# IMPLEMENTASI ALGORITMA *LONG SHORT-TERM MEMORY* UNTUK MENGONVERSI *SPEECH TO TEXT* BAHASA INDONESIA

## SKRIPSI


**NURUL SALSABILA SYAM**

**H071191054**

**PROGRAM STUDI SISTEM INFORMASI**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**MAKASSAR**


**2023**

# IMPLEMENTASI ALGORITMA *LONG SHORT-TERM MEMORY* UNTUK MENGONVERSI *SPEECH TO TEXT* BAHASA INDONESIA

## SKRIPSI

**Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi Departemen Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin**


**NURUL SALSABILA SYAM**

**H071191054**

**PROGRAM STUDI SISTEM INFORMASI**

**DEPARTEMEN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS HASANUDDIN**

**MAKASSAR**


**2023**

# LEMBAR PERNYATAAN KEOTENTIKAN

Yang bertandatangan di bawah ini :

Nama : Nurul Salsabila Syam

NIM : H071191054

Program Studi : Sistem Informasi

Jenjang : S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

**Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia**

adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alih tulisan orang lain, dan belum pernah dipublikasikan dalam bentuk apapun.

**Makassar, 18 Agustus 2023**


**Nurul Salsabila Syam**

**NIM. H071191054**

**Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia**

**Disusun dan diajukan oleh**

**NURUL SALSABILA SYAM**

**H071191054**

**Menyetujui,**

| Pembimbing Utama | Pembimbing Pertama |
|---|---|
| **Dr. Eng. Armin Lawi, S.Si., M.Eng.** | **Edy Saputra Rusdi, S.Si., M.Si.** |
| NIP. 197204231995121001 | NIP. 199104102020053001 |

**Ketua Program Studi**

**Dr. Hendra, S.Si., M.Kom.**
NIP. 197601022002121001

# HALAMAN PENGESAHAN

Skripsi ini diajukan oleh :

Nama : Nurul Salsabila Syam

NIM : H071191054

Program Studi : Sistem Informasi

Judul Skripsi : Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia

**Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Sistem Informasi Departemen Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.**

DEWAN PENGUJI

| | | Tanda Tangan |
|---|---|---|
| Ketua | : Dr. Eng. Armin Lawi, S.Si., M.Eng. | (.....) |
| Sekretaris | : Edy Saputra Rusdi, S.Si., M.Si. | (.....) |
| Anggota | : Muhammad Sadno, S.Si., M.Si. | (.....) |
| Anggota | : Ir. Eliyah Acantha Manapa Sampetoding, S.Kom., M.Kom. | (.....) |

Ditetapkan di : Makassar

Tanggal : 18 Agustus 2023

# KATA PENGANTAR

Puji syukur penulis panjatkan kepada Allah SWT, karena dengan iradat dan karunia-Nya sehingga penulisan skripsi dengan judul "Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia" dapat penulis rampungkan dengan baik. Penulis juga mengirimkan salam dan shalawat kepada Rasulullah, Muhammad SAW, yang telah mengantar dan membimbing umat manusia meninggalkan masa jahiliah ke masa yang dipenuhi cahaya Ilahi.

Skripsi yang merupakan tugas akhir ini disusun dan diajukan sebagai salah satu persyaratan dalam menyelesaikan Pendidikan Strata Satu (S1) Sarjana Komputer pada Program Studi Sistem Informasi, Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Dalam penyusunan skripsi ini, penulis banyak mendapatkan bantuan serta bimbingan dari berbagai pihak. Oleh karena ini, pada kesempatan ini penulis dengan kerendahan hati menyampaikan terima terima kasih yang sebesar-besarnya kepada Ayahanda **Prof. Dr. Syamsuddin Toaha, M.Sc.**, dan Ibunda **Nur Rahmatullah, S.Pd.**, sebagai orang tua yang penuh kesabaran dalam mengasuh dan mendidik penulis serta senantiasa memberikan cinta, kasih sayang, nasehat, dukungan, dan doa yang tulus sehingga penulis dapat menyelesaikan skripsi ini. Serta ucapan terima kasih kepada saudara penulis, **Nurul Syahri Ramadhani Syam** dan **Nurul Saffanah Ailah Syam** yang turut memberikan dukungan dan doa dari awal proses penulisan skripsi hingga selesai.

Tak lupa pula, dengan segala kerendahan hati penulis mengucapkan terima kasih yang setulus-tulusnya kepada:

1. Bapak **Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.**, selaku rektor Universitas Hasanuddin Makassar.
2. Bapak **Dr. Eng. Amiruddin, M.Si.**, selaku dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.
3. Bapak **Dr. Hendra, S.Si., M.Kom.**, selaku Ketua Program Studi Sistem Informasi yang senantiasa membantu dan mengarahkan selama masa studi penulis.

4. Bapak **Dr. Eng. Armin Lawi, S.Si., M.Eng.**, selaku dosen pembimbing utama yang telah membimbing dan memberikan arahan dengan penuh kesabaran selama masa penyusunan skripsi.
5. Bapak **Edy Saputra Rusdy, S.Si., M.Si.**, selaku dosen pembimbing pertama sekaligus dosen penasehat akademik yang telah membimbing dan memberikan arahan dengan penuh kesabaran selama masa studi penulis hingga penyusunan skripsi.
6. Bapak **Muhammad Sadno, S.Si., M.Si.**, dan Bapak **Ir. Eliyah Acantha Manapa Sampetoding, S.Kom., M.Kom.**, selaku dosen penguji yang telah memberikan masukan dan saran selama masa penyusunan skripsi.
7. Bapak dan Ibu **Dosen Program Studi Sistem Informasi, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin** untuk ilmu yang sangat bermanfaat yang telah diajarkan kepada penulis selama menempuh pendidikan.
8. Sahabat **"Nolep Menuju Lep"**, yaitu Bila, Septi, Nisa, Izza, dan Uly yang saling memberikan informasi, dukungan, motivasi, dan semangat satu sama lain selama masa masa perkuliahan hingga selesainya penyusunan skripsi.
9. Sahabat **"Ayam Kecil"**, yaitu Indah, Hana, Fathiyyah, Kia, Rizka, Vira, dan Aan yang memberikan dukungan dan menghibur penulis selama masa penyusunan skripsi.
10. Sahabat **"Niu Gengstah"**, yaitu Raniya, Atika, Caca, Aji, Ilham, Abid, dan Mushawwir yang memberikan semangat dan menghibur penulis selama masa penyusunan skripsi.
11. Seluruh teman-teman **Program Studi Sistem Informasi Angkatan 2019** yang telah memberikan bantuan, informasi, dan dukungan serta membersamai penulis selama masa perkuliahan hingga selesainya penyusunan skripsi.

Penulis menyadari bahwa skripsi yang telah dirampungkan ini masih jauh dari sempurna, baik dari aspek substansi maupun cara penyajiannya. Hal itu disebabkan karena keterbatasan pengetahuan dan pengalaman yang dimiliki

oleh penulis sampai saat ini. Oleh karena itu, ketidaksempurnaan skripsi ini mohon dimaklumi dan kritikan serta saran yang membangun sangat diharapkan.

Akhirnya kepada Allah yang Maha Penyayang jua penulis memanjatkan permohonan dan doa kiranya bekenan untuk membalas segala kebaikan dari semua pihak yang telah membantu dalam penyusunan skripsi ini. Harapannya, skripsi ini dapat memberi manfaat yang seluas-luasnya bagi pengembangan ilmu pengetahuan dan teknologi.

Makassar, 18 Agustus 2023

Nurul Salsabila Syam

## PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini :

| | |
|---|---|
| Nama | : Nurul Salsabila Syam |
| NIM | : H071191054 |
| Program Studi | : Sistem Informasi |
| Departemen | : Matematika |
| Fakultas | : Matematika dan Ilmu Pengetahuan Alam |
| Jenis Karya | : Skripsi |

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah berjudul :

**Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia**

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.
Dibuat di Makassar. Pada Tanggal 18 Agustus 2023

Yang menyatakan

(Nurul Salsabila Syam)

# ABSTRAK

Kemajuan sains dan teknologi sekarang ini sudah memungkinkan manusia berkomunikasi dengan komputer, pengirim dan penerima pesan dapat dilakukan oleh manusia dan komputer. Salah satu kecerdasan buatan dalam mengubah suara menjadi teks adalah *speech to text* yang merupakan bagian dari *speech recognition* dan digunakan untuk mengonversi suara menjadi teks. Hasil dari proses s*peech to text* dapat membantu untuk mendapatkan informasi dalam bentuk teks. Pada penelitian ini, suatu model dibangun untuk mengonversi *speech to text* pada data audio dalam Bahasa Indonesia. Penelitian ini menggunakan data primer berupa rekaman audio yang berisi kalimat sederhana dalam Bahasa Indonesia. Ada sebanyak 60 narasumber yang dilibatkan untuk mendapatkan data. Total data rekaman yang digunakan adalah sebanyak 1800 data audio yang dibagi menjadi data latih sebanyak 1440 dan data uji sebanyak 360. Penelitian ini dimulai dari pengambilan data dan dilanjutkan dengan pre-processing data. Algoritma *Long Short-Term Memory* (LSTM) sebagai suatu model digunakan untuk mengonversi *speech to text*. Metrik *Word Error Rate* (WER) digunakan untuk menghitung tingkat kesalahan hasil prediksi. Dari hasil penelitian ini diperoleh rata-rata nilai WER sebesar 0,02 untuk data latih dan rata-rata WER sebesar 0,48 untuk data uji.

**Kata Kunci**: Speech to text, Long Short-Term Memory, Word Error Rate

# ABSTRACT

The advancement of science and technology now allow humans to communicate with computers, sending and receiving messages can be done by humans and computers. One of the artificial intelligences in converting voice into text is Speech to Text which is part of speech recognition and is used to convert voice into text. The results of the Speech to Text process can help to get information in text form. In this research, a model is built to convert speech to text on audio data in Bahasa Indonesia. This research uses primary data in the form of audio recordings containing simple sentences in Indonesian. There were 60 interviewees involved to get the data. The total recording data used is 1800 audio data which is divided into 1400 training data and 360 test data. This research starts with data collection and continues with data pre-processing. Long Short-Term Memory (LSTM) algorithm as a model is used to convert speech to text. Word Error Rate (WER) metric is used to calculate the error rate of prediction results. The results of this study obtained an average WER value of 0.02 for training data and an average WER of 0.48 for test data.

**Keywords**: Speech to text, Long Short-Term Memory, Word Error Rate

# DAFTAR ISI

# DAFTAR GAMBAR

# DAFTAR TABEL

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Komunikasi merupakan suatu proses transmisi informasi dari pengirim kepada penerimanya dengan menggunakan simbol, kata, atau semacamnya melalui suatu media. Dalam berkomunikasi, pengirim dan penerima pesan dapat menggunakan bahasa tulisan, bahasa verbal, ataupun dengan bahasa isyarat. Bentuk komunikasi sehari-hari yang paling sering dilakukan dan efektif adalah dengan menggunakan bahasa verbal. Ketika komunikasi dilakukan dengan menggunakan bahasa verbal, maka organ tubuh yang bertugas dalam pembentukan suara dan menangkap suara mempunyai peran yang sangat penting dalam mengolah dan menangkap pesan yang disampaikan.

Dengan kemajuan sains dan teknologi, pengirim dan penerima pesan tidak hanya dilakukan antar manusia, tetapi sekarang ini pengirim dan penerima pesan dapat dilakukan oleh manusia dan komputer. Proses komunikasi antar komputer dan manusia memerlukan perangkat *input* dan serangkaian proses yang mengolah pesan sehingga diperoleh *output* yang dapat dipahami oleh manusia. Proses mengolah pesan tersebut membutuhkan teknologi yang dapat mengoversi suara manusia supaya dikenali oleh komputer. Salah satu kecerdasan buatan yang merupakan teknologi dalam mengubah suara menjadi teks adalah *Speech to Text*. *Speech to Text* merupakan bagian dari *speech recognition* yang digunakan untuk mengubah suara menjadi teks atau tulisan yang memungkinkan manusia dapat berkomunikasi dengan komputer. Teks atau kalimat yang dihasilkan dari proses *speech to text* dapat membantu manusia mendapatkan informasi dalam bentuk tulisan.

Penerapan *Speech to Text* mulai banyak digunakan di dalam kehidupan sehari-hari, seperti pada *smart home*, aplikasi penerjemah, asisten virtual (Siri, Google Assistant, Alexa, Cortana), rekam medis, dan lain-lain. Salah satu pemanfaatan *speech to text* yang sering dijumpai adalah subtitle di video YouTube. Audio dari video di YouTube sebagai input untuk diproses dan dikonversi menjadi teks, kemudian sistem memberikan output berupa subtitle yang akan muncul pada video.

Subtitle pada video memudahkan penonton untuk memahami isi dari video yang sedang ditonton.

Dalam melakukan penelitian mengenai *speech to text* pada Bahasa Indonesia, terdapat beberapa tantangan tersendiri. Berdasarkan data Litbang Kompas, jumlah penutur Bahasa Indonesia pada tahun 2022 mencapai 269 juta jiwa (Kompas, 2022). Karena Indonesia merupakan negara kepulauan dan memiliki keanekaragaman budaya yang sangat kaya, pengucapan Bahasa Indonesia dapat dipengaruhi oleh logat atau dialek dan intonasi khas dari daerah masing-masing penutur. Hal ini menyebabkan adanya variasi pada cara pengucapan Bahasa Indonesia. Selain itu, Bahasa Indonesia juga memiliki jenis kalimat yang cukup banyak, diantaranya terdapat kalimat pernyataan, kalimat interogatif atau pertanyaan, kalimat imperatif atau perintah, dan lain-lain. Dengan uraian di atas, penulis menyimpulkan bahwa penelitian mengenai konversi *speech to text* pada Bahasa Indonesia masih sangat relevan untuk dikembangkan. Algoritma atau model yang digunakan pada penelitian mengenai konversi *speech to text* diharapkan dapat mengatasi tantangan tersebut.

Penelitian terkait mengenai konversi *speech to text* telah dilakukan dalam berbagai bahasa, antara lain dalam Bahasa Inggris, Bahasa Jepang, Bahasa India, dan Bahasa Indonesia. Penelitian terdahulu terkait dengan *speech recognition* untuk bahasa Indonesia menggunakan algoritma *Deep Neural Network* (DNN) dengan ekstraksi ciri *Mel Frequency Cepstral Coefficient* (MFCC) menghasilkan akurasi sebesar 65% untuk data latih sebanyak 500 kata dan data uji sebanyak 50 kata (Laksono, 2018). Metode Hidden Markov Model juga telah diimplementasikan untuk mengonversi *speech to text* dalam Bahasa Indonesia menggunakan ekstraksi ciri yang sama, yaitu *Mel Frequency Cepstral Coefficient* (MFCC) (Afkari et al., 2019). Ekstraksi ciri MFCC banyak digunakan pada bidang *speech recognition* untuk mengonversi sinyal suara menjadi beberapa parameter karena MFCC melakukan adaptasi dari sistem pendengaran manusia (Putra & Resmawan, 2011).

Berdasarkan latar belakang yang telah diuraikan, penulis tertarik untuk melakukan penelitian mengenai konversi *speech to text* pada Bahasa Indonesia dengan judul "Implementasi Algoritma *Long Short-Term Memory* untuk Mengonversi *Speech to Text* Bahasa Indonesia". Model ini dapat digunakan dalam

perancangan sistem informasi untuk mengonversi *speech to text* pada Bahasa Indonesia. Penelitian ini sekaligus merupakan salah satu syarat untuk menyelesaikan studi pada Program Studi Sistem Informasi, Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

## 1.2 Rumusan Masalah

Berdasarkan judul penelitian, rumusan masalah pada penelitian ini adalah:

1. Bagaimana tahapan pengambilan dan penyiapan data yang diperlukan dalam penelitian ini?
2. Bagaimana membangun model untuk mengonversi *speech to text* pada Bahasa Indonesia?
3. Bagaimana hasil evaluasi kinerja model yang mengonversi *speech to text* pada Bahasa Indonesia?

## 1.3 Batasan Masalah

Batasan masalah pada penelitian adalah:

1. Data yang digunakan pada penelitian ini adalah data primer berupa data rekaman audio dalam Bahasa Indonesia berupa kalimat standar dengan pola S-P-O-K (subjek, predikat, objek, dan keterangan). Narasumber yang dilibatkan untuk mendapatkan data adalah sebanyak 60 orang yang terdiri atas 30 laki-laki dan 30 perempuan. Setiap narasumber diminta menyebutkan sepuluh kalimat yang sudah disiapkan dan diulang sebanyak tiga kali.
2. Dalam melakukan ekstraksi ciri sebagai tahap mengambil informasi penting dari data audio digunakan *Mel Frequency Cepstral Coefficient*
3. Algoritma untuk mengonversi speech to text dalam Bahasa Indonesia digunakan *Long Short-Term Memory (LSTM)*.

## 1.4 Tujuan Penelitian

Tujuan dari penelitian ini antara lain:

1. Untuk mengetahui tahapan pengambilan dan penyiapan data yang digunakan dalam mengonversi *speech to text* pada data audio bahasa Indonesia.
2. Untuk membangun model yang mengonversi *speech to text* pada data audio bahasa Indonesia.
3. Untuk mengetahui hasil evaluasi kinerja model yang mengonversi *speech to text* pada data audio bahasa Indonesia.

# BAB II
# TINJAUAN PUSTAKA

## *2.1 Speech to Text*

*Speech to Text* merupakan metode yang memungkinkan komputer mengenal suara dengan cara mendeteksi kata-kata dan frasa pada input audio yang dihasilkan oleh seseorang atau mesin, dan mengubahnya menjadi format teks yang dapat dibaca (Nagdewani & Jain, 2020). Metode ini memudahkan manusia untuk berkomunikasi dengan sesama manusia maupun dengan mesin atau komputer. Komputer yang dilengkapi dengan aplikasi *speech to text* dapat mengenali dan memahami ucapan manusia melalui perintah suara dan mengubahnya menjadi suatu teks (Khilari & P., 2015).

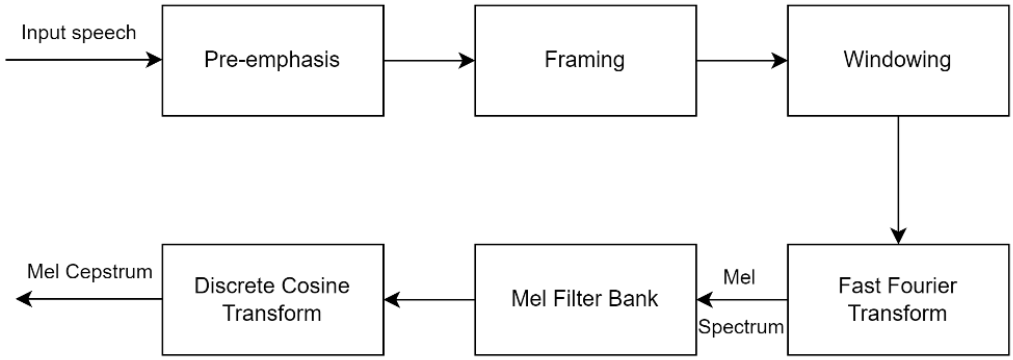
Penggunaan teknologi *speech to text* banyak dimanfaatkan di berbagai aplikasi, seperti pada tugas transkripsi. *Speech to text* digunakan untuk mengubah rekaman suara menjadi teks tertulis, seperti proses transkripsi rekaman seminar. Sehingga memudahkan pengguna aplikasi untuk menghemat waktu dan tenaga dalam mengubah rekaman suara menjadi teks tertulis secara manual. *Speech to text* juga dapat digunakan pada asisten virtual untuk mengontrol perangkat menggunakan suara, seperti mengirim pesan, memutar musik, atau mematikan lampu. Ini memungkinkan pengguna untuk mengontrol perangkat dengan lebih mudah dan praktis.

## 2.2 *Mel Frequency Cepstral Coefficient* (MFCC)

Ekstraksi ciri pada proses *speech to text* merupakan tahap mengekstraksi dan menangani informasi tersembunyi pada sinyal data mentah. Data dapat lebih mudah diolah dengan menerapkan ekstraksi ciri karena menghilangkan fitur-fitur yang tidak dibutuhkan dari data tanpa menghilangkan informasi penting dari data tersebut, (Abdul & Al-Talabani, 2022). Tujuan utama dilakukan ekstraksi ciri adalah untuk mendapatkan ciri dari sinyal suara dan menemukan representasi parameter terbaik dari sinyal.

Ekstraksi ciri yang akan digunakan pada proses konversi *speech to text* ini adalah *Mel Frequency Cepstral Coefficient* (MFCC). MFCC mengadaptasi struktur pendengaran manusia, di mana sinyal suara difilter secara linear untuk frekuensi

rendah (di bawah 1000 Hz) dan secara logaritmik untuk frekuensi tinggi (di atas 1000 Hz) (Putra & Resmawan, 2011). MFCC merupakan teknik ekstraksi ciri yang populer untuk *speech recognition* sehingga MFCC sudah teruji dengan akurasi yang tinggi untuk pengenalan suara (Desai, 2018). Tahapan ekstraksi ciri menggunakan MFCC dapat dilihat pada Gambar 2.1




Gambar 2.1 Tahapan ekstraksi ciri MFCC

1. *Pre-emphasis*

   Pre-emphasis merupakan tahap awal pada proses MFCC. Pre-emphasis merupakan salah satu filter yang digunakan sebelum sinyal diproses. Filter ini mempertahankan frekuensi-frekuensi tinggi pada sebuah spektrum, yang umumnya tereliminasi pada saat proses produksi suara (Tan & Jiang, 2013). Tahap ini dilakukan untuk meningkatkan kualitas sinyal dengan cara mengurangi noise pada sinyal. Pre-emphasis dihitung dengan persamaan sebagai berikut

$$y(n) = x(n) - \propto x(n-1),$$

   di mana $y(n)$ merupakan sinyal hasil pre-emphasis, $x(n)$ merupakan sinyal input, dan $\propto$ merupakan konstanta filter pre-emphasis dengan rentang nilai $0{,}9 \leq \propto \leq 1$.

2. *Frame Blocking*

   Pada tahap *frame blocking*, sinyal hasil pre-emphasis disegmentasi menjadi beberapa frame. Panjang setiap frame dalam kisaran 20 sampai

40 ms. Sinyal suara dibagi menjadi beberapa frame dari $N$ sampel. Frame yang berdekatan dipisahkan oleh $M$ ($M < N$) (Desai, 2018).

3. *Windowing*

   Tahap *frame blocking* pada sinyal suara menyebabkan efek *aliasing* atau munculnya sinyal baru yang memiliki frekuensi yang berbeda dengan sinyal aslinya. Efek tersebut dapat terjadi karena *sampling rate* yang rendah atau karena proses *frame blocking* menyebabkan sinyal menjadi diskontinu (Putra & Resmawan, 2011). *Windowing* dilakukan untuk mengurangi terjadinya efek *aliasing* yang terjadi pada tahap frame blocking. Jenis *windowing* yang paling sering digunakan adalah Hamming Window. Hamming Window diperoleh dengan persamaan sebagai berikut

$$w(n) = 0{,}54 - 0{,}46\ \cos\left(\frac{2\pi n}{N-1}\right),\ \ 0 \le n \le N-1,$$

   di mana $w(n)$ adalah hasil dari *windowing* dan $N$ adalah panjang frame.

4. *Fast Fourier Transform* (FFT)

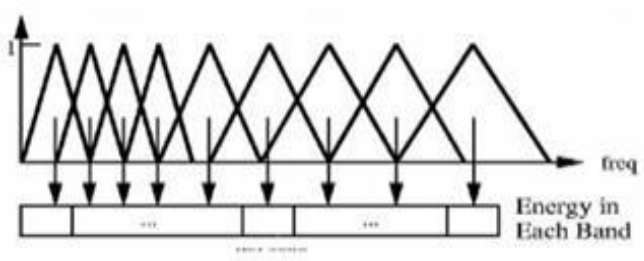
   FFT dilakukan untuk mengonversi sinyal suara dari domain waktu menjadi domain frekuensi. FFT diterapkan pada setiap frame $N$ sampel yang telah melalui tahap *windowing* untuk mendapatkan spektrum sinyal (Amin & Mahmood, 2008). FFT diperoleh dengan persamaan sebagai berikut

$$S_n = \sum_{k=0}^{N-1} S_k e^{-2\pi jkn/N},\ \ 0 \le n \le N-1,$$

   di mana $S_n$ adalah sinyal hasil FFT, $S_k$ adalah nilai sampel sinyal, $N$ adalah jumlah sinyal/sampel yang akan diproses, $j$ adalah bilangan imajiner. Hasil yang diperoleh pada tahap FFT disebut spektrum sinyal.

5. *Mel Filter Bank*

   *Filter bank* adalah salah satu bentuk dari filter yang dilakukan dengan tujuan untuk mengetahui ukuran energi dari frekuensi band tertentu dalam sinyal suara. *Filter bank* dapat diterapkan pada domain waktu maupun pada domain frekuensi. Namun, pada MFCC, *filter bank* harus diterapkan dalam domain frekuensi. *Filter bank* ditunjukkan seperti Gambar 2.2




Gambar 2.2 *Mel Filter Bank*

Sumber: (Desai, 2018)

Sinyal suara tidak mengikuti skala linear. Jadi untuk setiap nada dengan frekuensi, f, diukur dalam Hz, nada subjektif diukur pada skala yang disebut skala mel. Mel frequency scale adalah jarak frekuensi linear di bawah 1000 Hz dan jarak logaritmik di atas 1000 Hz. Persamaan untuk menghitung mel untuk frekuensi tertentu Hz.

$$F(mel) = 2595 \log 10 \left(1 + \frac{f}{700}\right),$$

di mana $F(mel)$ adalah fungsi mel scale.

6. *Discrete Cosine Transform* (DCT)

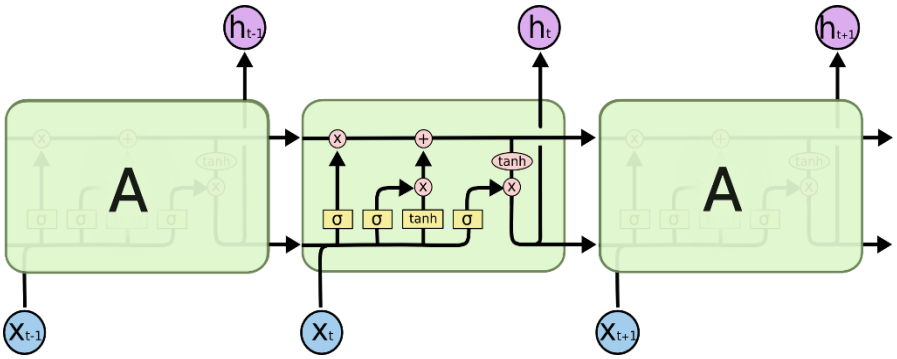
   DCT merupakan tahap akhir pada proses MFCC. Perhitungan menggunakan DCT dilakukan untuk mengonversi nilai *mel spectrum* kembali menjadi domain waktu. Hasil konversi tersebut disebut sebagai koefisien MFCC, (Desai, 2018).

## 2.3 *Recurrent Neural Network* (RNN)

*Recurrent Neural Network* (RNN) merupakan salah satu model *neural network* berulang yang melakukan tugas yang sama pada setiap bagian sekuens dan menyimpan informasi dari input sebelumnya untuk menghasilkan output urutan berikutnya. Nilai *neuron* pada *hidden layer* sebelumnya akan digunakan kembali sebagai data input. Penggunaan *neuron* pada *hidden layer* akan disimpan ke dalam sebuah layer yang dinamakan *context layer*. Nilai *neuron* pada *context layer* akan terus update hingga kondisi RNN terpenuhi (Rizal & Soraya, 2018).

## 2.4 *Long Short-Term Memory* (LSTM)

*Long Short-Term Memory* (LSTM) merupakan salah satu jenis arsitektur *Recurrent Neural Network* (RNN) yang dimodifikasi dengan menambahkan *memory cell* yang dapat menyimpan informasi jangka panjang dalam urutan data. LSTM dikembangkan oleh Hochrieter dan Schmidhuber pada tahun 1997 dan diusulkan sebagai solusi untuk mengatasi terjadinya *vanishing gradient* pada RNN pada saat memproses input data sekuen yang lebih kompleks.




Gambar 2.3 Arsitektur LSTM

Sumber: (Colah, 2015)

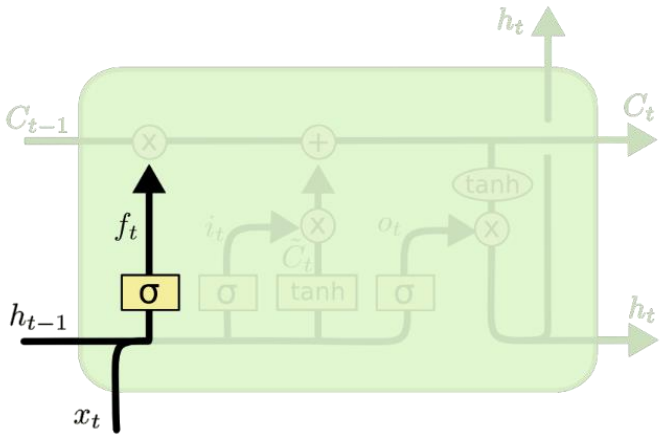
Arsitektur LSTM mencakup tiga gerbang utama, yaitu *forget gate*, *input gate*, dan *output gate*. Proses yang terjadi pada arsitektur LSTM adalah sebagai berikut (Colah, 2015).

1. *Forget Gate*

Gerbang ini menentukan informasi apa yang harus diabaikan dari *cell state*. Keputusan ini dibuat oleh lapisan sigmoid yang disebut *forget gate layer*, yang akan memproses $h_{t-1}$ dan $x_t$ sebagai *input* dan menghasilkan *output* berupa angka 0 dan 1 pada *cell state*. Persamaan *forget gate* $(f_t)$ diberikan oleh

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big),$$

di mana $f_t$ adalah *forget gate*, $\sigma$ adalah fungsi aktivasi *sigmoid*, $W_f$ adalah nilai *weight* untuk *forget gate*, $h_{t-1}$ adalah *state* sebelumnya $(t-1)$, $x_t$ adalah nilai *input* pada saat ini $(t)$, dan $b_f$ adalah nilai *bias* pada *forget gate*.




Gambar 2.4 *Forget Gate*

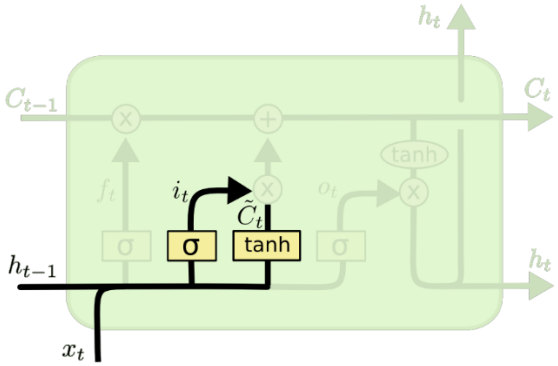
Sumber: (Colah, 2015)

2. *Input Gate*

   Gerbang ini menentukan informasi baru apa yang akan disimpan di *cell state*. Terdapat dua bagian pada gerbang ini, yaitu lapisan sigmoid yang disebut *input gate layer*, yang menentukan nilai mana yang akan diperbarui. Persamaan *input gate* $(i_t)$ diberikan oleh

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),$$

di mana $i_t$ adalah *input gate*, $\sigma$ adalah fungsi aktivasi *sigmoid*, $W_i$ adalah nilai *weight* untuk *input gate*, $h_{t-1}$ adalah *state* sebelumnya $(t-1)$, $x_t$ adalah nilai *input* pada saat ini $(t)$, dan $b_i$ adalah nilai *bias* pada *input gate*.




Gambar 2.5 *Input Gate*

Sumber: (Colah, 2015)

Bagian selanjutnya adalah lapisan *tanh* yang membuat vektor nilai baru yang dapat ditambahkan ke *cell state*. Pada langkah selanjutnya, keduanya digabungkan untuk membuat pembaruan pada *cell state*. Persamaan *candidate cell state* $(\hat{C}_t)$ diberikan oleh

$$\hat{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c),$$

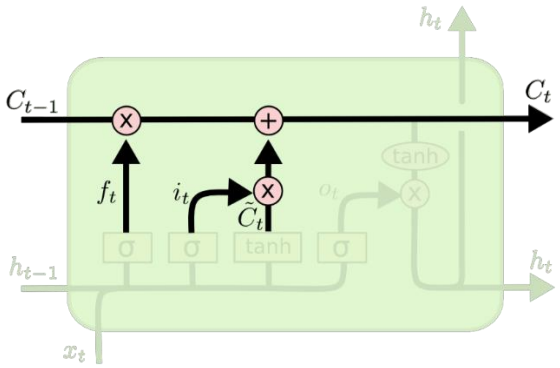
di mana $\hat{C}_t$ adalah *candidate cell state*, $W_c$ adalah nilai *weight* untuk *cell state*, $h_{t-1}$ adalah *state* sebelumnya $(t-1)$, $x_t$ adalah nilai *input* pada saat ini $(t)$, dan $b_c$ adalah nilai *bias* pada *cell state*.

3. *Memory Cell*

   Tahap ini memperbarui *cell state* yang lama dan melepaskan informasi lama kemudian menambahkan dengan informasi baru. Persamaan dari *cell state* $(C_t)$ adalah sebagai berikut.

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t$$

di mana $C_t$ adalah *cell state*, $f_t$ adalah *forget gate*, $C_{t-1}$ adalah *cell state* sebelumnya $(t-1)$, $i_t$ adalah *input gate*, dan $\hat{C}_t$ adalah *candidate cell state*.




Gambar 2.6 *Memory Cell*

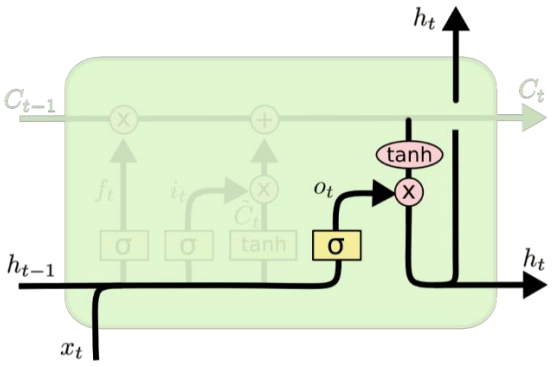
Sumber: (Colah, 2015)

4. *Output Gate*

   Gerbang ini menjalankan lapisan sigmoid yang menentukan bagian *cell state* mana yang menjadi output. Persamaan *output gate* $(O_t)$ diberikan oleh

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$

   di mana $O_t$ adalah *output gate*, $\sigma$ adalah fungsi aktivasi *sigmoid*, $W_o$ adalah nilai *weight* untuk *output gate*, $h_{t-1}$ adalah *state* sebelumnya $(t-1)$, $x_t$ adalah nilai *input* pada saat ini $(t)$, dan $b_o$ adalah nilai *bias* pada *output gate*.




Gambar 2.7 *Output Gate*

Sumber: (Colah, 2015)

Langkah selanjutnya adalah menempatkan *cell state* melalui fungsi aktivasi *tanh* (untuk mengubah nilai antara -1 dan 1) dan mengalikannya dengan output dari *sigmoid gate*, sehingga hanya mengeluarkan bagian-bagian yang diputuskan.

$$h_t = O_t * \tanh C_t$$

di mana $h_t$ adalah state pada saat ini $(t)$ atau *output* LSTM, $O_t$ adalah *output gate*, dan $C_t$ adalah *cell state*.

## 2.5 *Word Error Rate* (WER)

*Word Error Rate* (WER) merupakan metrik yang digunakan untuk mengukur tingkat keberhasilan model dengan menghitung tingkat kesalahan kata pada hasil konversi audio. WER menghitung banyaknya kata yang diganti, ditambahkan, dan dihilangkan dari hasil konversi audio.

$$WER = \frac{S+I+D}{N},$$

di mana simbol $S$ adalah penggantian, $I$ adalah penyisipan, $D$ adalah penghapusan, dan $N$ adalah jumlah kata yang diucapkan.

Simbol $S$ menunjukkan penggantian kata yang terjadi jika suatu kata diganti dari kata yang sebenarnya. Contohnya, ketika narasumber mengucapkan kata “saya” tetapi pada hasil konversi diganti menjadi kata “ayah”.

Simbol $I$ menunjukkan penyisipan kata yang terjadi jika suatu kata ditambahkan pada kalimat hasil konversi, tetapi kata tersebut tidak diucapkan oleh narasumber. Contohnya, ketika narasumber mengucapkan kalimat “adik bermain bola” tetapi kalimat hasil konversinya adalah “adik bermain bola basket”. Maka, kata “basket” merupakan kata yang disisipkan.

Simbol $D$ menunjukkan penghapusan kata yang terjadi jika pada suatu kalimat terdapat kata yang dihapus. Contohnya, ketika narasumber mengucapkan kalimat “saya menonton film di kamar”, tetapi kalimat hasil konversinya adalah “saya menonton film kamar". Maka, kata “di” merupakan kata yang dihapus.

Berikut diuraikan contoh perhitungan WER. Jika diberikan kalimat sebenarnya adalah “saya menonton film di kamar” dan kalimat hasil prediksi adalah “ibu menonton film di dapur”. Pada kalimat hasil konversi, terdapat dua kata yang

diganti, yaitu kata “saya” dan kata “kamar”. Sehingga didapatkan nilai WER dari kalimat hasil konversi adalah $WER = \frac{2}{5} = 0{,}4$.

### *2.6 Streamlit*

*Streamlit* adalah *open-source framework* berbasis *python* yang digunakan untuk memudahkan pengembang untuk membangun dan mengembangkan web dengan cepat pada model *machine learning* dan *data science* (Patil & Lokesha, 2022). *Streamlit* memungkinkan untuk membuat web yang interaktif untuk memberikan visualisasi hasil analissi dari model *machine learning*. Pada penelitian ini, dilakukan deployment model untuk melakukan konversi *speech to text* pada Bahasa Indonesia menggunakan *framework streamlit*.

### 2.7 S*oft System Methodology* (SSM)

*Soft System Methodology* (SSM) adalah metodologi sistematis yang digunakan untuk mengembangkan sistem informasi dengan pendekatan terstruktur untuk memahami suatu masalah, membangun mode konseptual, mendapatkan kelayakan dan perubahan yang diinginkan serta mengimplementasikannya.(Sumadyo, 2016)

Peter Checkland mengembangkan SSM yang terdiri dari tujuh tahapan, yaitu:

1. Identifikasi dan pahami situasi masalah
2. Deskripsikan situasi masalah melalui *rich picture*, yaitu membuat sketsa dari situasi masalah
3. Tentukan root definition untuk situasi masalah
4. Melakukan pembuatan model konseptual untuk mengatasi *root definition*
5. Membandingkan model konseptual (pada tahap 4) dengan situasi sebenarnya (pada tahap 2)
6. Identifikasi kemungkinan perubahan strategi untuk meningkatkan sistem
7. Melakukan pengembangan sistem yang sesuai dengan hasil identifikasi (pada tahap 6)

## 2.8 Penelitian Terkait

Penelitian yang berjudul "*Speech to Text* Menggunakan Metode *Hidden Markov Model*" yang dilakukan oleh Muhammad Fariz Taswarul Afkari, Mudhi Irawan dan Surya Michrandi Nasution pada tahun 2019 membuat aplikasi *speech to text* menggunakan metode *Hidden Markov Model* (HMM) hybrid dengan *Gaussian Mixture Model* (GMM) dan ekstraksi ciri dari input sinyal suara yang diterima menggunakan *Mel Frequency Cepstral Coefficient* (MFCC). Terdapat beberapa tahapan desain sistem pada penelitian ini, yaitu merekam audio, membagi data menjadi data latih dan data uji, melakukan ekstraksi ciri, pengenalan kata menggunakan model HMM/GMM, melakukan pencocokan ciri, dan text output akan tampil pada aplikasi. Setelah dilakukan pengujian dengan beberapa parameter terhadap sistem, hasil dari penelitian ini memperoleh akurasi 100% untuk mengenali 10 kata. Akurasi terbaik diperoleh dari hasil pengujian dengan parameter MFCC 13 *feature* dan GMM 6 *mixture*. Dapat disimpulkan bahwa menggunakan metode *hybrid* dapat memperbaiki tingkat performa sistem (Afkari et al., 2019).

Penelitian yang berjudul "*Speech to Text* untuk Bahasa Indonesia" yang dilakukan oleh Teguh Puji Laksono pada tahun 2018 membangun model *speech to text* untuk Bahasa Indonesia dan Bahasa Jawa menggunakan algoritma *Deep Neural Network* (DNN) dan *Connectionist Temporal Classification* (CTC), dan *Mel Frequency Cepstral Coefficient* (MFCC) sebagai ekstraksi ciri. Pada penelitian ini, dilakukan dua tahapan dalam membangun *speech to text*, yaitu tahap manual dan tahap utama. Pada tahap manual, dilakukan pengambilan data dengan merekam audio dari sepuluh narasumber, melabel suara, dan mengganti tipe file menjadi tipe *flac* dan *wav*. Selanjutnya, pada tahap utama dilakukan *pre-processing*, *training*, dan *testing*. Pada tahap *pre-processing*, dilakukan pemrosesan pada data suara dan label untuk diambil nilai MFCC *feature* dan nilai label. Pada tahap *training*, dilakukan perhitungan dari nilai MFCC *feature* dan nilai label untuk diproses menjadi model menggunakan metode *deep learning*. Pada *deep learning*, digunakan multilayer perceptron sebagai layer pembelajaran. Selanjutnya, pada tahap *testing* dilakukan pengujian pada model yang telah dibuat dan melihat akurasi berdasarkan kecocokan kata yang dihasilkan. Dengan data latih sebanyak 500 data

dan data uji sebanyak 50 data, hasil akurasi yang didapatkan untuk data Bahasa Indonesia adalah 65% dan untuk data Bahasa Jawa adalah 57% (Laksono, 2018).

Penelitian dengan judul "*Speech-to-Text Conversion in Indonesian Language Using a Deep Bidirectional Long Short-Term Memory Algorithm*" yang dilakukan oleh Suci Dwijayanti, Muhammad Abid Tami, dan Bhakti Yudho Suprapto pada tahun 2021 membangun model yang mengonversi *speech to text* untuk Bahasa Indonesia menggunakan metode *Deep Bidirectional* LSTM, ekstraksi ciri menggunakan *Spectrogram* dan MFCC, dan *language model 5-gram*. Terdapat beberapa tahapan yang dilakukan pada penelitian ini, dimulai dengan pengambilan data dengan merekam suara dari sepuluh narasumber dan membagi data menjadi data *training* dan data *testing*. Kemudian, pada tahap *preprocessing* dilakukan tiga tahapan, yaitu normalisasi, *silence removal*, dan *pre-emphasis* yang bertujuan untuk mengurangi *noise* dan menghapus *silence* pada sinyal suara. Selanjutnya dilakukan ekstraksi ciri untuk mendapatkan *spectrogram* dan MFCC. Hasil dari proses ekstraksi ciri menjadi input pada proses klasifikasi menggunakan *deep bidirectional* LSTM. *Connectionist Temporal Classification* (CTC) digunakan pada proses *training* untuk mengetahui *loss* pada jaringan *deep bidirectional* LSTM. Pembuatan *language model* menggunakan *corpus text* yang berasal dari surat kabar kompas sehingga diperoleh hasil dari *speech to text.* Hasil dari penelitian ini memperoleh akurasi terbaik pada nilai rata-rata WER sebesar 0,2745% untuk melakukan proses *speech to text* pada Bahasa Indonesia. Untuk ekstraksi ciri, ditunjukkan bahwa MFCC mampu mengekstraksi ciri lebih baik dibandingkan dengan *spectrogram* untuk proses *speech to text* (Dwijayanti et al., 2021).

Penelitian yang berjudul "*Deep Learning Long-Short Term Memory* (LSTM) *for Indonesian Speech Digit Recognition using LPC and MFCC Feature*" yang dilakukan oleh Ericks Rachmat Swedia, Achmad Benny Mutiara, Muhammad Subali, dan Ernastuti pada tahun 2018 membahas tentang bagaimana membangun model *speech recognition* untuk pengenalan digit dalam Bahasa Indonesia menggunakan algoritma LSTM dan ekstraksi ciri menggunakan metode LPC dan MFCC. Data yang digunakan sebagai data *training* sebanyak 7990 rekaman suara yang terdiri atas 389 suara perempuan dan 410 suara laki-laki yang mengucapkan angka 0-9. Penelitian ini dilakukan dengan 2 tahapan utama, yaitu tahap *training*

dan tahap *classification*. Pada tahap *training*, dilakukan ekstraksi ciri pada data menggunakan LPC dan MFCC. Metode LPC mengekstraksi ciri berdasarkan pitch atau frekuensi fundamental, sedangkan metode MFCC mengekstraksi ciri berdasarkan spektrum suara. Hasil ekstraksi fitur LPC dan MFCC digunakan sebagai input untuk menemukan model LSTM terbaik. Selanjutnya, model LSTM yang dihasilkan pada tahap *training* digunakan untuk tahap *classification.* Pada tahap *classification*, model LSTM yang diperoleh dari tahap training kemudian digunakan untuk mengklasifikasikan data *testing*. Data *testing* dimasukkan ke dalam model LSTM dan menghasilkan output teks berupa digit. Hasil dari penelitian ini menunjukkan hasil akurasi menggunakan ekstraksi fitur MFCC sebesar 96,58% pada detik ke 0,78 sedangkan dengan menggunakan ekstraksi fitur LPC sebesar 93,79% pada detik ke 0,70. Dengan demikian dapat disimpulkan pada penelitian ini bahwa model LSTM mampu mengenali digit dalam bahasa Indonesia dan metode MFCC dapat melakukan ekstraksi ciri lebih baik daripada metode LPC. (Swedia et al., 2016)

Penelitian dengan judul "*Nepali Speech Recognition using LSTM-CTC*" yang dilakukan oleh Rupesh Shrestha, Basanta Joshi, dan Suman Sharma pada tahun 2021 membahas tentang sistem pengenalan suara untuk bahasa Nepal. Terdapat beberapa tahap dalam mengerjakan penelitian ini, yaitu pengumpulan data, *pre-processing* data, ekstraksi fitur, pembuatan model menggunakan LSTM-CTC, pelatihan, pengujian, dan validasi data, dan evaluasi hasil. Data yang digunakan berasal dari Nepali Speech Corpus dengan total 2813 file audio yang diucapkan oleh tiga orang narasumber laki-laki. Beberapa audio rekaman tersebut memiliki noise yang kecil dan efek suara bergema karena kondisi saat merekam audio, sehingga dilakukan *pre-processing* untuk menghilangkan noise tersebut. Selanjutnya dilakukan ekstraksi fitur menggunakan metode MFCC pada file audio di mana setiap vektor merepresentasikan informasi. Fitur-fitur tersebut diteruskan ke neural network untuk melakukan pelatihan model. Algoritma LSTM melakukan pembelajaran pola berurutan jangka panjang. Metode CTC membantu menyelaraskan frame ucapan dan target label. Metode CTC kurang berhasil membantu dalam melatih LSTM dari data time series yang tidak selaras antara input frame ucapan dan target karakter label. Hasil dari penelitian ini memperoleh nilai

WER sebesar 40% untuk pengenalan kata tanpa layer CTC dan nilai WER sebesar 34,3% untuk pengenalan kata dengan layer CTC (Shrestha et al., 2021).