RANCANG BANGUN SISTEM NAVIGASI PADA APLIKASI *ROUTE GUIDANCE* UNTUK TUNANETRA BERBASIS *INDOOR POSITIONING*

PROPOSAL

Diajukan untuk melengkapi tugas-tugas dan
memenuhi syarat-syarat guna pelaksanaan penelitian Tugas Akhir

Oleh:

YUDA ADITYA
1608107010030

JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
JUNI, 2022

PENGESAHAN PROPOSAL

RANCANG BANGUN SISTEM NAVIGASI PADA APLIKASI *ROUTE GUIDANCE* UNTUK TUNANETRA BERBASIS *INDOOR POSITIONING*

DESIGN OF NAVIGATION SYSTEM FOR ROUTE GUIDANCE APPLICATION FOR VISUALLY IMPAIRED PERSON BASED ON INDOOR POSITIONING

Oleh:

Nama	: Yuda Aditya
NPM	: 1608107010030
Jurusan	: Informatika

Menyetujui:

Pembimbing I	Pembimbing II
Kurnia Saputra, S.T., M.Sc.	
NIP. 198003262014041001	NIP.

Mengetahui:

Ketua Jurusan Informatika FMIPA
Universitas Syiah Kuala,

Dr. Muhammad Subianto, S.Si., M.Si
NIP. 196812111994031005

KATA PENGANTAR

Segala puji dan syukur kepada Allah SWT yang telah melimpahkan rahmat serta hidayah-Nya kepada kita semua dan juga atas izin-Nya penulis dapat menyelesaikan penulisan proposal ini. Tak lupa Shalawat dan Salam penulis sanjung sajikan kepada Nabi Besar Muhammad SAW, karena beliau telah membawa kita semua dari alam jahiliah ke alam dengan ilmu pengetahuan.

Proposal yang berjudul **"Rancang Bangun Sistem Pengenalan Suara pada Aplikasi *Route Guidance* untuk tunanetra berbasis *Indoor Positioning*"** ini telah dapat diselesaikan atas bantuan banyak pihak. Oleh karena itu, melalui tulisan ini penulis ingin mengucapkan terima kasih dan penghargaan sebesar-besarnya kepada:

1. Bapak Kurnia Saputra, M.Sc., selaku Dosen Pembimbing I dan Bapak Alim Misbullah, S.Si., M.S., selaku Dosen Pembimbing II yang telah banyak memberikan bimbingan dan arahan kepada penulis, sehingga penulis dapat menyelesaikan Proposal ini.
2. Bapak Dr. Muhammad Subianto, M.Si., selaku Ketua Jurusan Informatika.
3. Bapak Muslim Amiren, S.Si., M.InfoTech., selaku Dosen Wali penulis.
4. Seluruh Dosen di Jurusan Informatika Fakultas MIPA atas ilmu dan didikannya selama perkuliahan.
5. Orang tua serta keluarga penulis yang telah membatu dan banyak memberikan dukungan secara spiritual, moral, dan material kepada penulis.
6. Andika Pratama, Budi Gunawan, M. Zikri Aksnana, dan Yuda Aditya selaku teman yang telah banyak memberikan dukungan, masukan serta ilmu yang cukup besar dan bermanfaat dalam penulisan Proposal ini.
7. Seluruh teman-teman seperjuangan Jurusan Informatika Unsyiah 2016 lainnya.

Penulis juga menyadari segala ketidaksempurnaan yang terdapat didalamnya baik dari segi materi, cara, ataupun bahasa yang disajikan. Seiring dengan ini penulis mengharapkan kritik dan saran dari pembaca yang sifatnya dapat berguna untuk kesempurnaan Proposal ini. Harapan penulis semoga tulisan ini dapat bermanfaat bagi banyak pihak dan untuk perkembangan ilmu pengetahuan.

Banda Aceh, November 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Komunikasi antar manusia dengan manusia merupakan cara penyampaian informasi yang efektif untuk saling terhubung dengan lingkungan sekitar. Cara berkomunikasi yang paling sering dilakukan oleh manusia adalah dengan menggunakan media suara atau ucapan, selain itu manusia juga memiliki media lainnya untuk berkomunikasi dengan menggunakan isyarat dan tulisan. Dengan adanya suara pula dapat membantu sejumlah besar orang yang memiliki keterbatasan penglihatan, sehingga sebagai pengganti indra penglihatan mereka mengandalkan atau mempertajam indra pendengaran mereka. Menurut *World Health Organization* dalam laporannya, secara global setidaknya 2,2 miliar orang memiliki gangguan penglihatan atau kebutaan, di mana setidaknya 1 miliar di antaranya memiliki gangguan penglihatan yang dapat dicegah atau belum ditangani (WHO, 2020). Karena atas dasar kemudahan dalam berkomunikasi, teknologi yang berkaitan dengan suara telah banyak dikembangkan. Kemudahan yang dimaksudkan adalah kemudahan manusia dalam berkomunikasi dengan alat teknologi, jadi alat teknologi tersebut dapat mengerti ucapan yang dikeluarkan oleh manusia. Teknologi ini juga dapat membantu orang–orang dengan gangguan penglihatan dalam berbagai aspek kehidupannya.

Speech recognition merupakan salah satu dari bentuk *Artificial Intelligence* atau AI. *Speech recognition* atau *Automatic Speech Recognition* (ASR) merupakan suatu pengembangan teknologi pada komputer agar dapat mengenali dan memahami kata dan frasa yang diucapkan oleh manusia. Kata dan frasa yang diucapkan tersebut akan didigitalisasikan dengan cara mengubah gelombang suara yang diterima menjadi suatu format yang dapat dibaca oleh alat teknologi. Alat teknologi yang berhasil membaca masukan kata dan frasa tersebut dapat mengidentifikasi serta memahami perintah yang diminta merupakan suatu konsep dari *voice command*. Pengenalan ucapan memiliki perbedaan dengan sistem *Text to Speech* (TTS), dimana pada TTS merupakan suatu sistem yang dapat mengubah suatu teks menjadi suara secara otomatis melalui transkripsi grafem-ke-fonem untuk kalimat yang diucapkan. Semakin berjalannya waktu, penelitian terkait dengan *speech recognition* dan *text to speech* semakin berkembang pula. Saat ini penerapan *speech recognition* yang terkenal adalah *Google Assistant* yang dikembangkan oleh perusahaan Google.

Ada dua fase yang dilibatkan dalam sistem ASR, yang pertama yaitu fase pelatihan dan yang kedua fase pengujian (Ouisaadane dkk., 2020). Untuk mendapatkan fitur-fitur yang berbeda seperti halnya konfigurasi kekuatan, nada dan

saluran vokal dari sinyal suara. Informasi yang akurat dari sinyal ucapan yang direkam dan harus lebih menunjukkan kekuatan atau perbedaan terhadap *noise* merupakan vektor fitur yang ideal (Dua dkk., 2018). Ekstraksi fitur dilakukan dengan berbagai teknik, namun para peneliti telah banyak menggunakan *Mel Frequency Cepstrum Coefficients* (MFCC) sebagai metode yang stabil dan terbukti mengekstrak karakteristik yang berbeda dari sinyal suara masukan (Dua dkk., 2018).

Selain ekstraksi fitur, model akustik juga dilakukan untuk mengoptimalkan hasil dari *speech recognition* tersebut. Model akustik merupakan model yang mewakili hubungan antara sinyal audio dan fonem atau unit linguistik yang membentuk ucapan. Model akustik dibuat dengan mengambil database suara yang besar atau yang disebut dengan *speech corpus*. Model akustik juga menggunakan algoritma pelatihan khusus untuk membuat representasi statistik dari tiap fonem. *Gaussian Mixture Models* (GMM) banyak digunakan untuk menentukan seberapa baik setiap status dari setiap *Hidden Markov Model* (HMM) sesuai dengan bingkai atau jendela pendek bingkai koefisien yang mewakili masukan akustik (Hinton dkk., 2012). Namun *Deep Neural Networks* (DNN) yang memiliki banyak *hidden layers* dan telah dilatih dengan metode baru terbukti mengungguli GMM dalam berbagai macam tolak ukur pengenalan ucapan (Hinton dkk., 2012).

Hal-hal yang telah dijabarkan di atas kemudian melatar belakangi penelitian ini. Penelitian ini akan membangun sistem pengenalan suara dengan menggunakan MFCC sebagai ekstraksi fitur serta DNN sebagai model akustik yang akan dipasangkan pada aplikasi *route guidance* untuk pengguna tuna netra berbasis *Indoor Positioning System* (IPS). Sejak tahun 2019 Tim jurusan Informatika Universitas Syiah Kuala mengembangkan aplikasi berbasis IPS dengan menggunakan suatu teknologi yang memastikan bahwa pengguna berada di dalam ruangan menggunakan *Bluetooth Low Energy* (BLE) (Puspitasari, 2020). Sehingga aplikasi *route guidance* yang akan dipasangkan sistem pengenalan suara ini merupakan pengembangan lebih lanjut dari penelitian *Indoor Positioning System* menggunakan BLE. Pada proses implementasinya akan dilakukan di Gedung A Fakultas Matematika dan Ilmu Pengetahuan Alam yang nantinya akan dipasang alat transmisi BLE yang disebut *Beacon*.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, maka masalah yang akan dikaji pada penelitian ini adalah:

1. Bagaimana membangun model sistem pengenalan suara yang dapat digunakan untuk aplikasi *route guidance* berbasis *indoor positioning*?

2. Bagaimana tingkat akurasi yang dihasilkan dari pembangunan model sistem pengenalan suara?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disebutkan sebelumnya, maka dapat dipaparkan tujuan dari penelitian ini adalah sebagai berikut:

1. Membangun model sistem pengenalan suara untuk digunakan pada aplikasi *route guidance* berbasis *indoor positioning*.
2. Menganalisis tingkat akurasi yang dihasilkan dari pembangunan model sistem pengenalan suara.

1.4. Manfaat Penelitian

Setelah penelitian ini dilakukan, akan didapatkan hasil dari model *speech to text* sistem pengenalan suara yang terbaik akan dipasangkan pada aplikasi *route guidance* berbasis *indoor positioning*.

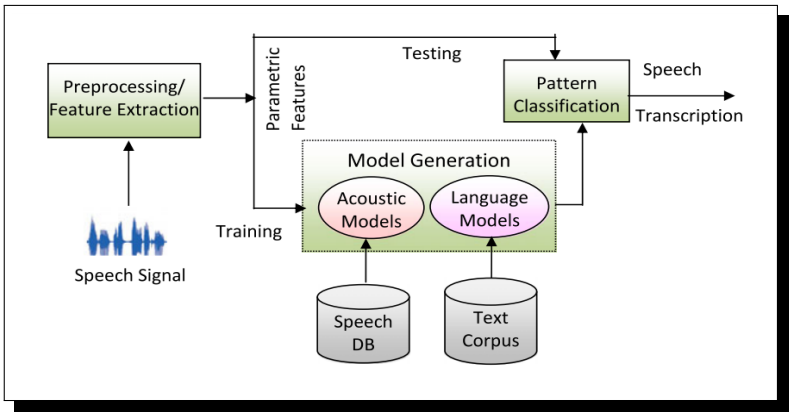
BAB II
TINJAUAN KEPUSTAKAAN

Untuk mendukung penelitian ini, maka dalam bab ini akan dikemukakan beberapa rumusan teori pendukung yang dikutip dari berbagai referensi baik dalam bentuk buku, jurnal, maupun tulisan karya ilmiah yang memiliki kaitan dengan penelitian yang dibuat.

2.1. *Speech Recognition*

Kemampuan suatu komputer agar dapat mengenali apa yang diucapkan oleh seseorang berdasarkan sinyal suara yang diucapkan oleh seseorang disebut sebagai *Automatic Speech Recognition* (ASR) (Azizah dkk., 2015). Selain untuk mengubah ucapan menjadi teks, ASR juga memiliki kemampuan untuk autentikasi *biometric*, yaitu suatu kemampuan untuk mengenali pengguna dari suaranya. Hasil dari proses pengenalan dari suara seseorang dapat digunakan untuk melakukan tugas berdasarkan instruksi (Mustikarini dkk., 2007). Menjadi suatu kemudahan bagi manusia jika komputer dapat memahami apa yang diucapkan oleh manusia dan karena hal itu juga manusia dapat dengan mudah mengoperasikan komputer karena adanya teknologi yang disebut sebagai *voice command* atau perintah suara (Widiyanto dan Endah, 2015).

Pada *speech recognition*, sinyal suara akustik dipetakan oleh komputer ke beberapa bentuk makna abstrak dari ucapan tersebut. Suara harus dicocokkan dengan potongan suara yang disimpan sebelumnya, namun pada proses ini memiliki kesulitan yang tinggi jika *sound bites* tidak cocok dengan potongan suara yang disimpan, sehingga analisis lebih lanjut harus dilakukan. Untuk mendapatkan kualitas *speech recognition* yang lebih baik, berbagai metode ekstraksi fitur dan teknik pencocokan pola digunakan karena memiliki peran penting untuk memaksimalkan tingkat pengenalan suara dari berbagai orang (Saksamudre dkk., 2015).



Gambar 2.1. Arsitektur ASR (Aggarwal dan Dave, 2012)

.

Ada beberapa langkah atau tahapan yang digunakan untuk melakukan *speech recognition* sesuai dengan gambar di atas, yaitu *pre-processing/digital processing*, ekstraksi fitur, *acoustic modelling*, *language modelling* dan *pattern classification/ decoding* (Aggarwal dan Dave, 2012). Pada pencocokan pola untuk *speech recognition* memiliki 5 pendekatan, yaitu pendekatan berbasis *template*, pendekatan berbasis pengetahuan, pendekatan berbasis jaringan saraf (*neural network*), pendekatan berbasis *Dynamic Time Warping* (DTW) dan pendekatan berbasis statistik (Saksamudre dkk., 2015).

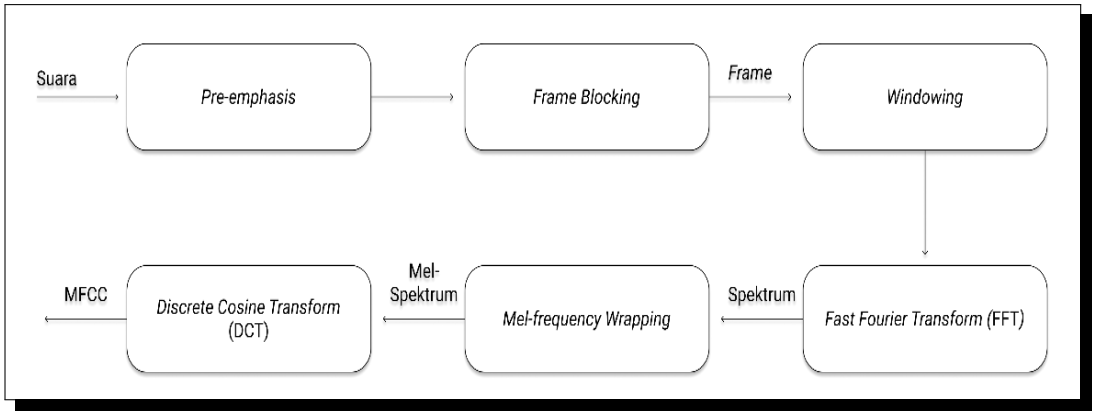
2.2. Ekstraksi Fitur

Salah satu bagian penting dalam proses pengenalan ucapan adalah ekstraksi fitur. Dengan adanya ekstraksi fitur, mesin pengenalan ucapan dapat membedakan antara satu ucapan dengan ucapan lainnya (Gupta dan Gupta, 2016). Ekstraksi fitur memiliki fungsi untuk menghapus informasi yang berlebihan dan tidak diinginkan dengan cara mendeteksi sekumpulan variabel dari sinyal suara yang dikorelasikan secara akustik, variabel tersebut disebut dengan fitur (Dua dkk., 2018). Ekstraksi fitur yang paling sering digunakan untuk pengolahan suara adalah Metode *Mel-Frequency Cepstral Coefficients* (MFCC), karena metode tersebut dapat mempresentasikan sinyal dengan baik (Umar dkk., 2019).

2.2.1. *Mel Frequency Cepstral Coefficient* (MFCC)

MFCC dapat berguna untuk mengoptimalkan sistem pengenalan suara dan menghasilkan hasil yang efisien, dengan dibangun suatu filter bank dari teknologi dan metode penelitian yang terus berubah. Dalam mengekstrak vektor fitur yang memiliki isi informasi tentang sinyal suara, MFCC menggunakan beberapa bagian

dari produksi ucapan dan persepsi ucapan (Dua dkk., 2018). Berikut adalah langkah-langkah dan diagram alir pada ekstraksi fitur dengan MFCC:



Gambar 2.2. Diagram Alir *Mel Frequency Cepstral Coefficient*

.

1. *Pre-emphasis*

   Pada tahapan awal dalam menggunakan MFCC yaitu dengan *pre-emphasis*. *Pre-emphasis* dilakukan untuk mengurangi *noise* karena sinyal suara yang didapat sering mengalami gangguan *noise* (Heriyanto dkk., 2018). Dalam mengurangi efek samping saat mekanisme produksi suara, digunakanlah *pre-emphasis* untuk menekan suara dengan frekuensi tinggi pada sinyal suara yang didapat. Fungsi lain *pre-emphasis* dapat menguatkan puncak spektrum suara berfrekuensi tinggi (Efendi, 2019). Secara matematis, *pre-emphasis* dapat dirumuskan seperti pada persamaan berikut.

$$y(n) = s(n) - as(n-1) \tag{2.1}$$

   Pada persamaan 2.1 dapat dijelaskan bahwa $y(n)$ adalah sinyal yang ditekan, $s(n)$ merupakan sinyal yang terdigitasi, serta a adalah sebuah ketetapan filter *pre-emhasis* atau sinyal yang diekstrak, dengan nilai 0.9 < a < 1.0.

2. *Frame blocking* dan *Windowing*

   Dalam menganalisis sinyal ucapan ke dalam bentuk *frame* dibutuhkan tahap *frame blocking* (Heriyanto dkk., 2018). Setelah sinyal ucapan melewati tahapan *pre-emphasis* maka sinyal tersebut dibagi menjadi beberapa *frame* dengan memuat N sampel sinyal pada masing-masing *frame* dan *frame* yang saling yang berdekatan akan dipisahkan sejauh M sampel (Laksono dkk., 2018). Setiap sampel dapat dibagi panjang *frame* menjadi beberapa *frame* berdasarkan waktu yang terletak di antara 20 ms hingga 40 ms (Laksono dkk.,

2018). Di antara bagian-bagian *frame* dengan *frame* lainnya terdapat bagian yang bertumpang tindih atau yang disebut *overlap*, hal ini berguna agar antar *frame* saling berkesinambungan (Efendi, 2019).

Untuk mencegah ketidaksinambungan sinyal suara dari ujung awal sampai ujung akhir dari proses *frame blocking*, maka dilakukanlah *windowing* (Efendi, 2019). Tujuan dari *windowing* adalah untuk mengurangi efek diskontinu pada ujung tiap *frame* (Heriyanto dkk., 2018). Ada dua fungsi yang biasa digunakan yaitu *Rectangular Window* dan *Hamming Window* (Laksono dkk., 2018). Fungsi *Rectangular Window* di tulis dalam persamaan 2.2 sebagai berikut.

$$W = \begin{cases} 1 & 0 \leq n \leq N \\ 0 & L \end{cases} \tag{2.2}$$

Fungsi di atas merupakan salah satu cara untuk menghindari diskontinu pada ujung *window*, dengan cara meruncingkan sinyal menjadi nol atau dekat dengan nol, hal ini dapat mengurangi kesalahan yang terjadi (Laksono dkk., 2018). Lalu untuk Fungsi *Hamming* sendiri dapat digambarkan seperti bentuk jendela dengan mempertimbangkan blok berikutnya di dalam rantai pemrosesan ekstraksi fitur serta menyatukan semua garis frekuensi yang terdekat (Laksono dkk., 2018). Fungsi *Hamming* di tulis dalam persamaan 2.3 sebagai berikut.

$$W(n) = \begin{cases} 0,54 - 0,46\cos(\frac{2\pi n}{N} - 1) & 0 \leq n \leq N - 1 \\ 0 & Otherwise \end{cases} \tag{2.3}$$

Pada persamaan 2.3 dapat dijelaskan bahwa $W(n)$ merupakan *Hamming Window*, jumlah dari sampel diinisialkan dengan N dan n menginisialkan pada sampel saat ini (Dua dkk., 2018). Lalu hasil persamaan *Hamming Window* dikalikan dengan sinyal masukan/ *innput* yang telah ditetapkan, perhitungan tersebut dapat dijelaskan pada persamaan 2.4 berikut ini (Laksono dkk., 2018).

$$Y(n) = y(n) \times w(n) \tag{2.4}$$

Pada persamaan 2.4 dapat dijelaskan bahwa n merupakan banyaknya sampel tiap *frame*, sinyal *output* diwakilkan dengan $Y(n)$, sinyal *input* diwakilkan dengan $y(n)$ dan $w(n)$ mewakilkan dari *Hamming Window*.

3. *Fast Fourier Transform* (FFT)

Untuk melakukan konversi sinyal dari domain waktu menjadi domain frekuensi dengan cepat digunakanlah *Fast Fourier Transform* (FFT) (Efendi, 2019). FFT berguna dalam mengubah konvolusi getaran celah suara dan respons yang didapat dari gelombang saluran suara dalam domain waktu (Laksono dkk., 2018). FFT sendiri dikembangkan karena suatu masalah dapat diselesaikan dengan mengubah atau memodelkan suatu permasalahan dalam representasi matematika dan mendapatkan keuntungan dalam segi efisiensi dibandingkan hal lainnya (Efendi, 2019).

4. *Mel-frequency Wrapping*

   Pada tahap ini spektrum yang dikeluarkan dari FFT di *wrapping* sehingga menghasilkan *mel-scale* agar resolusi frekuensi terhadap pendengaran manusia disesuaikan (Laksono dkk., 2018). Spektrum FFT memiliki nilai frekuensi yang sangat lebar, sehingga harus dipetakan ke dalam *mel-scale* agar dapat mengetahui energi yang tersedia pada setiap titik dengan bantuan *triangular* filter bank. Untuk mendapatkan batas-batas nilai tertinggi dan nilai terendah dari *mel-scale* berdasarkan frekuensi suara dalam *mel-frequency* pada setiap filter bank dihitung berdasarkan persamaan 2.5 berikut ini (Efendi, 2019).

$$mel = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{2.5}$$

   Pada persamaan 2.5 dapat dijelaskan bahwa mel merupakan skala *mel-frequency* dan frekuensi linear diwakilkan dengan f.

5. *Discrete Cosine Transform* (DCT)

   Tahap *Discrete Cosine Transform* (DCT) digunakan untuk mengubah nilai koefisien *mel-spectrum* menjadi ke domain waktu (Dua dkk., 2018).

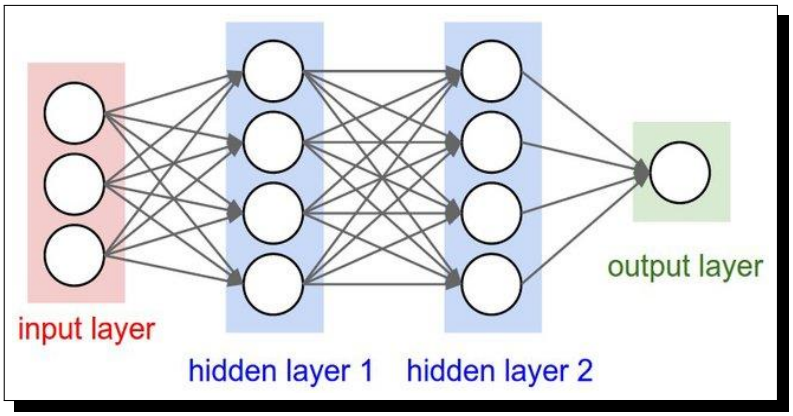
2.3. *Acoustic Model*

Terdapat dua komponen utama pada sistem *voice-to-text*, yang pertama adalah model akustik (*acoustic model*) dan model bahasa (*language model*). Model yang mengandung representasi statistik dari setiap suara yang berbeda dalam bentuk kata ataupun kalimat merupakan model akustik. Pada tiap-tiap representasi statistik tersebut dapat menentukan sebuah label yang disebut dengan suku kata (*phonemes*) (Misbullah dkk., 2020). Model akustik memiliki dua tipe yang berbeda, yaitu *phonemes* dan kata. Dua tipe tersebut diimplementasikan dengan berbagai metode seperti *Hidden Markov Model* (HMM), *Support Vector Machines* (SVM), *Dynamic Bayesian Networks* (DBN) dan *Artificial Neural Network* (ANN) (Gupta dan Gupta,

2016). Sebagian besar sistem pengenalan suara saat ini menggunakan HMM untuk menangani variabilitas temporal ucapan lalu menggunakan *Gaussian Mixture Models* (GMM) untuk menentukan seberapa baik status yang dihasilkan dari tiap HMM tersebut cocok dengan *frame* atau *short window of frames of coefficient* yang mewakili *input* dari akustik (Hinton dkk., 2012). Cara alternatif yang terbaik untuk mengevaluasi kecocokan HMM dengan menggunakan *Deep Neural Network* (DNN).

2.3.1. *Deep Neural Network* (DNN)

DNN memiliki cara kerja dengan mengambil beberapa koefisien dari frame sebagai masukannya dan menghasilkan keluaran berupa probabilitas posterior dari status HMM (Hinton dkk., 2012). *Deep Neural Network* merupakan salah satu cabang dari *machine learning* yang menggunakan konsep jaringan syaraf tiruan atau yang disebut dengan *Neural Network* (Laksono dkk., 2018). DNN merupakan *feed-forward*, *feed-forward* itu sendiri adalah salah satu bagian di *Artificial Neural Network* (ANN) yang memiliki lebih dari satu lapisan di *hidden units* antara masukan atau keluarannya (Hinton dkk., 2012). Pada umumnya terdapat tiga struktur dari *neural network*, yaitu *input layer*, *hidden layer* dan *output layer*. Pada *input layer* tiap *node* merepresentasikan vektor untuk jumlah data *input*, lalu pada *hidden layer* memiliki fungsi untuk mengontrol apakah informasi dapat diketahui atau tidak dari *input layer* yang akan diteruskan ke layer berikutnya dan yang terakhir setiap *node* pada *output layer* didefinisikan sebagai target dari kelas yang akan diprediksi (Misbullah dkk., 2020). DNN terbukti dapat mengungguli GMM dalam berbagai tolak ukur pengenalan ucapan serta terkadang dengan margin yang besar, karena DNN sendiri memiliki banyak *hidden layer* dan dibuktikan dengan dilatih menggunakan metode baru (Hinton dkk., 2012). Penggambaran model umum dari DNN ditunjukkan seperti pada Gambar 2.3, di mana variasi proses, jumlah dan urutan *hidden layer* tergantung pada arsitekturnya (Musiol, 2016).



Gambar 2.3. Model Umum dari *Deep Neural Network* (Musiol, 2016)

.

Neural Network sendiri memiliki dua arsitektur, yaitu *single layer perseptron* dan *multi layer perseptron*. *Single layer persepton* mempunyai satu *hidden layer* yang digunakan, lalu pada *multi layer perseptron* mempunyai lebih dari dua *hidden layer* yang digunakan (Laksono dkk., 2018). Untuk mendapatkan hasil yang akurat, pada *Deep Neural Network* atau *Deep Learning* menggunakan jumlah *hidden layer* yang sangat banyak (Laksono dkk., 2018).

2.4. Kaldi *Toolkit*

Salah satu *toolkit* untuk pengenalan suara yang *open-source* adalah Kaldi, Kaldi sendiri ditulis dalam bahasa C++ dan di bawah lisensi Apache v2.0 (Povey dkk., 2011). Kaldi bergantung dengan dua *library* dari luar yang tersedia secara bebas, *library* yang pertama adalah OpenFst digunakan untuk *finite-state framework* lalu untuk *library* aljabar linear ekstensif menggunakan "*Basic Linear Algebra Subrutin*" (BLAS) dan "*Linear Algebra PACKage*" (LAPACK) (Povey dkk., 2011). Fitur pada Kaldi yang dapat digunakan adalah ekstraksi fitur yang paling umum digunakan, pemodelan akustik dengan beberapa model umum namun dapat diperluas dengan model jenis baru, *phonetic decision tree* yang efisien untuk ukuran konteks arbitrer dan juga mendukung dengan berbagai pendekatan, pemodelan bahasa dapat menggunakan model bahasa apa pun yang dapat direpresentasikan sebagai *finite-state transducer* (FST) serta dapat menggunakan *toolkit* IRSTLM untuk membangun pemodelan Bahasa dari teks mentah, dan dapat membuat grafik decoding yang didasarkan pada *Weighted Finite State Transducers* (WFSTs) (Povey dkk., 2011). Kaldi terus dikembangkan dan sedang mengerjakan model Bahasa yang besar dalam kerangka FST, pembuatan kisi dan pelatihan diskriminatif (Povey dkk., 2011).

2.5. *Indoor Positioning System*

Indoor Positioning System (IPS) dapat menemukan lokasi orang atau objek di dalam ruangan dengan menggunakan gelombang radio, medan magnet, sinyal akustik serta informasi sensoris yang dikumpulkan oleh alat teknologi berupa *smartphone*, tablet, atau perangkat pintar lainnya (IndoorAtlas, 2020). IPS memiliki beberapa pendekatan yaitu, *positioning systems* berbasis WiFi (WPS), *positioning systems* berbasis *Bluetooth Low Energy* (BLE), sistem berbasis Identifikasi Frekuensi Radio (RFID) dan teknologi *Ultra-Wide Band* (UWB) atau *Visible Light Communication* (VLC) (Cantón Paterna dkk., 2017).

Bluetooth Low Energy (BLE) *Beacon* merupakan suatu perangkat yang berukuran kecil dan menggunakan baterai sebagai sumber tenaganya, BLE dapat mengirimkan sinyal di area yang terbatas sampai 20 meter / 70 kaki dan bereaksi terhadap lokasi seseorang yang berada dalam jangkauan (IndoorAtlas, 2020). BLE memiliki kelebihan yang dapat mengalahkan WPS yang telah popular dimasa lalu. BLE menawarkan harga yang lebih murah dan memiliki daya yang rendah, sehingga untuk infrastruktur yang tidak memungkinkan adanya WiFi dapat menjadi pilihan terbaik (Cantón Paterna dkk., 2017). BLE dikenal sebagai pilihan yang terbaik karena mampu menyiarkan data menggunakan daya yang minimal hal itu didasarkan karena teknologi *Bluetooth* yang digunakannya, serta BLE ideal untuk perangkat yang dapat berfungsi tanpa gangguan untuk jangka waktu yang lama karena menggunakan baterai kecil. Dari hal yang sudah dijelaskan sebelumnya BLE dapat berfungsi dengan baik untuk keperluan navigasi di dalam ruangan (Herrera Vargas, 2016).

BAB III
METODOLOGI PENELITIAN

3.1. Waktu dan Lokasi Penelitian

Penelitian ini dilakukan di Fakultas Matematika dan Ilmu Pengetahuan Alam Gedung A Universitas Syiah Kuala. Waktu yang dibutuhkan untuk penelitian ini adalah 4 bulan terhitung dari bulan Maret 2021 hingga Juli 2021.

3.2. Alat dan Bahan

Alat dan bahan yang digunakan pada penelitian ini meliputi perangkat keras, perangkat lunak, dan data. Perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut:

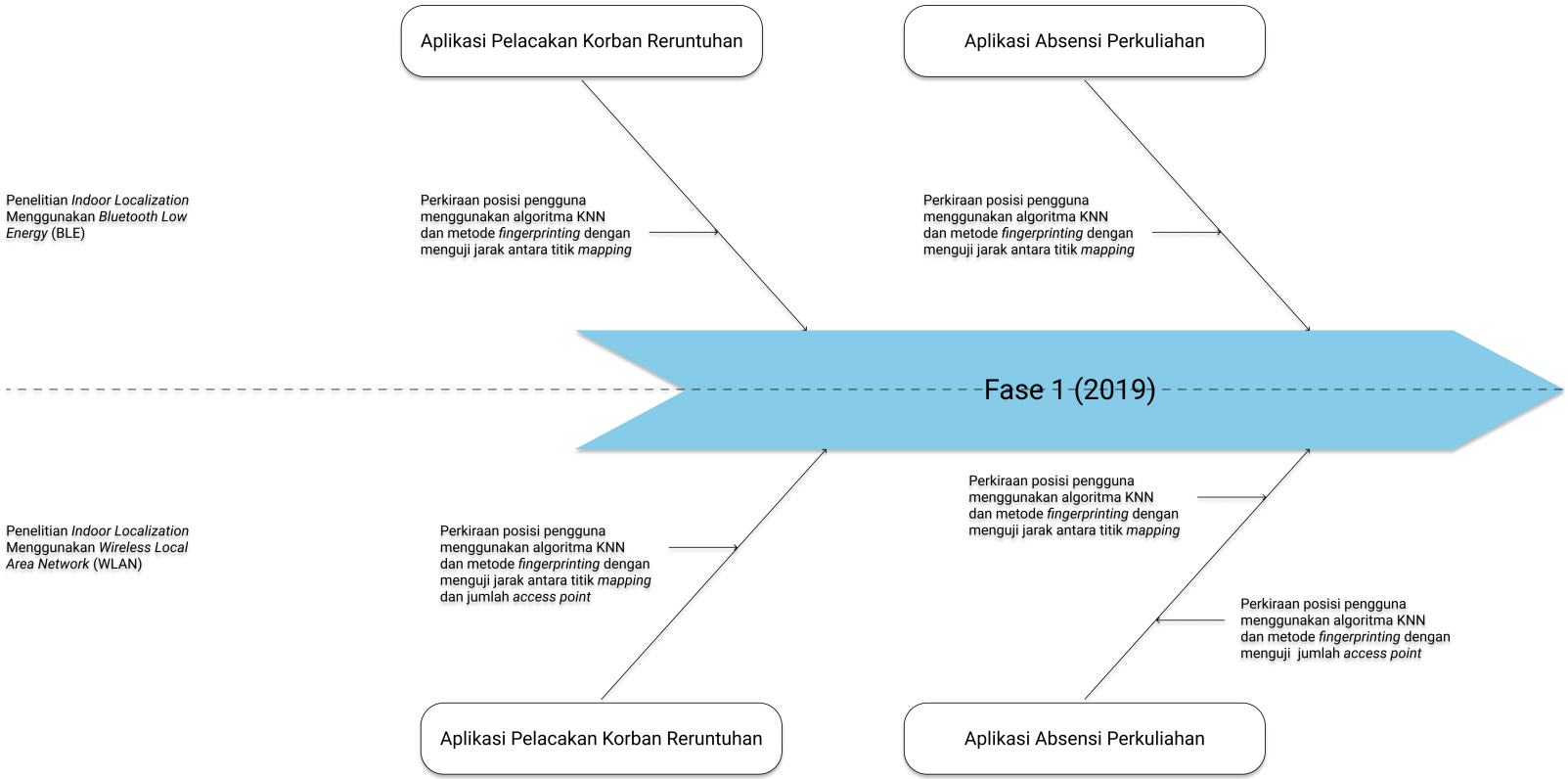
- Windows 10 64 bit
- Linux Ubuntu 20.04.1
- Kaldi
- Vosk-api
- Python
- Visual Studio Code 1.50.1
- Google Form
- Freemake Audio Converter 1.1.9
- SRILM 1.7.3

Sedangkan perangkat keras yang digunakan pada penelitian ini adalah 1 unit Laptop Acer Aspire A514-52G dengan RAM 12 GB, *Processor* Intel® Core™ i5-10210U @ 1.60 GHz (8 CPUs), 2.1GHz, Kartu grafis NVIDIA GeForce MX250 dan Harddisk 1 TB.

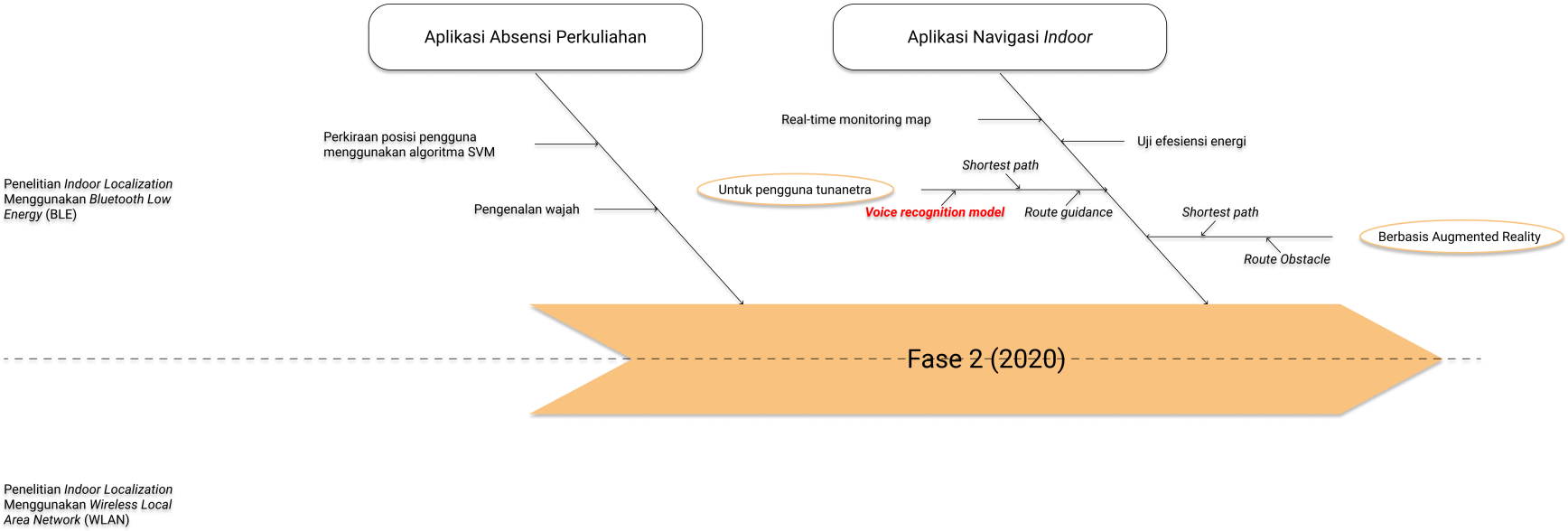
3.3. *Roadmap* Penelitian

Roadmap pada penelitian ini merupakan diagram yang menggambarkan rangkaian beberapa penelitian yang memiliki kesinambungan dalam rentang waktu 2019 sampai dengan 2020 yang dibagi menjadi 2 fase. Fase pertama pada tahun 2019 memiliki dua fokus penelitian *indoor localization* dengan menggunakan *Bluetooth Low Energy* (BLE) dan menggunakan *Wireless Local Area Network*

(WLAN) dapat dilihat pada gambar 3.1. Lalu pada fase kedua pada tahun 2020 lebih berfokus pada penelitian *indoor localization* dengan menggunakan BLE dapat dilihat pada gambar 3.2. Aplikasi Navigasi *Indoor* dibangun untuk Fakultas MIPA, karena memiliki gedung yang cukup luas dan memiliki banyak sub Gedung yang dibagi menjadi 6 bagian dari Gedung A sampai Gedung F. Aplikasi ini dapat berjalan dan memberikan arahan pada pengguna di dalam ruangan. Salah satu bagian dari aplikasi Navigasi *Indoor* ini adalah aplikasi *route guidance* untuk pengguna tuna netra. Aplikasi *route guidance* ini menggunakan *voice recognition* sebagai penghubung untuk mempermudah bagi pengguna tuna netra agar dapat menjalankan aplikasi ini. Penelitian ini terletak pada fase 2 di tahun 2020 dengan topik utama yaitu Aplikasi Navigasi *Indoor* dengan sub topik untuk pengguna tunanetra. Penelitian ini memiliki batasan berupa pembangunan model *voice recognition* seperti yang ditunjukkan pada gambar 3.2 dengan kata dicetak tebal berwarna merah sebagai berikut.



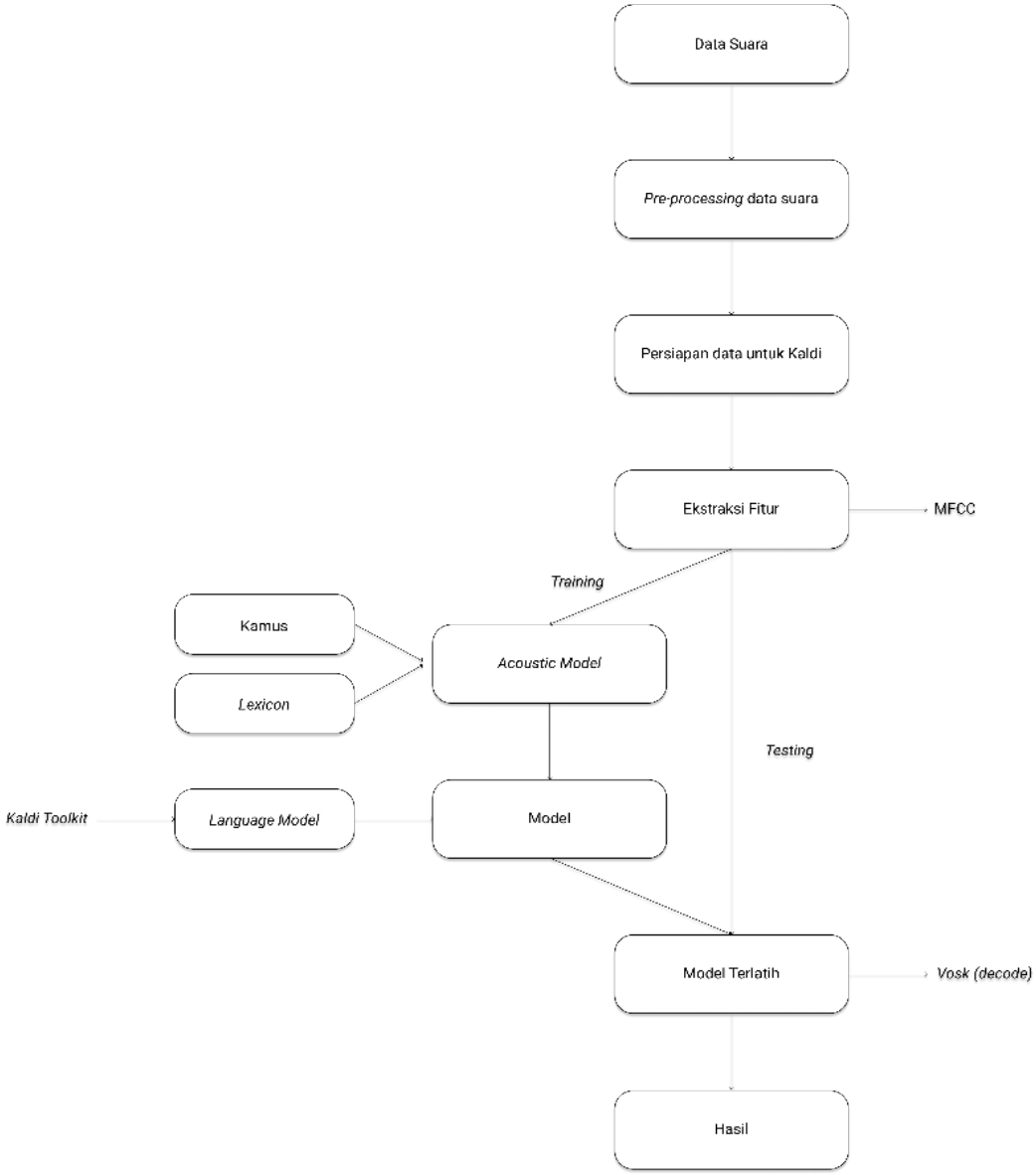
Gambar 3.1. *Roadmap* Penelitian Fase 1



Gambar 3.2. *Roadmap* Penelitian Fase 2

3.4. Metode Penelitian

Metode penelitian yang dilakukan dalam penelitian ini ditunjukkan pada Gambar 3.3.



Gambar 3.3. Diagram Perancangan Sistem Pengenalan Ucapan

3.4.1. Data Suara

Pada tahap ini memiliki 2 hal yang dilakukan, yang pertama pembuatan transkrip dari nama-nama ruangan pada Gedung A FMIPA Unsyiah dan pengambilan data suara dari narasumber. Pada penulisan transkrip dilakukan dengan menggali informasi nama-nama ruangan pada Gedung A FMIPA dari beberapa orang. Setelah itu dilakukan proses pengambilan data suara oleh beberapa narasumber yang berjumlah 20 orang, memiliki rentang umur 17 tahun sampai

dengan 58 tahun, memiliki 10 orang yang memiliki jenis kelamin laki-laki dan 10 orang memiliki jenis kelamin perempuan. Pengambilan suara dilakukan dengan 2 cara. Pertama melalui proses perekaman suara masing-masing melalui *smartphone* masing-masing lalu di *input* melalui *form* yang sudah dibuat dengan Google Form. Kedua dilakukan perekaman suara dengan narasumber secara langsung tanpa melalui Google Form. Sebelum proses perekaman suara, narasumber diberikan 56 daftar nama-nama ruangan pada gedung A FMIPA Unsyiah serta diberi pengarahan dalam pengucapan dari daftar yang telah diberikan.

3.4.2. *Pre-processing* Data Suara

Pada tahap ini dilakukan *pre-processing* terhadap data suara yang sudah dimasukkan oleh narasumber. Tahap ini dilakukan karena narasumber melakukan proses perekaman suara secara mandiri dengan *smartphone* masing-masing, jadi memiliki tipe *file* audio yang berbeda-beda. *File* audio diubah menjadi tipe *file* *.wav dengan *mono channel*, 16KHz *sample rate* dan *sample size* 16 bit. Aturan tersebut sesuai dengan yang diterima oleh Kaldi dan Vosk-api. Proses mengubah tipe file dengan menggunakan aplikasi Freemake Audio Converter, aplikasi ini dapat diinstalasi secara gratis dan dapat digunakan untuk sistem operasi Windows. Selain mengubah tipe *file*, pada tahapan ini juga dilakukan pencocokan data berdasarkan nama-nama ruangan yang dikumpulkan menjadi satu *folder*. Sehingga pada penelitian ini memiliki 20 *folder* berbeda yang berisikan 56 *file*. 56 *file* tersebut merupakan jumlah narasumber atau yang dapat kita sebut di sini adalah *speaker*.

3.4.3. Persiapan Data untuk Kaldi

Pada tahapan ini ada beberapa file yang harus disiapkan untuk membuat model *speech recognition*, yaitu ada persiapan data akustik untuk membuat *acoustic model* dan *language data* untuk membuat *language model*. Sebelum mempersiapkan beberapa *file* yang dibutuhkan, data suara yang telah dikumpulkan di bagi menjadi 2 bagian, di mana 80% *speaker* untuk *data training* dan 20% speaker untuk *data testing*. Berikut adalah penjelasan dari persiapan data akustik dan *language data*:

1. Persiapan data akustik

   Pada tahapan ini, ada 5 file yang harus dibuat agar Kaldi dapat memahami data audio yang akan diproses. Berikut adalah penjelasan file yang harus dibuat:

   - Spk2gender

     Pada file ini berisikan informasi tentang nama yang diasumsikan sebagai *id* dari *speaker* dan jenis kelamin *speaker* tersebut. *File* ini memiliki

pattern <*speakerID*> <jenis kelamin>, di mana jenis kelamin diinisialkan f sebagai perempuan dan m sebagai pria.

- Wav.scp

  Pada *file* ini berisikan informasi *utteranceID* dan lokasi audio tersebut ditambahkan dengan nama file tersebut, dalam hal ini penulisan *utteranceID* merupakan *speakerID* yang disambung dengan nama-nama ruangan yang diucapkan pada audio tersebut. Sehingga file ini memiliki *pattern* <*uterranceID*> <lokasi_file_audio>.

- *Text*

  Pada *file* ini berisikan informasi tentang *utteranceID* dan transkrip dari nama–nama ruangan di Gedung A FMIPA Unsyiah yang diucapkan oleh *speaker*. Sehingga file ini memiliki *pattern* <*uterranceID*> <*text_transcription*>.

- Utt2spk

  Pada *file* ini berisikan informasi berupa *utteranceID* dan *speakerID*, agar sistem pengenalan suara dapat mengetahui nama-nama ruangan yang diucapkan oleh *speaker* tertentu. Sehingga file ini memiliki *pattern* <*uterranceID*> <*speakerID*>.

- *Corpus*

  Pada *file* ini berisikan informasi berupa transkrip dari nama-nama ruangan di Gedung A FMIPA Unsyiah yang diucapkan oleh *speaker*, namun *file* ini disimpan pada lokasi yang berbeda dari beberapa *file* yang sudah disebutkan sebelumnya. *File* ini memiliki *pattern* <*text_transcription*> yang ditulis per baris, sehingga *file* pada penelitian ini terdapat 56 baris.

2. Persiapan *language data*

   Pada tahapan ini, ada 4 *file* berkaitan dengan *language model*. Berikut adalah penjelasan *file* yang harus dibuat:

   - *Lexicon*

     Pertama membuat *file* yang bernama *lexicon*.txt. Pada *file* ini berisikan setiap kata dari *dictionary* dengan penambahan *phone transcriptions* atau penyebutan dari kata tersebut. Seperti contoh, jika ada kata ‘satu‘ maka penyebutan dari kata tersebut adalah sa tu. Namun, jika ada suara yang diam akan dideteksi sebagai *silence phone* dan penyebutannya akan ditandai dengan ‘sil‘. Sehingga file ini memiliki *pattern* <*word*> <*phone* 1> <*phone* 2>.

- *Nonsilence phones*

  Lalu pada *file* kedua diberi nama *nonsilence_phones*.txt. Pada *file* ini berisikan *list phone transcriptions* yang *nonsilence phones*. Yang termasuk pada *file* ini adalah semua *phones* pada *file lexicon* kecuali *phones* seperti ‘sil‘ atau diam. *File* ini memiliki *pattern*: <*phone*>.

- *Silence phones*

  Pada *file* ketiga diberi nama *silence_phones*.txt. Pada *file* ini berisikan *list phone transcription* yang diam atau ‘sil‘ pada bagian *phone*. *File* ini memiliki *attern*: <*phone*>.

- *Optional silence*

  Pada *file* keempat diberi nama *optional_silence*.txt. Pada *file* ini berisikan *list* pilihan lainnya dari *silence phone transcription*. *File* ini memiliki *pattern*: <*phone*>.

3.4.4. Ekstraksi Fitur

Setelah tahapan sebelumnya selesai dilakukan, selanjutnya masuk dalam tahapan ekstraksi fitur. Ekstraksi fitur dilakukan dengan menggunakan metode *Mel Frequency Ceptral Coefficient* (MFCC), agar dapat mengidentifikasi konten linguistik dan membuang suara latar belakang serta suara yang tidak dibutuhkan dalam proses ini. Ekstraksi fitur pada penelitian ini memanfaatkan *open source toolkit* untuk pengolahan suara yang tersedia yaitu Kaldi Toolkit. Pada perhitungan MFCC untuk penelitian ini ada hal yang harus ditentukan, yaitu menentukan jumlah *ceptral coeffient* sebanyak 13 *ceptral coeffient* dan jumlah *frame* dalam suatu *file* sepanjang 25 *millisecond* dan memiliki *window step* sebesar 10 *millisecond*.

3.4.5. Akustik Model

Pada bagian ini merupakan tahapan untuk melakukan model akustik pada data training yang telah melewati proses ekstraksi fitur dengan MFCC. Tahapan ini menggunakan proses *machine learning* dengan metode *Deep Neural Network* (DNN) dan dilakukan juga pengambilan nilai *CTC Loss* yang merepresentasikan akurasi dari *training*. Pada tahap awal *training* dimulai dengan menentukan jumlah *file* yang masuk ke dalam perhitungan ke *deep learning* sebagai bahan dari pembelajaran atau pembagi data. Setelah itu nilai dari MFCC *feature* dimasukkan, lalu label untuk proses pembelajaran dimasukkan juga. Berikutnya dilakukan inisialisasi *learning rate*, maksimal *epoch* dan minimum *loss* yang digunakan. Lalu pada proses perhitungan *neural network* diinisialkan bobot yang dipakai. Selanjutnya metode *deep learning* dengan *input* nilai MFCC *feature* dan target

dihitung. Terakhir menampilkan nilai *loss* dari hasil perhitungan yang didapat. Hal tersebut dilakukan sampai mendapatkan minimum nilai *loss* dan maksimum nilai *epoch*. Pada penelitian ini digunakan *multilayer perseptron* sebagai layer pembelajarannya.

3.4.6. Model

Setelah mendapatkan hasil dari *acoustic modelling* dengan Kaldi, selanjutnya melakukan *language modelling* dengan menggunakan SRILM (*SRI Language Modeling Toolkit*). *Language modelling* berguna agar dapat membedakan kata dan frasa yang terdengar serupa pada ucapannya sehingga didapatkan model pengenalan ucapan yang terlatih.

3.4.7. Model Terlatih

Setelah model pengenalan ucapan sudah selesai dibangun, lalu model tersebut diletakan pada Vosk yang merupakan *open source speech recognition toolkit*, Vosk dapat berjalan pada Android dengan API secara *offline*. Data *testing* yang sebelumnya sudah dipisahkan, diuji pada model tersebut sehingga mendapatkan keluaran berupa *text*. Model ini akan dapat berjalan pada aplikasi Android *route guidance* untuk tuna netra berbasis *Indoor Positioning* yang akan dibangun.

DAFTAR KEPUSTAKAAN

Aggarwal, R. K. dan Dave, M. (2012). Integration of multiple acoustic and language models for improved hindi speech recognition system. *International Journal of Speech Technology*, 15(2):165–180.

Azizah, R. S., Nurjanah, D., dan Sari, F. D. (2015). Sistem automatic speech recognition menggunakan metode mfcc dan hmms untuk deteksi kesalahan pengucapan kata bahasa inggris. *eProceedings of Engineering*, 2(3).

Cantón Paterna, V., Calveras Auge, A., Paradells Aspas, J., dan Perez Bullones, M. A. (2017). A bluetooth low energy indoor positioning system with channel diversity, weighted trilateration and kalman filtering. *Sensors*, 17(12):2927.

Dua, M., Aggarwal, R. K., dan Biswas, M. (2018). Performance evaluation of hindi speech recognition system using optimized filterbanks. *Engineering Science and Technology, an International Journal*, 21(3):389–398.

Efendi, R. (2019). Automatic speech recognition bahasa indonesia menggunakan bidirectional long short-term memory dan connectionist temporal classification.

Gupta, H. dan Gupta, D. (2016). Lpc and lpcc method of feature extraction in speech recognition system. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, pages 498–502. IEEE.

Heriyanto, H., Hartati, S., dan Putra, A. E. (2018). Ekstraksii ciri mel frequency ceptral coefficient (mfcc) dan rerata coefficient untuk pengecekan bacaan al-qur'an. *Telematika: Jurnal Informatika dan Teknologi Informasi*, 15(2):99–108.

Herrera Vargas, M. (2016). Indoor navigation using bluetooth low energy (ble) beacons.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., dkk. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

IndoorAtlas (accessed November 29, 2020). A 2016 global research report on the indoor positioning market. *http://www.indooratlas.com/wp-content/uploads/2016/09/A-2016-Global-Research-Report-On-The-Indoor-Positioning-Market.pdf*.

Laksono, T. P. dkk. (2018). Speech to text untuk bahasa indonesia.

Misbullah, A., Nazaruddin, N., Marzuki, M., dan Zulfan, Z. (2020). Penerapan time delay neural network pada model akustik untuk sistem voice-to-text berbahasa sunda. *Journal of Data Analysis*, 2(2):61–70.

Musiol, M. (2016). Speeding up deep learning.

Mustikarini, W., Hidayat, R., dan Bejo, A. (2007). Real-time indonesian language speech recognition with mfcc algorithms and python-based svm. *IJITEE (International Journal of Information Technology and Electrical Engineering)*, 3(2):55–60.

Ouisaadane, A., Safi, S., dan Frikel, M. (2020). Arabic digits speech recognition and speaker identification in noisy environment using a hybrid model of vq and gmm. *TELKOMNIKA*, 18(4):2193–2204.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., dkk. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Puspitasari, R. (2020). Rancang Bangun Aplikasi Kehadiran Perkuliahan Berbasis Teknologi Indoor Positioning System Menggunakan Bluetooth Low Energy dan Metode Klasifikasi K-NN. *ETD Unsyiah*.

Saksamudre, S. K., Shrishrimal, P., dan Deshmukh, R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications*, 115(22).

Umar, R., Riadi, I., dan Hanif, A. (2019). Analisis bentuk pola suara menggunakan ekstraksi ciri mel-frequencey cepstral coefficients (mfcc). *CogITo Smart Journal*, 4(2):294–304.

WHO (accessed Oktober 23, 2020). World report on vision. *https://www.who.int/publications-detail-redirect/world-report-on-vision*.

Widiyanto, E. dan Endah, S. N. (2015). *Aplikasi Speech to Text Berbahasa Indonesia Menggunakan Mel-Frequency Cepstral Coefficient Dan Hidden Markov Model*. PhD thesis, Universitas Diponegoro.