Teknologi: Jurnal Ilmiah Sistem Informasi 15 (1) 2025 32-43

ISSN 2527-3671 (*Online*)
ISSN 2087-8893 (*Print*)

Tersedia *online* di www.journal.unipdu.ac.id

Halaman jurnal di www.journal.unipdu.ac.id/index.php/teknologi

Pengenalan Ucapan Pelo Menggunakan *Short Time Fourier Transform* dengan Pemodelan *Lightweight Convolutional Neural Network*

Andi Sugandi[a], Henry Ardian Irianta[b], Muhammad Fauzan Gustafi[c]
abc Program Studi Informatika, Universitas Siber Muhammadiyah, Yogyakarta, Indonesia

email: [a,]andi@sibermu.ac.id [b,*] henryai@sibermu.ac.id [c,*]muhammadfauzangustafi@sibermu.ac.id*

*Korespondensi *andi@sibermu.ac.id*

Dikirim 21 Mei 2025; **Direvisi** 01 Juni 2025; **Diterima** 09 Juni 2025; **Diterbitkan** 15 Juni 2025

Abstrak

Disartria adalah gangguan neurologis yang menghambat penderita untuk mengucapkan kata-kata dengan benar. Saat ini, *Speech Command Recognition* (SCR) turunan rumpun ilmu *Automatic Speech Recognition* (ASR) sedang diteliti dan dikembangkan untuk membantu penderita disartria atau pelo sehingga sulit berkomunikasi. Salah satu tahap dasar dalam membangun SCR adalah proses pengenalan kata, klasifikasi dan prediksi ucapan kata. Penelitian ini bertujuan untuk membangun pemodelan *Deep learning,* untuk pengenalan ucapan pelo, dengan desain arsitektur *deep learning* efisien yaitu *Lightweight Convolutional Neural Network* (LCNN), menggunakan ekstarksi ciri *Short Time Fourier Transform STFT* (STFT-LCNN). Augmentasi data dengan *noise posisition* dan *re-pitch* agar menambah variasi dataset yang efisien. Pendekatan arsitektur LCNN dihgunakan untuk implementasi *edge devices* kedepanya. Dalam pengimplementasiannya, model menggunakan 2 layer konvolusi, ekstraksi ciri menggunakan spektrogram, dan mennggunakan mini dataset dari *EasyCall* untuk pendeteksian, 5 kelas klasifikasi dengan penutur pelo atau disartria. Metode ini menghasilkan akurasi pemodelan sebesar 82%.

Kata Kunci: *Dysarthic speech Recognition*, Deteksi Pelo, *SCR, ASR*, *Keyword Spotting*, LCNN, LCNN-STFT

Dysarthia Speech Recognition Using Short Time Fourier Transform and Lightweight Convolutional Neural Network Modeling

Abstract

Dysarthic Speech is a neurological disorder that impairs an individual's ability to articulate words correctly. Currently, Speech Command Recognition (SCR), a branch of Automatic Speech Recognition (ASR), is being researched and developed to assist individuals with dysarthria or Speech disorder, who face difficulties in communication. A fundamental stage in building SCR involves word recognition, classification, and prediction of spoken words. This research aims to develop a lightweight deep learning model named for recognizing Dysarthic speech. The design features a Lightweight Convolutional Neural Network (LCNN) architecture, utilizing Short Time Fourier Transform (STFT) for feature extraction (STFT-LCNN). Re-pitching and noise posisition used for data augmented to add variety of of efficient dataset. The LCNN architecture is chosen for its suitability for implementation on edge devices in the future . In its implementation, the PeloNet model uses 2 convolutional layers, feature extraction using spectrograms, and utilizes a mini dataset from EasyCall for detection, with 5 classification classes including dysarthric speakers and Controlled Dysarthria speakers. This method achieves a modeling accuracy of 82%.

Keywords: *Dysarthic speech Recognition, Deteksi Pelo, SCR, ASR, Keyword Spotting, LCNN, LCNN-STFT*

Untuk mengutip artikelini dengan APAStyle:

Sugandi. A., Irianta. H.R., Gustafi.M.F. (2025). Pengenalan Ucapan Pelo Menggunakan Short Time Fourier Transform dengan Pemodelan Lightweight Convolutional Neural Network. TEKNOLOGI: Jurnal Ilmiah Sistem Informasi,15(1), 32-43: https://doi.org/10.26594/teknologi.v15i1.5625.

1. Pendahuluan

Disartria, atau ucapan pelo, merupakan gangguan bicara yang diakibatkan oleh kerusakan pada sistem saraf yang mempengaruhi kontrol otot yang terlibat dalam produksi suara. Kondisi ini sering kali muncul akibat berbagai penyakit neurologis seperti *stroke*, penyakit Parkinson, atau cedera otak traumatis.

ISSN 2502-3357 (*Online*)
A. Sugandi dkk./TEKNOLOGI 15 (1) 2025 19-31 ISSN 2503-0477 (*Print*)

Dysarthria dapat menyebabkan kesulitan dalam artikulasi, kejelasan suara, dan ritme bicara, yang pada gilirannya mengurangi intelligibility atau keterbacaan ucapan, sehingga mempengaruhi kemampuan

individu untuk berkomunikasi secara efektif (Li, 2023; R, 2023). Penelitian menunjukkan bahwa individu dengan dysarthria sering mengalami frustrasi dalam berkomunikasi, dan hal ini berdampak pada kualitas hidup mereka (Calvo et al., 2020).

Automatic Speech Recognition (ASR) telah muncul sebagai solusi yang menjanjikan untuk membantu individu dengan dysarthria dalam berkomunikasi. ASR dapat mengubah ucapan menjadi teks atau perintah yang dapat diproses oleh perangkat digital, sehingga memberikan alternatif bagi mereka yang mengalami kesulitan berbicara. Namun, tantangan utama dalam pengembangan ASR untuk *dysarthric speech* adalah variabilitas yang tinggi dalam pola bicara individu, yang disebabkan oleh perbedaan dalam tingkat *severity* level pengucapan saat disartria, dan karakteristik bicara masing-masing individu, dan implementasi dalam clinical trial, seperti *Augmentative and alternative communication* (AAC), atau sebagai media rehabilitasi dan fisioterapi (Duffy et al., 2023; Jaddoh et al., 2022). Penelitian terbaru menunjukkan bahwa sistem seperti *speech command recognition* (SCR) dalam rumpun ilmu ASR yang ada saat ini menjadi salah satu topik perhatian peneliti di era nya *smart wearable devices* sudah banyak di gunakan oleh user, karena metode ini dalam ASR mengimplementasikan *keyword spotting* (KWS) menggunakan *deep learning* dalam mengenal pola corpus pada kata secara *lightweight model* atau model yang dirancang se efisien mungkin untuk perangkat *mobile*, internet of things (IoT), dan *edge.*

Dalam konteks rehabilitasi, penggunaan teknologi SCR dan perangkat edge dapat memberikan manfaat signifikan bagi individu dengan dysarthria. Misalnya, sistem SCR yang dirancang khusus untuk mengenali pola bicara dysarthric dapat untuk memberi perintah alat fisioterapi untuk melakukan aktivitas *Activity daily living* (ADL), atau aktivitas sehari hari, seperti penelitian oleh kang, dalam merancang alat fisioterapi *exo glove* (Kang et al., 2019), sementara perangkat edge dan IoT memungkinkan penggunaan teknologi ini dalam memonitoring aktivitas telerehabilitasi dan *feedback* data yang berguna untuk pasien dalam mengevaluasi aktivitas rehabilitasi (Kang et al., 2019). Penelitian lebih lanjut diperlukan untuk mengeksplorasi berbagai metode dan teknologi yang dapat diintegrasikan ke dalam sistem SCR untuk *dysarthric speech*, serta bagaimana perangkat *edge* dapat dioptimalkan untuk mendukung pengguna atau pasien dengan gangguan bicara.

2. State Of The Art (Sudah Zotero)

Pengenalan ucapan bagi individu dengan gangguan bicara, seperti disartria atau pelo, merupakan bidang penelitian yang semakin berkembang, terutama dengan kemajuan dalam teknologi pemodelan *lightweight* dan *edge*. Model-model seperti *MobileNet*, *YOLO*, dan KWS telah menunjukkan potensi yang signifikan dalam meningkatkan akurasi pengenalan ucapan pada perangkat *edge device* (Bouraoui et al., 2017). Penelitian sebelumnya telah mengeksplorasi berbagai dataset untuk melatih model-model ini, termasuk *TORGO*, *UASpeech*, dan *MOCHA-TIMIT*, yang masing-masing memiliki karakteristik unik yang mendukung pengembangan sistem pengenalan ucapan yang lebih baik.

Penelitian yang dilakukan oleh Mahmoud et al. yang membandingkan berbagai platform pengenalan ucapan untuk penilaian afasia, menggunakan TORGO menemukan bahwa model yang dilatih dengan data dari pembicara disartria dapat mencapai akurasi yang signifikan (Li, 2023; Mahmoud et al., 2023).

Dalam konteks model *lightweight*, penelitian oleh Woszczyk et al. menunjukkan bahwa penggunaan Domain *Adversarial Neural Networks* dapat meningkatkan akurasi pengenalan ucapan disartria (Woszczyk et al., 2020). Penelitian ini menyoroti pentingnya teknik adaptasi domain untuk meningkatkan kinerja model dalam menghadapi variasi dalam data ucapan yang disebabkan oleh gangguan bicara.

Selain itu, Shabber menggunakan dataset *TORGO* dan *UASpeech* untuk mengembangkan model klasifikasi yang mencapai akurasi 85%, menunjukkan efektivitas pendekatan *speech biomarker* (Shabber, 2024). Cheng et al. juga menyoroti pentingnya pengenalan ucapan dalam konteks interaksi manusia-mesin, dengan menggunakan model lip-reading yang dapat berfungsi sebagai alternatif bagi individu dengan gangguan bicara (Cheng et al., 2023).

Penelitian ini menunjukkan bahwa teknologi *lip-reading* dapat meningkatkan akurasi pengenalan ucapan dalam situasi di mana suara tidak dapat digunakan, memberikan solusi yang inovatif untuk komunikasi bagi individu dengan disartria. Selain itu, penelitian oleh Kim et al. yang menggunakan dataset *NKI CCRT* dan *TORGO* menunjukkan bahwa pendekatan berbasis pembelajaran mendalam dapat meningkatkan kemampuan klasifikasi ucapan patologis, dengan akurasi yang bervariasi tergantung pada dataset yang digunakan (Kim et al., 2015).

Lebih lanjut, penelitian oleh Hernandez et al. menunjukkan bahwa representasi suara yang diperoleh dari model multibahasa dapat mengurangi *Word Error Rate* (WER) hingga 22% pada dataset *UASpeech*, yang mencakup pembicara dengan disartria akibat *cerebral palsy* (Hernandez et al., 2022).

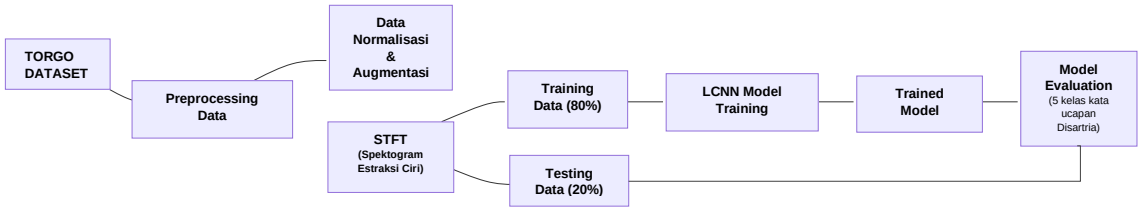
ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Penelitian oleh Morgan et al. menekankan pentingnya evaluasi kata dalam konteks anak-anak dengan gangguan pendengaran, yang menunjukkan bahwa pengenalan ucapan dapat berfungsi sebagai alat bantu komunikasi yang efektif (Morgan et al., 2024). Dengan menggunakan dataset yang relevan dan teknik pemodelan yang tepat, sistem pengenalan ucapan dapat dirancang untuk memenuhi kebutuhan spesifik individu dengan gangguan bicara, memberikan kontribusi yang signifikan terhadap peningkatan kualitas hidup mereka.

Secara keseluruhan, penelitian terdahulu menunjukkan bahwa dengan menggunakan model *lightweight* dan dataset yang sesuai, pengenalan ucapan bagi individu dengan gangguan bicara dapat ditingkatkan secara signifikan. Dengan terus mengembangkan dan mengadaptasi teknologi ini, kita dapat menciptakan solusi yang lebih inklusif dan efektif untuk membantu individu dengan disartria dalam berkomunikasi.

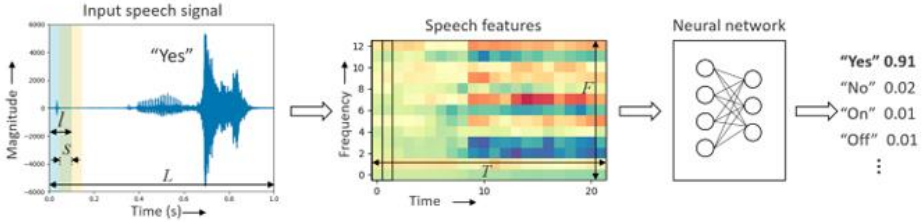
3. Metode Penelitian

Penelitian ini bertujuan untuk mengembangkan model klasifikasi disartria dengan menggunakan arsitektur *Lightweight Convolutional Neural Network* (LCNN), yang dirancang khusus untuk mengidentifikasi lima kelas kata yang diucapkan oleh individu dengan disartria. Dataset yang digunakan dalam penelitian ini *Easycall* Dataset **(Turrisi et al., 2021)**, yang merupakan kumpulan data yang berisi rekaman ucapan dari individu dengan dan tanpa disartria. Proses penelitian dimulai dengan tahapan *preprocessing data*, yang bertujuan untuk normalisasi audio, augmentasi dengan menambah background *noise* agar sampel kata yang di ucap diharapkan dapat berlaku dengan lingkungan *noise* yang terjadi di sekitaran pengucap. Proses ini penting karena kualitas sinyal yang baik akan berkontribusi pada akurasi model yang lebih tinggi **(Mahmoud et al., 2023)**.



Gambar 1. Blok diagram rancangan diagram STFT-LCNN Model untuk klasifikasi 5 kelas kata dengan disartria

Setelah tahap *preprocessing*, data dinormalisasi dan dilakukan augmentasi. Augmentasi dilakukan dengan cara *noise positioning* dengan *background noise*, dan *pitch shifting* yang bertujuan untuk meningkatkan variasi data, dan membuat model lebih *robust* terhadap *noise* lingkungan. Penelitian sebelumnya menunjukkan bahwa teknik augmentasi dapat meningkatkan kinerja model dalam situasi nyata di mana kebisingan sering terjadi (Musalia, 2023). Ekstraksi fitur dilakukan menggunakan *Short-Time Fourier Transform (STFT)*, yang merupakan metode menghasilkan citra spektrogram yang cukup efektif untuk mendapatkan representasi spektral dari sinyal audio. Parameter STFT yang digunakan dalam penelitian ini adalah *windowing*, *frame size*, *stride*, dan *hop size* yang dirancang untuk menghasilkan resolusi spektrogram yang lebih kompak tanpa kehilangan informasi penting dari data suara ***(Zhang et al., 2018)***.



Gambar 2. Diagram pengenalan ucapan dengan KWS *pipeline* (Zhang et al., 2018)

ISSN 2502-3357 (*Online*)
A. Sugandi dkk./TEKNOLOGI 15 (1) 2025 19-31 ISSN 2503-0477 (*Print*)

Setelah tahap ekstraksi fitur, dataset dibagi menjadi dua bagian: 80% untuk data pelatihan dan 20% untuk data pengujian. Model LCNN yang digunakan memiliki arsitektur yang kompak dan efisien, dirancang untuk menangkap pola spektral yang relevan dari data suara. Penelitian sebelumnya menunjukkan bahwa model berbasis LCNN dapat mencapai akurasi yang tinggi dalam pengenalan ucapan, terutama ketika diterapkan pada data yang telah diproses dengan baik (Dong et al., 2020). Setelah model dilatih, evaluasi dilakukan menggunakan data pengujian untuk mengukur akurasi model dalam mengklasifikasikan lima kelas kata KWS pada *TORGO* dataset pada ucapan penderita disartria, yaitu kata *'no', 'stop', 'uno', 'zero',* untuk total subjek pengucap adalah 36 subjek, dengan kondisi disartria dan tanpa disartria atau *Controlled*.

Evaluasi model dilakukan menggunakan *confusion matrix*, yang merupakan alat penting untuk menilai kinerja model klasifikasi. *Confusion matrix* memberikan gambaran yang jelas tentang jumlah prediksi yang benar dan salah untuk setiap kelas, serta membantu dalam menghitung metrik evaluasi lainnya seperti *precision*, *recall*, dan *F1-score*. Dengan menggunakan *confusion matrix*, kita dapat menganalisis di mana model mengalami kesulitan dalam klasifikasi dan melakukan perbaikan yang diperlukan (Mondal et al., 2022). Selain itu, analisis lebih lanjut dapat dilakukan untuk mengevaluasi dampak dari augmentasi data dan teknik ekstraksi fitur terhadap kinerja model.

Detail pada metodologi yang digunakan dalam penelitian ini mencakup, *preprocessing* data, augmentasi, ekstraksi fitur, hingga pelatihan dan evaluasi model akan di dijelaskan di sub bab selanjutnya. Setiap langkah dirancang untuk memastikan bahwa model yang dihasilkan tidak hanya akurat tetapi juga *robust* terhadap variasi dalam data ucapan yang dihasilkan oleh individu dengan disartria.

3.1. Dataset dan *Preprocessing Dataset*

3.1.1 Dataset

Dataset yang digunakan pada penelitian ini adalah *Easy call Dataset* Yang dikembangkan oleh Universitas Toronto dan *Holland Bloorview Kids Rehabilitation Hospital* untuk penelitian disartria. Subjek terdiri dari 7 individu laki laki dan Perempuan, Setiap pembicara diberi kode dan memiliki direktori masing-masing. Pembicara perempuan memiliki kode yang dimulai dengan huruf 'F', sedangkan pembicara laki-laki memiliki kode yang dimulai dengan huruf 'M'. Jika pembicara adalah anggota kelompok kontrol (yaitu, mereka tidak memiliki disartria), maka huruf 'C' mengikuti kode gender. dengan disartria akibat *cerebral palsy* dan 1 individu dengan ALS, berusia 16–50 tahun (Turrisi et al., 2021). Data akustik dikumpulkan menggunakan mikrofon, sementara data artikulatori direkam dengan *electromagnetic articulography* (EMA). Materi ucapan dalam dataset mencakup tiga kategori: ***Non-Word***, untuk mengukur kontrol prosodi dan konsonan peledak; ***Short Word***, untuk analisis akustik seperti frekuensi formant dan energi suara; serta ***Restricted sentences***, yang digunakan untuk merekam sintaksis lengkap untuk pengenalan suara otomatis (ASR). Stimulus diambil dari berbagai sumber, seperti *TIMIT* (Garofolo et al., n.d.) dan *Yorkston-Beukelman Assessment* (Yorkston & Beukelman, 1978), serta mencakup kombinasi bunyi, kata umum, dan kalimat kompleks. Data ini dirancang untuk mendukung penelitian pengembangan teknologi asistif wicara.

Pade penelitian ini, akan menggunakan ***Short Word*** sebagai kata yang akan di identifikasi, yaitu *'no', 'stop', 'uno', 'zero',* untuk subjek pengucap dengan *code* F01-F09 (*Female* dengan disartria), FC01-09 (*Female Controlled /* tanpa disartria), M01-M09 (*male* dengan disartria), dan MC01-09 (*Male Controlled /* tanpa disartria), dengan total 3912 kata. Format audio yang digunakan adalah .*wav, 16 bit, mono, dan *sampling rate* 16 khz, selama 1 detik, pada pada kondisi level *noise* berkisar 35db dengan format audio .wav. Dataset final dikumpulkan menjadi 1 *folder* dataset dengan rasio 80:10:10 untuk *training*, validasi dan *testing* yang akan di ekstraksi menjadi spektrogram sebagai transormasi citra gambar dari gelombang suara untuk ekstraksi ciri pemodelan LCNN. Augmentasi pertama menggabungkan kelas kata yang di pilih dengan penutur disartia dan kontrol (non disartria), lalu augmentasi kedua, menggabungkan kelas kata yang dipilih dengan background 6 jenis rekaman *background noise*, sebagai tambahan augmentasi data, 2 jenis ucapan lainya dengan total 1304 rekaman dari kata lain yang ditargetkan menjadi *invalid class, dan pitch shifiting* dengan *scale pitch* atas dan bawah. Total dataset yang digunakan untuk pelatihan model adalah 3912, dengan 6 target class ka yaitu vokal *'no', 'stop', 'uno', 'zero'* dan *'invalid'*. Berikut tabel parameter dataset final yang bisa di lihat di tabel 1.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

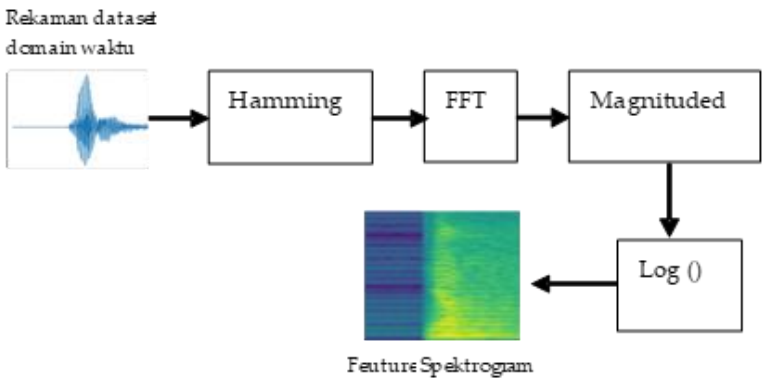
Tabel 1. Properti dataset ucapan kata disartria yang digunakan

VARIABEL DATASET UCAPAN KATA DISARTRIA	INFORMASI
Total Inisial Dataset *Easycall* 5 kelas kata	1800
Total Dataset setelah Augmentasi	3912
Durasi Rekaman	1 s
Lokasi Rekaman	Dalam Ruangan
Sample Rate	16.000 Hz
Bit Depth	16 Bit
Channel	1/Mono
Byterate	32 Kbyte/s
Format	*.wav

Tabel 2. Jumlah dataset ucapan kata disartria per-kelas

KELAS KATA DISARTRIA	*CODENAME* PENUTUR	JUMLAH PENUTUR	DATA LATIH	DATA UJI	DATA VALIDASI
no	*Male (M)*	M = 9	522	65	65
stop	*Male Controlled (Mc)*	Mc = 9	522	65	65
uno	*Female (F)*	F = 9	522	65	65
zero	*Female Controlled (Fc)*	Fc = 9	522	65	65
Invalid (2 kata)			1044	130	130
Jumlah total		36	3132	390	390

diskrit. Beberapa contohnya adalah transformasi *Short Time Fourier Transform* . STFT adalah transformasi dengan *discreate fourier transform* (DFT) yang digunakan untuk menentukan frekuensi sinusoidal pada bagian lokal dari sinyal seiring dengan berubahnya sinyal tersebut terhadap waktu. Dengan kata lain, STFT adalah transformasi fourier di sinyal berjendela (*windowed signal*). STFT memberikan informasi lokal (terhadap waktu) dari komponen frekuensi, berbeda dengan transformasi fourier standar yang menyediakan informasi frekuensi di sepanjang interval waktunya (Goodwin, n.d.)



Gambar 4. Blok diagram Transformasi STFT untuk *Feuture* Pemodelan

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Pada penelitian ini konfigurasi komputasi operator dengan STFT dilakukan menggunakan *library numpy* dan *scipy* pada python dengan konfigurasi fungsi *w*(n) menggunakan jenis *hamming window* dengan overlap 50% di setiap *frame* atau *window size* , *sampling rate* menggunakan 16 khz, setiap *window* memiliki *window size* 320 titik Amplitudo untuk durasi cuplik per *frame blocking* adalah 20ms tanpa *overlap* , dengan overlap 50% disetiap *frame blocking* nya, didapatkan nilai *stride* pada amplitudo atau stride size adalah 160 titik amplitudo ,dicuplik 10 ms untuk setiap frekuensi bin, lalu setiap konfigurasi dilakukan *windowing* untuk waktu interal sinyal suara selama 1 detik, lalu representasi frekuensi bin melalui komputasi, untuk di transormasikan dengan algoritma STFT menjadi besaraan spektral berupa spektrogram. Output spektrogram kemudian di *upscalling* secara *logarithmic* pada spektogram (log-spektogram) dilakukan, menggunakan *average pooling*, pengurangan pixel dengan kernel 1x6 dimaksudkkan meningkatkan representasi fitur dari spektrogram pixel awal menjadi 99x43 pixel sebagai input utama pada model CNN dengan *recall* semua atribut konfigurasi tersebut didialam satu method *function get spectrogram*. Berikut *pseuducode function* nya :

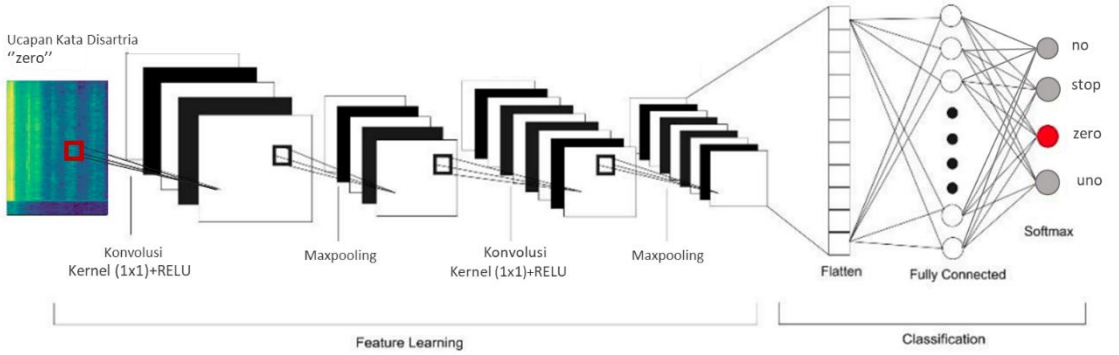
```
def get_spectrogram(audio):
    # normalisasi audio
    audio = audio - np.mean(audio)
    audio = audio / np.max(np.abs(audio))
    # konfigurasi spectrogram
    spectrogram = audio_ops.audio_spectrogram(audio,
                                              window_size=320,
                                              stride=160,
                                              magnitude_squared=True).numpy()
    # Log Spektrogram
    spectrogram = tf.nn.pool(
        input=tf.expand_dims(spectrogram, -1),
        window_shape=[1, 6],
        strides=[1, 6],
        pooling_type='AVG',
        padding='SAME')
    spectrogram = tf.squeeze(spectrogram, axis=0)
    spectrogram = np.log10(spectrogram + 1e-6)
    return spectrogram

```

Gambar 5. Pseudo Code untuk *generate* spectrogram

2.1.3 Convolutional Neural Network (CNN)

CNN adalah salah satu jenis jaringan saraf yang biasa digunakan dalam pengolahan data citra (Tian et al., 2020). CNN juga dapat digunakan untuk mendeteksi dan mengenali objek berupa gambar. CNN tidak jauh berbeda dari jaringan saraf biasa yang terdiri dari neuron yang memiliki bobot, bias dan fungsi aktivasi. CNN termasuk dalam jenis *deep neural network* karena memiliki tingkat jaringan lebih banyak dan secara luas banyak diterapkan pada data citra. Metode CNN memiliki dua tahapan, yaitu tahap klasifikasi menggunakan *feedforward* dan tahap pembelajaran menggunakan *backpropagation*. Cara kerja CNN memiliki kesamaan dengan *multi layer perceptron* (MLP), akan tetapi setiap neuron pada CNN disajikan dalam bentuk dua dimensi. Berbeda halnya dengan MLP, dimana setiap neuron hanya memilki satu dimensi.

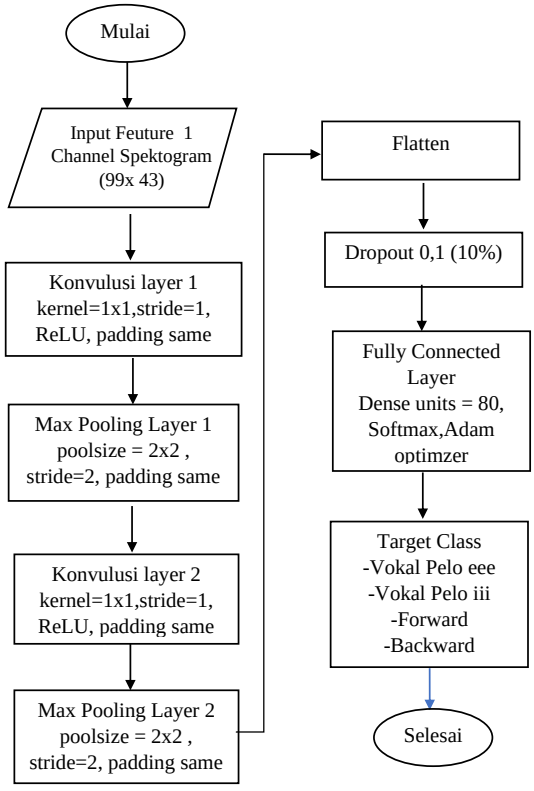


Gambar 6. Blok Diagram Arsitektur Pemodelan *STFT - LCNN*

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

2.1.4 Arsitektur CNN dan Training pemodelan

Pada penelitian ini arsitektur pemodelan di bangun menggunakan *jupyter notebook* dengan *library* tensorflow dan keras, konfigurasi pemodelan terinspirasi dari hasil *rule of thumb* yang dilakukan oleh *google keyword spotting.* (Sainath & Parada, 2015).



Gambar 7. Diagram Alir Arsitektur Pemodelan *STFT-LCNN*

kemudian peneliti melakukan *forking* terhadap arsitektur tersebut. Konfigurasi lalu disusun oleh peneliti berurutan atau *sequential* dengan beberapa lapisan *layer* diantaranya, 2 layer konvolusi dengan 4 filter dengan kernel 3x3, 2 layer *maxpooling* untuk *downsampling output shape* dengan *pool Size* 2x2 dan fungsi aktifasi ReLU, layer selanjutnya adalah *fully connected layer* atau MLP diantaranya, *flatten* layer, *dense* layer dengan jumlah neuron, 80 *node unit* dan *dropout* bobot neuron sebesar 0.1 atau 10% dan fungsi aktifasi yang digunakan adalah *Softmax*. Pencegahan *overfitting* selain melaksanakan augmentasi dataset dan konfigurasi pada MLP, pada arsitektur digunakan metode kernel *tuning* untuk menentukan dan membatasi nilai bobot neuron pada proses konvulusi atau *weight intillizer* salah satunya, regulisasi pada bobot kernel atau filter dengan L2 *regulizer*, dan diketauhi menggunakan *threshold* dengan ukuran kernel 1x1 pada proses konvulusi. Adapun citra spektrogram pada dataset akhir atau *feuture* untuk pemodelan adalah 99 x 43 pixel. Kemudian model di *fitting* dengan nomer *epoch* adalah 10 dan dilakukan per 30 *batch*, Dan berikut hasil *feuture map* dari dengan ringkasan urutan algoritma:

a) Input Feuture

Input pada arsitektur CNN, merupakan citra yang berukuran [99 x 43]. Citra input tersebut adalah representasi sinyal suara yang diambil dari operasi STFT dengan durasi total 1s dan konfigurasi yang telah di ternagkan sebelumnya.

b) Layer Konvulusi Pertama

operasi konvulusi pada penelitian ini adalah *single channel* dilakuakan di setiap layer konvulusi sehingga menghasilkan nilai *feuture map*. Pada Layer Konvulusi yang pertama, citra input akan

ISSN 2502-3357 (*Online*)
A. Sugandi dkk./TEKNOLOGI 15 (1) 2025 19-31 ISSN 2503-0477 (*Print*)

dikonvolusikan oleh kernel berdimensi [1 x 1], dengan jumlah filter 4, *stride* berukuran [1 x 1], dan *padding* "*same*"dengan nilai 0. Jika *Padding* yang digunakan adalah "*same*", maka sistem akan memberlakukan aturan *zero padding* terhadap input matriks.

Dikarenakan *filter size* yang digunakan l2 regulizer, dari persamaan *output shape* pada konvulusi didapatkan dan ukuran stride adalah [1 x 1], maka *output shape* dari konvulusi layer pertama adalah sama dengan ukuran input shape dari citra input yakni [99 x 43]. Fungsi aktifasi menggunakan ReLU layer yang bertujuan untuk mengubah nilai minus pada output proses konvolusi menjadi nol, dikarenakan non linearitas.

c) Layer *Maxpooling* Pertama

Pada Layer *Maxpooling* pertama *input shape* nya adalah *output shape* konvulusi pertama. Pada layer ini, gambar akan di *downsampling* menggunakan *pool size* berukuran [2 x 2], stride [2 x 2], dan *PaddingMode* 'same'. Nilai dari *padding* pada layer ini seperti yang terdapat pada layer konvulusi. Sedangkan *output size* pada *max pooling layer* adalah [49 x 21].

b) Layer Konvulusi Kedua

Proses Kovolusi kedua yaitu meneruskan hasil dari proses pooling pertama yakni dengan input matriks gambar sebesar [49 x 21] dengan konfigurasi yang sama dengan konvulusi pertama sehingga *output shape* tetap [49 x 21].

c) Layer *Maxpooling* Kedua

Pada Layer *Maxpooling* kedua *input shape* nya adalah *output shape* konvulusi kedua , menggunakan konfigurasi yang sama dengan *maxpooling* pertama dan menghasilkan *output shape* [24 x 10].

c) Layer *Backpropagation*

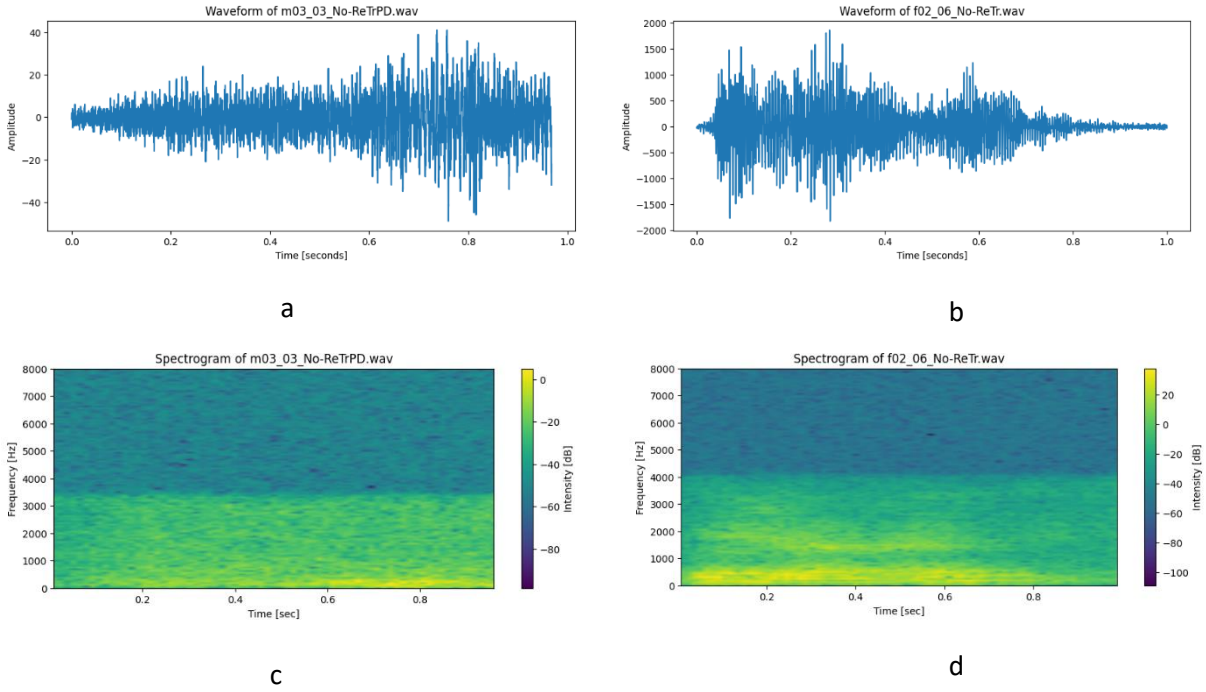
Selanjutnya proses *training* layer *fully connected* dengan perhitungan final *dot product* bobot dan bias dengan teknik *backpropagation*, dengan parameter *loss cross entropy* dan autokorelasi *dot product* dengan operasi *stoastic gradient descent* (SGD), serta transformasi bobot dengan fungsi *Flatten*. Pada tahap ini digunakan mengubah *output pooling* layer menjadi sebuah vektor 1 dimensi. proses propagasi dan klasifikasi atau memprediksi gambar, juga dilakukan proses *Dropout* denga nilai 0.1 atau 10%, sebuah teknik regulasi *neural network* dengan tujuan memilih beberapa neuron secara acak dan tidak akan dipakai selama proses pelatihan, dengan kata lain neuron-neuron tersebut dibuang secara acak. Tujuan dari proses ini yaitu mengurangi *overfitting* pada saat proses *training*. Selanjutnya proses *fully conncected* lainya menggunakan *dense* dengan 80 *units node* dan layer menggunakan aktifasi *softmax*, dan layer ini menjadi layer terakhir yang akan menghitung probabilitas gambar input terhadap semua kelas target yang memungkinkan dan kemudian akan menentukan kelas target berdasarkan input yang diberikan.

4. Hasil dan Pembahasan

3.1 Citra Spektrogram

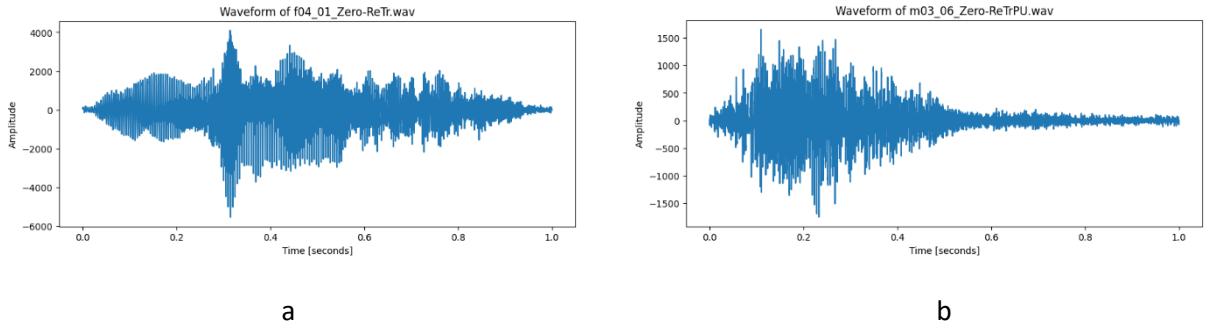
Data pada penelitian ini yang ditampilkan, hanya proses awal dan akhir pada mekanisme STFT berupa sinyal pada domain waktu dari dataset dan hasil transormasi sinyal dengan algoritma STFT pada sinyal untuk *feuture* pemodelan pada total waktu 1 detik. Seperti pada grafik-grafik dibawah adalah hasil representasi mekanisme berupa spektrogram yang dilakukan komputasi tanpa *downsampling* yang dapat di identifikasi secara pola gambar nya, ataupun dari nilai citra pixelnya untuk ML. Identitas ciri setiap sinyal akan berbeda dengan ditentukanya nlai besaran *envelope* amplitude atau frekuensi bin pada setiap sinyal sehingga layer konvulusi, pembelajaran *backpropragation* di layer MLP atau *fully connected* hingga layer output akan mendapatkan prediksi nilai bobot neuron pada *target class* nya masing-masing.

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

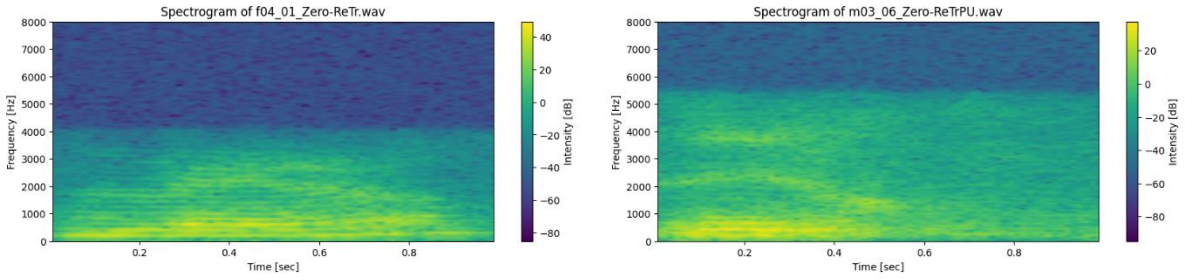


Gambar 8. (a) Sinyal Domain Waktu *''No-M03''* . (b) Sinyal Domain Waktu *"No-M06"*. (c) Spektrogram *'No-M03''*. (d) Spektrogram *"No-M06"*.

Pada transformasi grafik sinyal suara untuk sampel pertama ucapan *forward* diatas (3.1), jika dilihat dari spektrogram, aktifitas amplitudo yang tinggi berada pada interval frekuensi bin 50-2000 Hz di waktu 200-550 ms, untuk sampel ucapan ke 2 berada pada interval frekuensi bin 50-2000 Hz di waktu 200-850 ms , terdapat sedikit perbedaan terhadap besaran amplitudo terhadap waktu tetapi tetap terlihat indentik, sedangkan aktifitas amplitudo lainya pada waktu lainya juga merepresentasikan ciri pada sinyal sampel suara *"No-m03''* tersebut saat dilakukan komputasi terhadap bobot yang akan digunakan pada pemodelan. Sehingga jika dibandingkan diantara 2 sampel ucapan *"No-m03''* pada plot spectrogram-1 dan spectrogram-2 dapat pada analisa tanpa menggunakan pemodelan dapat dilihat ciri karakter sinyal suara pada *output shape* nya.



ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)



c d

Gambar 9. (a) Sinyal Domain Waktu *" Zero-F04"*. (b) Sinyal Domain Waktu *"Zero-M03"*. (c) Spektrogram *"Zero-F04"*. (d) Spektrogram *"Zero-M03"*.

untuk sampel pertama ucapan *"Zero-F04"* diatas, jika dilihat dapat mudah dibedakan secara pola dengan vokal tutur non pelo diatas. Pola gambar yang terlihat konstan dikarenakan fonem tunggal diucapkan seperti tunak atau kontinu pada level frekuensi yang dengan intensitas vokal yang bisa selaras ataupun tidak, menyesuaikan kondisi pengucap. Tetapi jika dilihat aktifitas amplitudonya, untuk tutur pertama, yang tinggi berada pada interval frekuensi bin 10-256 Hz di waktu 200-800 ms, untuk sampel ucapan ke 2 berada pada interval frekuensi bin 10-200 Hz di waktu 20-80 ms , terdapat sedikit perbedaan terhadap besaran amplitudo pada frekuensi tertentu menandakan intensitas pengucap yang bisa tinggi/keras ataupun lemah, tetapi aktifitas amplitudo yang identik dapat terlihat pada frekuensi bin 256Hz dan 3kHz untuk meidentifikasi ucapan tutur vokal pelo *"Zero-F04"* dari tutur vokal pelo yang *"Zero-M03"*.

3.2 Hasil *Training* dan Validasi pemodelan

Hasil *training* pemodelan CNN yang dilakukan pada penelitian di setiap batch dan epoch atau iterasi dengan jumlah 10 *epoch* dan melakukan *data sate* dibagi per *batch size* sebanyak 32 *state* di setiap *epoch*.

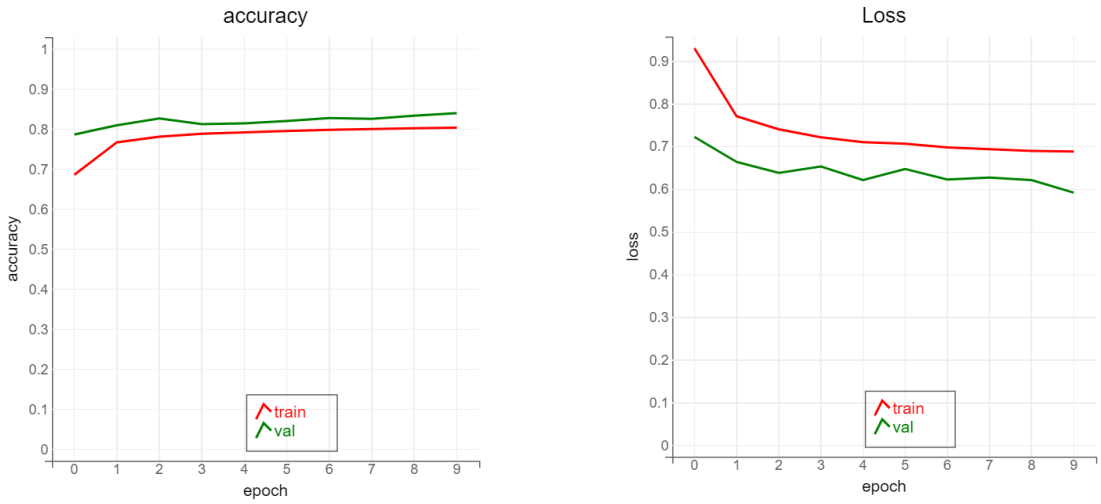
Tabel 3. Hasil Akurasi Dan Loss pada training pemodelan

EPOCH	Loss (%)	Val-loss(%)	Akurasi (%)	Val-Akurasi (S)
0	0.910559	0.673758	0.693069	0.920348
1	0.737471	0.625774	0.782089	0.953298
2	0.708824	0.624831	0.796149	0.957895
3	0.690617	0.601953	0.802896	0.960427
4	0.681507	0.596912	0.805931	0.961901
5	0.674795	0.583675	0.810072	0.962857
6	0.667854	0.592213	0.812292	0.963857
7	0.666737	0.592199	0.812673	0.963991
8	0.662839	0.554603	0.813053	0.964530
9	0.658238	0.557422	0.815582	0.965149

Diketauhui bahwa pada hasil tabel diatas (3.1) dan pada grafik dibawah (3.5) adalah akumulasi hasil data *training* dan data validasi dengan rasio 80:10% di setiap *epoch* yang menunjukkan nilai data validasi tidak lebih besar dari nilai data *training* atau yang direpresentasikan pada tabel adalah Akurasi dan *loss* sedangkan data validasi adalah validaasi-*loss* dan validasi-*akurasi* , dan hal tersebut menyatakan bahwa

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

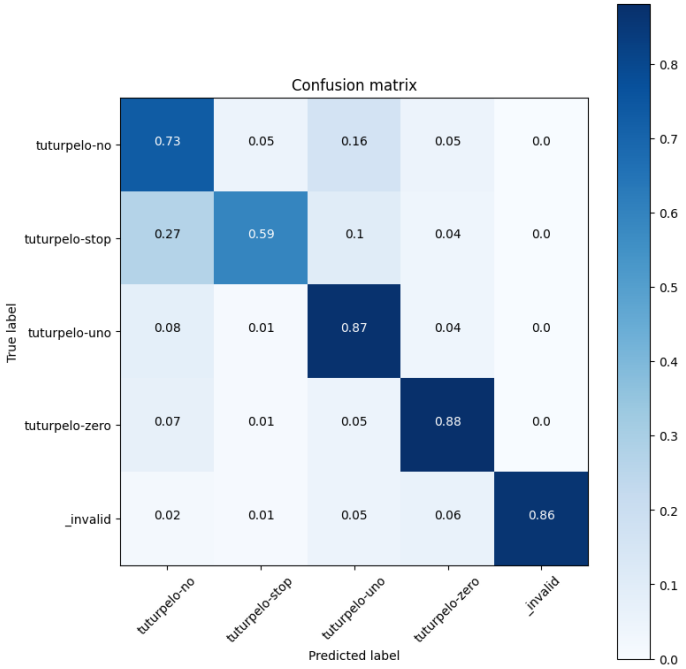
pemodelan dengan arsitektur yang digunakan selain mendapatkan nilai akurasi tertinggi adaalah 94% dengan nilai loss terkecil 3.2%, artinya pemodelan pada hasil *training* tidak mengalami *overfitting.*



Gambar 10. Grafik nilai akurasi dan *loss* pemodelan STFT-LCNN.

3.2 Evaluasi *Testing* Pemodelan

Seperti telah disebutkan pada sub bab persiapan dataset, komposisi data *training*, *validasi* dan *testing* terbagi menjadi rasio 80:10:10 persen. 10 % pada data *testing* tidak sama dengan 10% pada data validasi maupun *training*.



Gambar 11. Matriks *Confusion matrix* Pemodelan STFT-LCNN.

ISSN 2502-3357 (*Online*)
A. Sugandi dkk./TEKNOLOGI 15 (1) 2025 19-31 ISSN 2503-0477 (*Print*)

Dengan pendekatan *lightwetight* model mengunakan 2 konvulusi, dan 2 *maxpooling*, dengan total parameter 77,473, model sudah cukup mampu mendeteksi jenis ucapan kata dengan disartria dengan

akurasi 82%, adapun pendekatan *tuning* model dengan *hyperparameter tuning* menggunakan *gridsearch* , ataupun teknik teknik *preprocessing* dan augmentasi data lainya yang memiliki step komputasi yang lebih banyak seperti esktraksi ciri suara dengan *Mel Frequency Cepstral Coefficient* MFCC dan *Linear Predictive Coding* LPC akan menjadi perhatian penulis untuk di digunakan dalam meningkatkan akurasi sampai dengan diatas 90 %, Jiga dengan teknik evaluasi seperti *F1 score*, *Recall*, *Precision, Word of Error* (WER) , juga akan di pertimbangkan agar dapat menyajikan informasi hasil pemodelan yang lebih akurat dan lengkap.

5. Kesimpulan

Berdasarkan confusion matrix yang ditampilkan, dapat disimpulkan bahwa model klasifikasi secara umum telah menunjukkan kinerja yang cukup baik dalam mengenali lima kelas utama, yaitu *"tuturpelo-no"*, *"tuturpelo-stop", "tuturpelo-uno", "tuturpelo-zero", dan "_invalid"*. Kelas dengan tingkat akurasi tertinggi adalah tuturpelo-zero dengan 88% prediksi yang benar, diikuti oleh *"tuturpelo-uno"* dan *"_invalid*" masing-masing dengan 87% dan 86%. Namun, kelas *"tuturpelo-stop"* menunjukkan performa yang paling rendah, dengan hanya 59% prediksi yang benar dan sebagian besar kesalahan terjadi karena model sering mengklasifikasikan kelas ini sebagai *"tuturpelo-no"*. Hal ini mengindikasikan adanya kemiripan fitur antara kedua kelas tersebut atau kemungkinan distribusi data pelatihan yang tidak seimbang. Kesalahan klasifikasi lainnya juga terjadi dalam jumlah kecil antara kelas-kelas yang lain, namun masih dalam batas yang wajar. Secara keseluruhan, model sudah cukup mampu membedakan masing-masing kelas, namun perlu dilakukan evaluasi lebih lanjut terutama pada kelas tuturpelo-stop, baik dari sisi representasi fitur maupun keseimbangan data, agar akurasi model dapat ditingkatkan secara merata di semua kelas.

6. Referensi

Bouraoui, H., Jerad, C., Chattopadhyay, A., & Hadj-Alouane, N. B. (2017). Hardware Architectures for Embedded Speaker Recognition Applications: A Survey. *ACM Transactions on Embedded Computing Systems*, *16*(3), 1–28. https://doi.org/10.1145/2975161

Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A. B., Grajzer, M., Gilardone, M., & Corbo, M. (2020). Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech. *Folia Phoniatrica Et Logopaedica*. https://doi.org/10.1159/000511042

Cheng, L., Fang, G., Wei, L., Gao, W., Wang, X., Lv, Z., Xu, W., Ding, C., Wu, H., Zhang, W.-A., & Liu, A. (2023). Laser-Induced Graphene Strain Sensor for Conformable Lip-Reading Recognition and Human–Machine Interaction. *Acs Applied Nano Materials*. https://doi.org/10.1021/acsanm.3c00410

Duffy, J. R., Martin, P. R., Clark, H. M., Utianski, R. L., Strand, E. A., Whitwell, J. L., & Josephs, K. A. (2023). The Apraxia of Speech Rating Scale: Reliability, Validity, and Utility. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/2022_ajslp-22-00148

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (n.d.). *Acoustic-Phonetic Continuous Speech Corpus CD-ROM*.

Goodwin, M. M. (n.d.). The STFT, Sin 12. The STFT, Sinusoidal Models, and Speech Modification. *Part B*.

Hernandez, A., Pérez-Toro, P. A., Nöth, E., Orozco-Arroyave, J. R., Maier, A., & Yang, S. H. (2022). *Cross-Lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition*. https://doi.org/10.48550/arxiv.2204.01670

ISSN 2502-3357 (*Online*)
ISSN 2503-0477 (*Print*)

Jaddoh, A., Loizides, F., & Rana, O. (2022). Interaction Between People With Dysarthria and Speech Recognition Systems: A Review. *Assistive Technology*. https://doi.org/10.1080/10400435.2022.2061085

Kang, B. B., Choi, H., Lee, H., & Cho, K.-J. (2019). Exo-Glove Poly II: A Polymer-Based Soft Wearable Robot for the Hand with a Tendon-Driven Actuation System. *Soft Robotics*, *6*(2), 214–227. https://doi.org/10.1089/soro.2018.0006

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. (2015). Automatic Intelligibility Classification of Sentence-Level Pathological Speech. *Computer Speech & Language*. https://doi.org/10.1016/j.csl.2014.02.001

Li, Y. (2023). Improving Text-Independent Forced Alignment to Support Speech-Language Pathologists With Phonetic Transcription. *Sensors*. https://doi.org/10.3390/s23249650

Mahmoud, S. A., Pallaud, R. F., Kumar, A., Faisal, S., Wang, Y., & Fang, Q. (2023). A Comparative Investigation of Automatic Speech Recognition Platforms for Aphasia Assessment Batteries. *Sensors*. https://doi.org/10.3390/s23020857

Morgan, A. E., El-Geidy, S., Amer, A., El-Tawwab, M. A., & Ismail, E. I. (2024). Word Recognition in Relation to Phoniatric Evaluation in Aided Hearing-Impaired Children. *Egyptian Journal of Ear Nose Throat and Allied Sciences*. https://doi.org/10.21608/ejentas.2023.188115.1602

R, V. (2023). *Empowering Dysarthric Communication: Hybrid Transformer-CTC Based Speech Recognition System*. https://doi.org/10.21203/rs.3.rs-3736965/v1

Sainath, T. N., & Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. *Interspeech 2015*, 1478–1482. https://doi.org/10.21437/Interspeech.2015-352

Shabber, S. M. (2024). AFM Signal Model for Dysarthric Speech Classification Using Speech Biomarkers. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2024.1346297

Tian, C., Zhuge, R., Wu, Z., Xu, Y., Zuo, W., Chen, C., & Lin, C.-W. (2020). Lightweight image super-resolution with enhanced CNN. *Knowledge-Based Systems*, *205*, 106235. https://doi.org/10.1016/j.knosys.2020.106235

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., Fadiga, L., & Badino, L. (2021). EasyCall corpus: A dysarthric speech dataset. *Interspeech 2021*, 41–45. https://doi.org/10.21437/Interspeech.2021-549

Woszczyk, D., Petridis, S., & Millard, D. E. (2020). *Domain Adversarial Neural Networks for Dysarthric Speech Recognition*. https://doi.org/10.21437/interspeech.2020-2845

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders*, *11*(6), 499–512. https://doi.org/10.1016/0021-9924(78)90024-2

Zhang, Y., Suda, N., Lai, L., & Chandra, V. (2018). *Hello Edge: Keyword Spotting on Microcontrollers* (arXiv:1711.07128). arXiv. https://doi.org/10.48550/arXiv.1711.07128