teknologi yang semakin berkembang dapat membantu permasalahan tersebut dalam membaca gerak bibir. *Convolutional Neural Network* atau CNN telah berkembang pesat dan menjadi salah satu metode yang paling populer dalam bidang pengenalan citra dan pemrosesan video karena kemampuannya untuk secara otomatis mempelajari fitur dari data masukan. Penelitian ini bertujuan melakukan pembacaan gerak bibir menggunakan metode CNN, *Long Short-Term Memory* (LSTM) dan *Connectionist Temporal Classification* (CTC) dalam bahasa inggris. Penelitian ini menggunakan dataset dari *The Grid audiovisual sentence corpus* sebanyak 1000 video dan 1000 teks. Pada tahapan preprocessing terdiri dari dua bagian yaitu *preprocessing* video dan *preprocessing* teks. Tahapan *preprocessing* video meliputi konversi *grayscale, cropping frame, augmentas*i dan normalisasi. Tahapan preprocessing teks dilakukan proses encoding pada dataset alignments. Tahapan klasifikasi menggunakan metode Convolutional Neural Networks, Long Short-Term Memory dan *Connectionist Temporal Classification Loss Function*. Hasil evaluasi mendapatkan nilai akurasi terbaik sebesar 96,9%, *Word Error Rate* (WER) sebesar 0,66%, dan *Character Error Rate* (CER) sebesar 0,16% dengan menggunakan model yang dengan skenario data 80:20 dan batch size 2.

Kata kunci : LipNet, CNN, LSTM, CTC

Pendahuluan

Tunarungu adalah kondisi yang mempengaruhi kemampuan seseorang untuk mendengar sebagian atau sepenuhnya, yang dapat mengakibatkan kesulitan dalam berkomunikasi dengan orang lain. Dampak langsung dari ketunarunguan adalah terhambatnya komunikasi verbal/lisan, baik secara ekspresif (berbicara) maupun reseptif (memahami pembicaraan orang lain), sehingga sulit berkomunikasi dengan lingkungan orang mendengar yang lazim menggunakan bahasa verbal sebagai alat komunikasi [1]. Bagi individu tunarungu, berkomunikasi menggunakan bahasa isyarat dan membaca gerakan bibir adalah metode utama untuk berinteraksi. Namun, membaca gerakan bibir memerlukan keahlian khusus dan dapat dipengaruhi oleh faktor seperti pencahayaan dan sudut pandang.

Dalam era digital saat ini, teknologi pengenalan ucapan telah menjadi semakin penting dalam berbagai aplikasi, termasuk dalam pengenalan kata-kata dari gerakan bibir. Teknik-teknik tradisional untuk pengenalan ucapan sering kali mengalami batasan dalam menangani variasi akustik dan lingkungan. Oleh karena itu, penggunaan teknik deep learning, seperti *Convolutional Neural Networks* (CNN) dan L*ong Short-Term Memory* (LSTM), telah menarik minat sebagai pendekatan yang lebih efektif untuk mengatasi masalah ini.

Pengenalan gerak bibir merupakan bidang penelitian yang penting dalam pengenalan ucapan, terutama dalam situasi di mana audio tidak tersedia atau sulit diakses. Dalam konteks ini, menggunakan citra atau video gerak bibir men-

DOI : http://dx.doi.org/10.32409/jikstik.24.1.3658, *) Penulis Korespondensi

jadi alternatif yang menjanjikan, memungkinkan sistem untuk menguraikan kata-kata yang diucapkan berdasarkan gerakan bibir seseorang.

Convolutional Neural Networks (CNN) telah menjadi standar dalam pengenalan gambar dan video. Dengan kemampuannya untuk mengekstraksi fitur secara hierarkis dari gambar. CNN termasuk dalam jenis *Deep Neural Network* karena kedalaman jaringan yang tinggi dan banyak diaplikasikan pada data citra [2]. CNN dapat digunakan untuk mengekstraksi fitur dari citra gerak bibir dengan mempertimbangkan konteks spasial dari bibir. *Long Short Term Memory* (LSTM) merupakan salah satu pengembangan *Recurrent Neural Network* (RNN) untuk mengatasi masalah difusi gradien [3]. *Long Short-Term Memory* (LSTM) adalah jenis arsitektur jaringan saraf rekuren (RNN) yang mampu menangani masalah pengenalan urutan, seperti pengenalan kata-kata dari urutan gerakan bibir. Dengan mempertahankan informasi jangka panjang, LSTM cocok untuk memodelkan urutan waktu dalam data gerak bibir. Connectionist Temporal Classification (CTC) adalah kriteria yang banyak digunakan untuk melatih model sequence-to-sequence yang diawasi [4]. Dalam konteks pengenalan gerak bibir, CTC *Loss Function* dapat digunakan untuk melatih model secara *end-to-end* tanpa perlu pemetaan yang tepat antara gerakan bibir dan kata-kata.

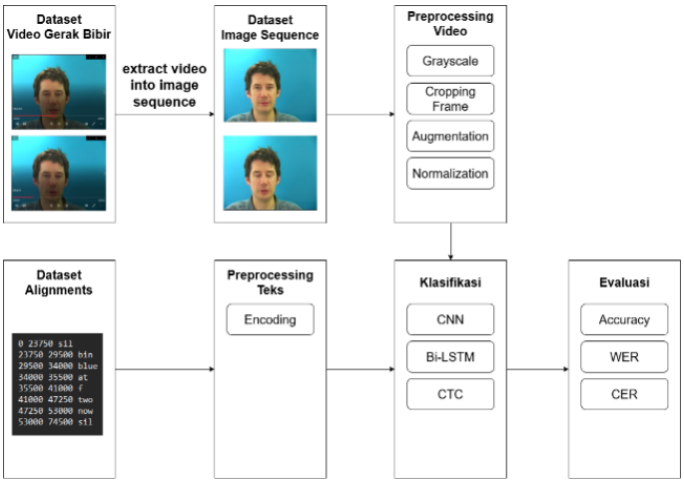
Metode CNN, LSTM dan CTC *loss fuction* banyak digunakan dalam penelitian membaca gerak bibir dan mendapatkan akurasi *Word Error Rate* (WER) sebesar 4,8% dan *Character Error Rate* (CER) 1,9% [5]. Metode CNN sendiri juga digunakan pada penelitian lain dan mendapatkan akurasi sebesar 90% [6]. Metode LSTM juga digunakan pada penelitian membaca gerak bibir lainnya dengan akurasi sebesar 85% [7]. Metode CTC *loss function* juga digunakan dalam penelitian untuk membaca gerak bibir yang mendapatkan akurasi WER sebesar 25,5% dan CER 8% [8].

Penelitian ini bertujuan mengembangkan metode CNN, LSTM dan CTC l*oss function* untuk pembacaan gerak bibir dengan melihat kelemahan dan kelebihan dari peneliti sebelumnya. Hasil penelitian ini diharapkan bahwa dengan menggabungkan metode-metode ini, sistem pengenalan gerak bibir dapat mencapai tingkat akurasi yang tinggi dalam menghadapi variasi dalam gerakan bibir.

Metode Penelitian

Secara umum, sistem ini dibuat untuk membaca gerak bibir berdasarkan data berupa video. Sistem ini dapat memprediksi kata atau huruf pada suatu video yang berbahasa inggris. Gambar 1 merupakan gambaran umum sistem yang terdiri dari beberapa tahapan yaitu *preprocessing* video yang meliputi konversi *grayscale, cropping frame, augmentasi* dan normalisasi yang bertujuan untuk menghasilkan citra yang lebih baik untuk diproses. Kemudian dilakukan proses *pre-processing text* dengan melakukan encoding pada dataset *alignments.*

Proses klasifikasi menggunakan *Convolutional Neural Networks, Long Short-Term Memory* dan *Connectionist Temporal Classification Loss Function.* Kemudian melakukan proses pengujian model dengan mengukur akurasi dan loss. Setelah selesai melakukan semua proses, maka sistem akan memberikan hasil model dan analisis. Hasil evaluasi menggunakan nilai akurasi, WER dan CER.



Gambar 1: Metode Penelitian

Dataset

Dataset yang digunakan pada penelitian ini berupa video seorang pria yang mengucapkan beberapa kata dan huruf dalam bahasa inggris berjumlah 1000 video dan data berisikan kata dan huruf yang diucapkan dalam video sebanyak 1000 file. Dataset yang digunakan ini di unduh dari *https://spandh.dcs.shef.ac.uk/gridcorpus.*

Dataset yang telah diunduh akan dipindahkan ke dalam google drive dan diekstrak yang akan memudahkan sistem untuk mengakses dataset dari google colab ke google drive.

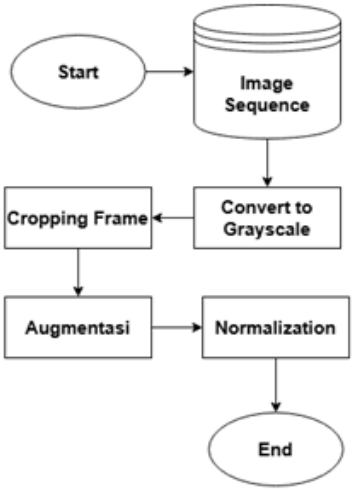
Extract Video To Image Sequence

Image sequence melibatkan pengambilan *frame* dari video untuk diolah lebih lanjut. Pada tahap ini, video yang ada pada dataset diubah menjadi serangkaian *frame.* Setiap frame ini kemudian diurutkan berdasarkan waktu pengambilan sehingga urutan kejadian dalam video tetap terjaga. Dengan cara ini, dapat memfokuskan analisis pada perubahan gerakan bibir dari satu *frame* ke *frame* berikutnya, memungkinkan model untuk belajar dari data visual yang ada.

Pada Tabel 1 merupakan contoh dari *image sequence* yang disesuaikan dengan kata yang di ucapkan oleh pembicara pada dataset.

Tabel 1: Teks Sesuai *Image Sequence*

Teks	Frame
Bin	
Blue	
Two	
At	



Gambar 2: *Tahap Preprocessing Video*

Dataset yang telah diekstrak diolah melalui beberapa tahapan dalam diagram alur yang ditunjukkan pada Gambar 2. Pertama, konversi grayscale untuk mengurangi kompleksitas warna, kemudian dilakukan *cropping frame* untuk mengambil area bibir. Selanjutnya, dilakukan augmentasi data untuk meningkatkan ukuran dan keragaman dataset dengan *flip* secara *horizontal*. Setelah itu, data dinormalisasi untuk memastikan semua fitur berada dalam skala yang sama, sehingga model pembelajaran mesin dapat berfungsi lebih efektif.

Grayscale

Metode *grayscale* adalah teknik yang digunakan untuk mengatur tingkat kecerahan dan kontras gambar dengan mengulangi siklus antara gambar*grayscale* asli dan gambar dengan kecerahan dan kontras yang telah disesuaikan [9]. Tahap *grayscale* adalah tahap dimana setiap *frame* dari video diubah dari format warna *Red Green Blue* (RGB) menjadi format skala abu-abu. Hal Ini dilakukan untuk mengurangi kompleksitas data dengan menghilangkan informasi warna yang mungkin tidak relevan untuk tugas klasifikasi tertentu.

Konversi ini juga membantu dalam mengurangi ukuran data dan mempercepat proses pemrosesan tanpa kehilangan informasi penting yang dapat diambil dari intensitas cahaya dalam *frame*.

Cropping Frame

Tahap *cropping frame* adalah proses memotong bagian bibir dari setiap frame video untuk menghilangkan area yang tidak relevan dan fokus pada bagian penting yang diperlukan untuk klasifikasi. Ini membantu dalam meningkatkan akurasi model dengan memastikan bahwa hanya informasi yang relevan yang digunakan dalam proses pelatihan dan pengujian.

Cropping juga dapat membantu dalam mengurangi ukuran data, yang selanjutnya mengurangi kebutuhan komputasi dan memori, sehingga mempercepat keseluruhan proses klasifikasi.

Augmentasi Data

Augmentasi adalah proses yang digunakan untuk menambah jumlah data dengan cara membuat data baru dari data yang sudah ada [10].

Dalam konteks video, augmentasi dapat mencakup *flipping horizontal* yang bertujuan untuk meningkatkan keragaman data pelatihan. Ini membantu model untuk generalisasi lebih baik dengan melihat berbagai variasi dari data asli.

Normalisasi Data

Normalisasi data adalah proses yang memastikan rentang nilai dari beberapa variabel menjadi seragam, tanpa ada yang terlalu besar atau terlalu kecil, sehingga mempermudah analisis statistik [11]. Tahap terakhir yaitu normalisasi data, ini sangat penting dalam pelatihan model *machine learning* karena membantu dalam skala nilai-nilai piksel ke rentang yang lebih konsisten, yang membuat proses pelatihan lebih stabil dan cepat. Normalisasi data menggunakan beberapa rumus untuk mendapatkan hasil yang optimal seperti rumus mean, standar deviasi.

$$mean = \frac{1}{N}\sum_{i=1}^{N} frame_i \tag{1}$$

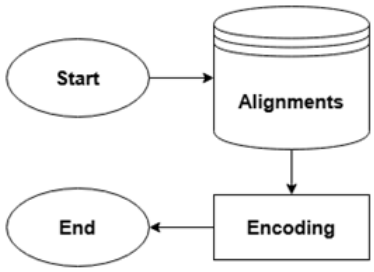
$$std = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(frame_i - mean)^2} \tag{2}$$

$$normalized = \frac{frames - mean(frames)}{std(frames)} \tag{3}$$

Pada persamaan (1), (2) dan (3), merupakan rumus normalisasi dengan cara z-*score* yang masing-masing nilai pada fitur dikurangi dengan rata-rata fitur kemudian dibagi dengan standar deviasi.

Tahapan *Preprocessing* Teks

File yang berformat **.align* berisikan kata atau huruf yang diucapkan oleh pembicara pada video. Kata atau huruf tersebut tidak dapat diproses oleh machine learning. Oleh karena itu, kata atau huruf akan melakukan proses *encoding* untuk dapat diproses pada tahap klasifikasi yang dapat dilihat pada Gambar 3.



Gambar 3: *Tahap Preprocessing Teks*

Encoding

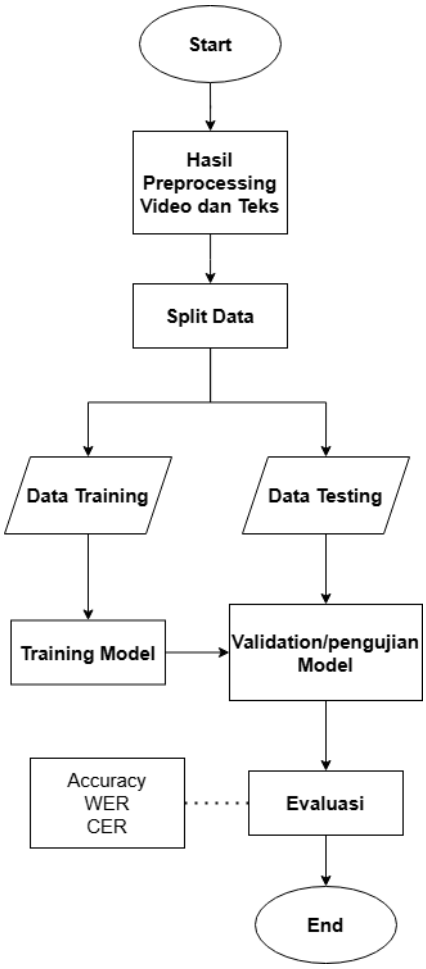
Encode merupakan proses mengubah teks yang dapat dibaca dan dipahami oleh manusia menjadi kriptografi atau teks yang tidak dapat dibaca dan dipahami oleh manusia [12]. Pada tahap ini file alignment akan di-*preprocessing* dengan melakukan encoding yaitu merubah teks menjadi bilangan bulat. Pada file alignment terdapat kata-kata yang diucapkan oleh pembicara. Dikarenakan menggunakan CTC *Loss Function* kata-kata tersebut akan di-*encode* per huruf dari setiap kata menjadi ("a= 1), ("b" = 2), ("c" = 3) dan seterusnya. Hasil dari *encode* dapat dilihat pada Tabel 2.

Tabel 2: Hasil Pengujian *Satisfaction*

Karakter	Encode	Karakter	Encode
a	1	u	21
b	2	v	22
c	3	w	23
d	4	x	24
e	5	y	25
f	6	z	26
g	7	!	27
h	8	?	28
i	9	!	29
j	10	1	30
k	11	2	31
l	12	3	32
m	13	4	33
n	14	5	34
o	15	6	35
p	16	7	36
q	17	8	37
r	18	9	38
s	19	space	39
t	20		

Klasifikasi

Pada Gambar 4, penulis melakukan *split* data untuk membuat beberapa skenario yang akan dilakukan dalam penelitian ini, menghasilkan data *training* dan data *testing*. Dari data tersebut akan dibuat sebuah model yang akan digunakan untuk membaca gerak bibir, arsitektur yang menjadi referensi pembuatan model adalah arsitektur *LipNet*. Arsitektur ini terdiri dari *Convolutional Neural Network* (CNN) untuk mengekstraksi fitur, *Long Short-Term Memory* (LSTM) untuk menangani data urutan dan *Connectionist Temporal Classification* (CTC) untuk pelabelan urutan dengan panjang variabel.



Gambar 4: Alur Klasifikasi

Model yang sudah dibuat akan dievaluasi berdasarkan akurasi, *Word Error Rate* (WER) dan *Character Error Rate* (CER) untuk untuk menilai kinerjanya dalam pengenalan ucapan dari gerakan bibir. Akurasi mengukur persentase prediksi yang benar dari total prediksi yang dibuat oleh model. *Word Error Rate* (WER) mengukur kesalahan pada tingkat kata dengan menghitung jumlah substitusi, penyisipan, dan penghapusan kata yang diperlukan untuk mengubah prediksi model menjadi teks referensi. *Character Error Rate* (CER) mengukur

kesalahan pada tingkat karakter dengan menghitung jumlah substitusi, penyisipan, dan penghapusan karakter yang diperlukan untuk mengubah prediksi model menjadi teks referensi.

Skenario Model

Pelatihan model dilakukan dengan 3 skenario yaitu distribusi data 80:20 dengan *batch size* 2, distribusi data 70:30 dengan batch size 2 dan distribusi data 80:20 dengan batch size 2. Skenario ini dibuat bertujuan untuk melihat perbandingan akurasi dan menggunakan *batch size* 2 untuk mengurangi komputasi yang besar.

Oleh karena itu, pembagian data adalah proses membagi data yang telah proses sebelumnya menjadi beberapa bagian. Pembagian data ini dilakukan dengan parameter tertentu. *Split* data terdiri dari dua bagian yaitu *train* data dan *test* data. Data train adalah bagian dataset yang dilatih untuk membuat prediksi algoritma untuk mencapai tujuan, sedangkan validation data adalah bagian dataset yang digunakan untuk proses validasi model dan mencegah *overfitting.*

Skenario pengujian dilakukan dengan rasio 60%, 70% dan 80% untuk data pelatih serta 40%, 30% dan 20% untuk data uji seperti yang ada pada Tabel 3.

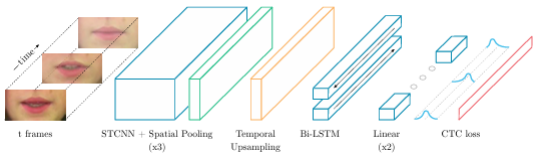
Tabel 3: Skenario Pelatihan Model

Skenario	Split Data	Batch Size	Jumlah Data Latih	Jumlah Data Uji
1	60:40	2	600	400
2	70:30	2	700	300
3	80:20	2	800	200

Modeling

Dalam pembuatan model, model ini menggunakan arsitektur *LipNet* yang sering digunakan untuk membuat jaringan saraf membaca gerak bibir *(Lipreading*) yang memetakan urutan panjang variabel frame video ke urutan teks, dan dilatih dari awal ke akhir.

Pada Gambar 5 mengilustrasikan arsitektur *LipNet*, yang terdiri dari 3 lapisan *Spatiotemporal Convolutional Neural Networks* (STCNN). Selanjutnya diikuti oleh Bi-LSTM, Bi-LSTM sangat penting untuk agregasi lebih lanjut yang efisien dari output STCNN. Terakhir, jaringan *feed-forward* diterapkan pada setiap langkah waktu, diikuti oleh softmax atas kosakata yang ditambah dengan CTC kosong, dan kemudian CTC *loss.* Semua lapisan menggunakan fungsi aktivasi *Rectified Linear Unit* (ReLU).



Gambar 5: Arsitektur *LipNet*

Convolutional Neural Networks

Model ini menggunakan beberapa lapisan *Convolutional Neural Network* (CNN) untuk mengekstraksi fitur dari video input. Tiga lapisan Conv3D digunakan dengan filter yang bertambah dari 128, 256 dan 75. Setiap lapisan konvolusi diikuti oleh aktivasi ReLU dan lapisan *MaxPooling*3D untuk mengurangi dimensi spasial.

Untuk menghubungkan CNN dengan Bi-LSTM, digunakan lapisan TimeDistributed digunakan untuk meratakan output dari lapisan CNN sehingga dapat diterima dengan baik oleh lapisan LSTM.

Long Short-Term Memory

Pada lapisan *Long Short-Term Memory* (LSTM), model menggunakan dua lapisan *Bidirectional* LSTM untuk menangkap dependensi temporal dari *frame* video. LSTM ini memiliki 128 unit dan menggunakan inisialisasi kernel *Orthogonal.* Untuk mencegah overfitting maka diterapkan Dropout sebesar 50% pada setiap lapisan LSTM. Lapisan output dari model ini adalah lapisan *Dense* dengan aktivasi softmax yang memetakan *output* LSTM ke urutan karakter.

Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) merupakan algoritma yang digunakan untuk melatih jaringan saraf dalam tugas seperti pengenalan ucapan dan pengenalan tulisan tangan, serta masalah sekuensial dimana tidak ada informasi eksplisit tentang penyelarasan antara masukan dan keluaran [13].

Model ini menghasilkan urutan distribusi diskrit atas kelas-kelas token (kosakata) yang ditambah dengan token "kosong" khusus, CTC menghitung probabilitas dari sebuah urutan dengan memarginalkan semua urutan yang didefinisikan sebagai ekuivalen dengan urutan ini. Hal ini secara bersamaan menghilangkan kebutuhan untuk penyelarasan dan mengatasi urutan panjang variabel.

Metode Evaluasi

Evaluasi adalah langkah terakhir dalam penelitian ini. Pada tahap ini, model akan diukur dalam akurasi berdasarkan file video yang dijalankan, *Word Error Rate* (WER) dan *Character Error Rate* (CER).

Akurasi

Akurasi dihitung sebagai persentase jumlah prediksi yang benar (Kalimat yang diprediksi sesuai dengan label sebenarnya) dibagi dengan jumlah total video dalam dataset. Adapun rumus mencari akurasi pada persamaan (4).

$$Akurasi = \frac{JumlahPrediksiBenar}{JumlahTotalSampel} X\,100\% \quad (4)$$

Word Error Rate (WER)

WER adalah rasio jumlah operasi pengeditan (penyisipan, penghapusan dan substitusi) yang diperlukan untuk mengubah prediksi menjadi label yang benar, dibandingkan dengan jumlah kata dalam label yang benar. Rumus WER dan keterangannya ada pada persamaan (5).

$$WER = \frac{D}{N} X\,100\% \quad (5)$$

dimana :

1. D adalah jumlah operasi pengeditan yang diperlukan (dihitung menggunakan jarak Levenshtein).
2. N adalah jumlah kata dalam label yang benar.

Character Error Rate (CER)

CER adalah rasio jumlah operasi pengeditan (penyisipan, penghapusan, dan substitusi) yang diperlukan untuk mengubah prediksi menjadi label yang benar, dibandingkan dengan jumlah karakter dalam label yang benar. Rumus CER dan keterangan ada pada persamaan (6).

$$CER = \frac{D}{N} X\,100\% \quad (6)$$

dimana :

1. D adalah jumlah operasi pengeditan yang diperlukan (dihitung menggunakan jarak Levenshtein).
2. N adalah jumlah karakter dalam label yang benar.

Hasil dan Evaluasi

Pada bagian ini menjelaskan hasil eksperimen yang telah dilakukan pada penelitian tersebut. Hasil penelitian akan menjelaskan hasil akurasi model yang didapatkan pada model yang diteliti.

Hasil *Image Sequence*

Tahap pertama dalam proses video adalah mengubah video menjadi serangkaian gambar atau frame. Setiap frame diambil dari video pada interval waktu tertentu sehingga mencakup seluruh durasi video. Gambar 6 menunjukkan hasil dari video yang telah diproses menjadi image sequence.

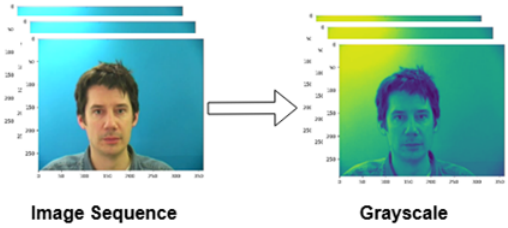
Gambar 6: Hasil *Image Sequence*

Hasil *Preprocessing* Video

Tahap preprocessing melibatkan beberapa langkah untuk mempersiapkan data sebelum digunakan dalam pelatihan model. Langkah-langkah ini dirancang untuk mengurangi kompleksitas data dan menyoroti fitur-fitur penting yang relevan untuk analisis gerakan bibir. *Preprocessing* yang dilakukan pada file berformat *.mpg atau video berupa *grayscale, cropping frame, augmentasi* dan normalisasi video pada bagian bibir.

Hasil *Grayscale*

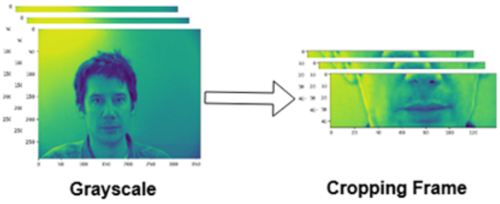
Tahap *grayscale* mengubah setiap frame dari video menjadi format skala abu-abu. Proses ini menghilangkan informasi warna yang tidak diperlukan dan fokus pada intensitas cahaya, sehingga mengurangi kompleksitas data. Gambar 7 menunjukkan hasil p*reprocessing grayscale* di mana gambar berwarna diubah menjadi gambar skala abu-abu, menyoroti fitur-fitur penting dari wajah yang dapat digunakan untuk analisis lebih lanjut.



Gambar 7: Hasil *Preprocessing Grayscale*

Hasil *Cropping Frame*

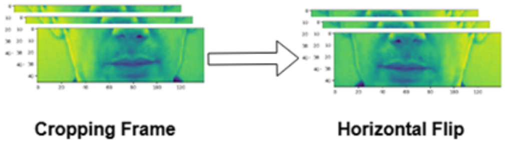
Tahap *cropping frame* memotong bagian video untuk fokus pada area sekitar bibir, yang merupakan area penting untuk pengenalan gerakan bibir. Proses ini menghilangkan bagian-bagian gambar yang tidak relevan dan mengurangi dimensi data, memungkinkan model untuk lebih efisien memproses dan menganalisis gerakan bibir. Gambar 8 menunjukkan hasil cropping frame, di mana bagian yang tidak relevan dari gambar dihilangkan, dan area bibir dipertahankan untuk analisis selanjutnya.



Gambar 8: Hasil P*reprocessing Cropping Frame*

Hasil *Augmentasi*

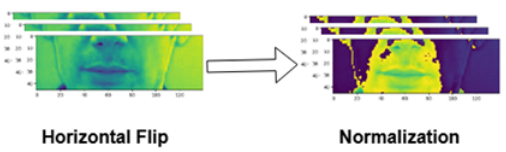
Proses *augmentasi* data melibatkan penerapan berbagai transformasi pada data asli untuk meningkatkan ukuran dan keragaman dataset. Transformasi ini mencakup pembalikan horizontal. Tujuannya adalah untuk membuat model lebih kuat terhadap variasi dalam data, sehingga dapat mengenali pola dalam data baru yang belum pernah dilihat sebelumnya. Gambar 9 menunjukkan hasil dari proses augmentasi menggunakan flip horizontal.



Gambar 9: Hasil *Augmentasi*

Hasil Normalisasi

Tahap normalisasi dilakukan untuk memastikan bahwa semua fitur dalam data berada dalam skala yang sama. Proses ini penting untuk meningkatkan kinerja model pembelajaran mesin, karena data yang tidak dinormalisasi dapat menyebabkan model mengalami kesulitan dalam menemukan pola yang konsisten. Gambar 10 menunjukkan hasil normalisasi, di mana nilai piksel dari gambar yang telah diproses diubah ke skala standar, memungkinkan model untuk memproses data dengan lebih efektif dan efisien.

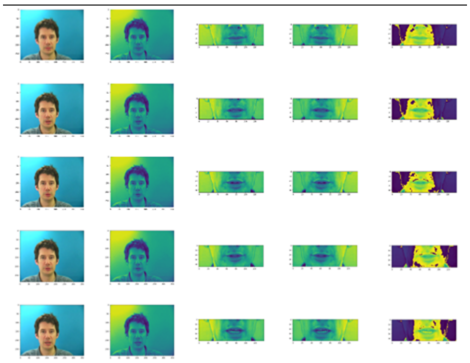


Gambar 10: Hasil Normalisasi

Tabel 4 menunjukkan hasil dari berbagai tahap preprocessing, augmentasi dan normaliasai yang diterapkan pada video. Kolom pertama menampilkan gambar asli dari video. Kolom kedua menampilkan hasil grayscale, di mana gambar diubah menjadi skala abu-abu untuk mengurangi kompleksitas data. Kolom ketiga menunjukkan hasil *cropping,* yang memfokuskan gambar pada area bibir untuk meningkatkan akurasi pengenalan gerakan bibir. Kolom keempat menampilkan hasil *flip horizontal*, yang mencerminkan gambar untuk menciptakan variasi data. Kolom kelima menunjukkan hasil normalisasi, di mana nilai piksel gambar distandarisasi untuk memastikan konsistensi dalam input data. Proses ini bertujuan untuk mempersiapkan data video sehingga model pembelajaran mesin dapat menganalisisnya dengan lebih efektif dan akurat.

Tabel 4: Hasil *Preprocessing* Video

Original	Greyscale	Cropping Frame	Augmentasi	Normalisasi

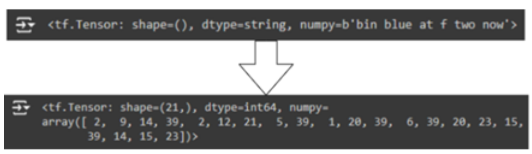


Hasil *Preprocessing* Teks

Preprocessing yang dilakukan pada file berformat **.align* atau teks berupa encode, yaitu merubah teks yang awalnya huruf menjadi sebuah angka. Proses ini penting untuk mengubah data teks menjadi format numerik yang dapat dipahami oleh model pembelajaran mesin.

Hasil *Encoding*

Proses ini penting untuk mengubah data teks menjadi format numerik yang dapat dipahami oleh model pembelajaran mesin. Gambar 11 menunjukkan hasil dari proses *encoding,* di mana teks asli telah dikonversi menjadi urutan angka yang sesuai dengan setiap karakter dalam teks. Proses ini memastikan bahwa model dapat memproses dan menganalisis data teks dengan cara yang konsisten dan terstruktur.



Gambar 11: Hasil *Encoding* Teks

Hasil Evaluasi

Proses pelatihan model ini menggunakan algoritma *Convolutional Neural Networks* (CNN),

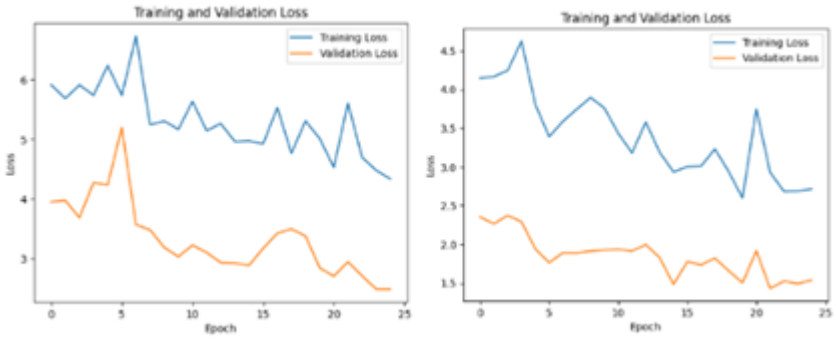
B*idirectional Long Short-Term Memory* (Bi-LSTM) dan *Connectionist Temporal Classification* (CTC) *Loss Function* dilakukan untuk membaca gerak bibir secara benar dengan evaluasi menggunakan *Word Error Rate* (WER) dan *Character Error Rate* (CER). Pada Gambar 12 dapat dilihat lapisan yang digunakan untuk membuat model dengan input *shape* sebesar (75, 46, 140, 1) dari hasil *preprocessing* video yang telah dilakukan sebelumnya. Dengan layer yang digunakan sebanyak 3 lapisan *conv3d + max_pooling3d* dan 2 lapisan Bi-LSTM yang dihubungkan dengan TimeDistributed untuk menghubungkan 2 algoritma tersebut.

Pada penelitian ini akan dilakukan epoch sebanyak 8 x 25 epoch pada masing-masing skenario yang telah dibuat sebelumnya. Hasil dari setiap skenario dapat dilihat pada Tabel 5. Berdasarkan Tabel 5 hasil yang terbaik didapatkan pada skenario *split* data 80:20 *Batch size* 2 memiliki akurasi terbesar yaitu 83.4%, WER terkecil yaitu 3% dan CER terkecil yaitu 0.2%. Pada Gambar 13 merupakan hasil dari proses menjalankan pelatihan dengan memvisualisasikan nilai dari loss dan validation loss pada langkah pelatihan.

Tabel 5: *Hasil Pelatihan Model*

Skenario	Epoch	Akurasi	WER	CER
Split Data 60:40	100	51,7%	94,69%	23%
	125	54,7%	92,74%	22,5%
	150	58,6%	43,26%	11,8%
	175	62,3%	20,19%	4,93%
	200	85,7%	3,42%	0,8%
Split Data 70:30	100	1%	28,66%	12.1%
	125	21,7%	15,42%	4%
	150	48,4%	9,25%	2%
	175	76%	5%	1,25%
	200	87.8%	2,84%	0,6%
Split Data 80:20	100	48,87%	9,62%	2.7%
	125	70,1%	6,24%	1.53%
	150	84.2%	3,27%	0.8%
	175	93%	1,59%	0,3%
	200	**96,9%**	**0,66%**	**0,16%**

Gambar 12: Rancangan Model



Gambar 13: Grafik *Loss* Dan *Validation Loss*

Perbandingan Hasil Penelitian Terdahulu

Pada Tabel 6 didapatkan hasil membaca gerak bibir dengan model yang digunakan dari penelitian sebelumnya. Hasil yang didapatkan dilihat dari akurasi, *Word Error Rate* (WER) dan *Character Error Rate* (CER). Hasil uji coba penelitian membaca gerak bibir dengan model *Convolutional Neural Networks* (CNN), *Long Short-Term Memory* (LSTM) dan *Connectionist Temporal Classification* (CTC) mendapatkan akurasi sebesar 96,9%, WER sebesar 0,66% dan CER sebesar 0,16%. Hal ini menunjukkan peningkatan nilai WER dan CER dibandingkan penelitian sebelumnya yaitu penelitian yang menggunakan metode *LipNet* (CNN, Bi-GRU, CTC) dengan menghasilkan akurasi WER sebesar 4,8% dan CER 1,9% [5]. Penelitian yang menggunakan metode 3D Conv + ResNet18 + CTC mendapatkan akurasi WER sebesar 25,5% dan CER 8% [8]. Penelitian yang menggunakan metode 3D + E3D-LSTM + CTC *loss* mendapatkan akurasi sebesar 38,96% [14]. Penelitian yang menggunakan metode LSTM mendapatkan akurasi sebesar 77,81% [15]. Penelitian yang menggunakan metode 2D+3D Conv mendapatkan akurasi sebesar 38,91% [16]. Penelitian yang menggunakan dataset GRID dengan metode CNN+Bi-LSTM+CTC mendapatkan akurasi sebesar 98,7% [17].

Hasil Penelitian ini menunjukkan bahwa metode yang digunakan memberikan hasil yang lebih baik dibandingkan metode yang telah ada. Arsitektur LipNet menggunakan algoritma CNN, Bi-LSTM dan CTC terbukti dapat meningkatkan tingkat akurasi pada proses membaca gerak bibir. Hasil akurasi juga berpengaruh pada dataset yang digunakan, metode *preprocessing* yang dilakukan dan jumlah data yang digunakan.

Tabel 6: Perbandingan Penelitian Terdahulu

Peneliti	Klasifikasi	Jenis dan Jumlah Data	Akurasi	WER	CER
Assael et al, 2016	LipNet (CNN, Bi-GRU, CTC)	GRID, 32.746	-	4,8%	1,9%
Ma et al, 2022	3D Conv + ResNet 18 + CTC	LRW+LRS, 296.301	-	25,5%	8%
Yang et al, 2019	2D + 3D Conv	LRW-1000, 718.018	38,19%	-	-
Hao et al, 2020	3D + 2D CNN, Bi-LSTM, CTC	Grid, 34.000	98,7 %	-	-
Mudaliar et al, 2020	RestNet + 3D Conv + GRU	LRW, 173.510	90 %	-	-
Deshmuk	RestNe	Custom,	85%	-	-
Peneliti	**Klasifikasi**	**Jenis dan Jumlah Data**	**Akurasi**	**WER**	**CER**
h et al, 2021	t50 + LSTM	540			
Bi et al, 2019	3D+E3D-LSTM +CTC Loss	LRW-1000, 128.520	38,96%	-	-
Berkol et al, 2023	Bi-GRU dan LSTM	Custom, 4.726	77,81%	-	-
Our Method	**3D CNN + Bi-LSTM + CTC**	**Grid, 1.000**	**96,9 %**	**0,66%**	**0,16 %**

Penutup

Berdasarkan penelitian dan pembahasan mengenai membaca gerak bibir menggunakan *Convolutional Neural Network* (CNN), *Bidirectional Long Short-Term Memory* (Bi-LSTM) dan *Connectionis Temporal Classification* (CTC) didapatkan akurasi sebesar 96,9%, WER sebesar 0,66% dan CER sebesar 0,16% dengan menggunakan skenario *split* data 80:20 sebesar 800 data *training* dan 200 data t*esting* serta *batch size* sebesar 2.

Hasil ini menunjukkan bahwa kombinasi CNN, Bi-LSTM, dan CTC efektif dalam menangkap dan menganalisis gerakan bibir untuk keperluan pengenalan ucapan. Model ini mampu mencapai tingkat akurasi yang tinggi, dengan kesalahan pada tingkat kata dan karakter yang sangat rendah, menunjukkan kemampuan yang kuat dalam mengatasi variasi dan kompleksitas dalam data video. Dengan demikian, pendekatan ini dapat diandalkan untuk aplikasi pengenalan gerak bibir, memberikan kontribusi signifikan dalam membantu komunikasi bagi individu tunarungu dan dalam pengembangan teknologi pengenalan ucapan secara umum.

Daftar Pustaka

[1] N. Haliza, E. Kuntarto, and A. Kusmana, "Pemerolehan Bahasa Anak Berkebutuhan Khusus (Tunarungu) Dalam Memahami Bahasa," Jurnal Metabasa, vol. 2, no. 1, http://dx.doi.org/10.26555/jg.v2i1.2051, 2020.

[2] Fani Nurona Cahya, Nila Hardi, Dwiza Riana, dan Sri Hadiyanti., "SISTEMASI: Jurnal Sistem Informasi Klasifikasi Penyakit Mata Menggunakan Convolutional Neural Network (CNN)," Sistemasi: Jurnal Sistem Informasi, Vol.10, No. 3, http://dx.doi.org/10.32520/stmsi.v10i3.1248, 2021.

[3] D. I. Af'idah, D. Dairoh, S. F. Handayani, R. W. Pratiwi,dan S. I. Sari, "Sentimen

Ulasan Destinasi Wisata Pulau Bali Menggunakan Bidirectional Long Short Term Memory," MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 21, no. 3, pp. 607–618, https://doi.org/10.30812/matrik.v21i3.1402, Jul. 2022.

[4] Eliya Segev, Maya Alroy, Ronen Katsir, Noam Wies, Ayana Shenhav, Yael Ben-Oren, David Zar, Oren Tadmor, Jacob Bitterman, Amnon Shashua, and Tal Rosenwein, "Align With Purpose: Optimize Desired Properties in CTC Models with a General Plug-and-Play Framework," arXiv Computer Science, https://doi.org/10.48550/arXiv.2307.01715, Jul. 2023.

[5] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv Computer Science, https://doi.org/10.48550/arXiv.1611.01599, Nov. 2016.

[6] Mudaliar Navin Kumar, Hegde Kavita, Ramesh Anand, and Patil Varsha, "Visual Speech Recognition: A Deep Learning Approach," 2020 5th International Conference on Communication and Electronics Systems (ICCES), https://doi.org/10.1109/ICCES48766.2020.9137926, 2020.

[7] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali, and K. Warkari, "Vision based Lip Reading System using Deep Learning," in 2021 International Conference on Computing, Communication and Green Engineering, Institute of Electrical and Electronics Engineers Inc., https://doi.org/10.1109/CCGE50943.2021.9776430, 2021.

[8] P. Ma, S. Petridis, and M. Pantic, "Visual Speech Recognition for Multiple Languages in the Wild," Nat Mach Intell 4, 930–939, https://doi.org/10.1038/s42256-022-00550-z, Feb. 2022.

[9] R. Pramudiya, C. Asyraq, A. Kadafi, and R. P. Sardika, "Analisis Gambar Menggunakan Metode Grayscale Dan Hsv (Hue, Saturation, Value)," Just IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer, Volume 14 No 3, https://doi.org/10.24853/justit.14.3.174-180, 2024.

[10] W. M. Pradnya D and A. P. Kusumaningtyas, "Analisis Pengaruh Data Augmentasi Pada Klasifikasi Bumbu Dapur Menggunakan Convolutional Neural Network," Jurnal Media Informatika Budidarma, vol. 6, no. 4, p. 2022, http://dx.doi.org/10.30865/mib.v6i4.4201, Oct. 2022.

[11] I. Yati Beti and H. Juliansa, "Penerapan Normalisasi Data Metode Decimal Scaling Dan Metode K-Means Dalam Mengelompokkan Kasus Demam Berdarah," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 4, no. 6, pp. 2928–2936, doi: 10.30865/klik.v4i6.1925, 2024.

[12] Mochammad Syahrul Kurniawan, I Gede Ardi Sukaryadi Putra, I Made Agastya Maheswara, Reynaldus Yoseph Maria Neto Labamaking, I Made Edy Listartha, & Gede Arna Jude Saskara, "Analisis Efektivitas Dan Efisiensi Metode Encoding Dan Decoding Algoritma Base64," Jurnal Informatika Dan Tekonologi Komputer (JITEK), vol. 3, no. 1, doi: 10.55606/jitek.v3i1.897, 2023.

[13] Zhang, Z., Lu, N., Liao, M., Huang, Y., Li, C., Wang, M., & Peng, W, "Self-Distillation Regularized Connectionist Temporal Classification Loss for Text Recognition: A Simple Yet Effective Approach", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 7, pp. 7441-7449, doi: https://doi.org/10.48550/arXiv.2308.08806, 2023.

[14] C. Bi, D. Zhang, L. Yang, and P. Chen, "An Lipreading Modle With DenseNet And E3D-LSTM," 2019 6th International Conference on Systems and Informatics (ICSAI), pp. 511-515, doi: 10.1109/ICSAI48974.2019.9010432, 2019.

[15] A. Berkol, T. Tümer Sivri, And H. Erdem, "Lip Reading Using Various Deep Learning Models with Visual Turkish Data," Gazi University Journal Of Science, vol. 37, doi: 10.35378/gujs.1239207, 2023.

[16] Yang, Shuang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, S. Shan and Xilin Chen. "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild." 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019),pp.: 1-8. doi: 10.1109/FG.2019.8756582, 2018.

[17] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A survey of research on lipreading technology," in IEEE Access, vol. 8, pp. 204518-204544, doi: 10.1109/ACCESS.2020.3036865, 2020.