BAB 3
METODOLOGI PENELITIAN

3.1 Studi Literatur

Studi literatur dilakukan dengan mencari karya-karya tulis yang berkaitan dengan augmentasi audio, model yang dapat digunakan pada sistem ASR, dan riset terkait untuk sistem deteksi suara.

3.2 Pengumpulan Data

Penelitian ini menggunakan *dataset* yang berasal dari OpenSLR. *Dataset* ini berfokus kepada bahasa Jawa dan Sunda [49].

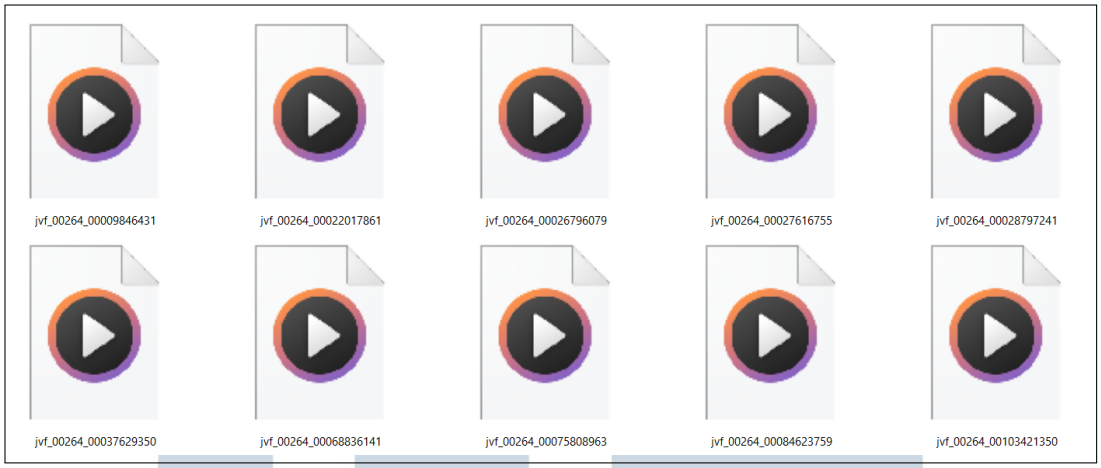
A Dataset Bahasa Jawa dan Sunda

Pada *dataset* yang dikhususkan untuk bahasa Jawa dan Sunda, walaupun kedua *dataset* disimpan secara terpisah, setiap *dataset* memiliki struktur yang mirip, dimana terdapat sebuah *folder* bernama 'wavs' untuk *file-file* audio dan sebuah *file* 'line_index.tsv' yang memuat transkripsi atau metadata. Lebih lanjut, *dataset* untuk bahasa Jawa dan Sunda masing-masing dibagi lagi berdasarkan jenis kelamin pembicara, yaitu versi laki-laki dan versi perempuan. Namun, perlu diperhatikan bahwa *file* 'line_index.tsv' tidak menyertakan baris judul (*header*) seperti yang dapat dilihat pada Gambar 3.1.

```
jvm_00027_00020497515   2a67beac-d68b-4ff3-882c-2bd7b73f181c   arep njupuk pensiunan saka p_letter t_letter taspen kanggo mbah uyut
jvm_00027_00036040997   cb3f999b-9734-4f2c-9c54-ab5c287dc77f   rokok sing mbok tuku wingi kae produksine bentoel group apa duduk
jvm_00027_00054599939   ff1cfd4e-b95c-4d63-9fcb-1142b2ca24a9   paklik bidal dhateng dubai nitih emirate airlines
jvm_00027_00058492106   73c29786-3f45-406a-ae95-f5886cf2ebe0   titi kamal nyilakani yugane piyambak ing dalan
jvm_00027_00096169948   bf7f4578-ea58-4eca-b265-334093a405bb   nekmu kepingin lunga ndelok petronas wis koyok wong arep nglairna
jvm_00027_00099501928   f376c5bf-85e8-4c71-9bf2-f464b164ba9c   ashley tidale iku kongkonen mingkem ta aja ngowoh ae
```

Gambar 3.1. Gambaran data bahasa jawa laki-laki

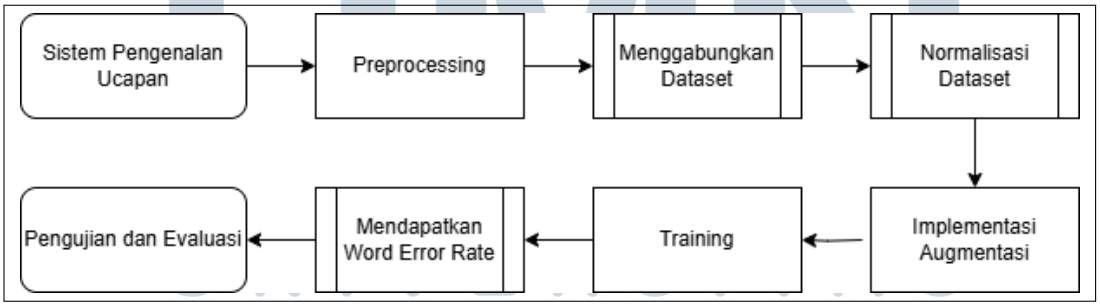
Selain file line_index, *dataset* ini juga memiliki data audio yang berformat WAV seperti apda gambar 3.2.



Gambar 3.2. Gambaran data audio

3.3 Pembangunan Model

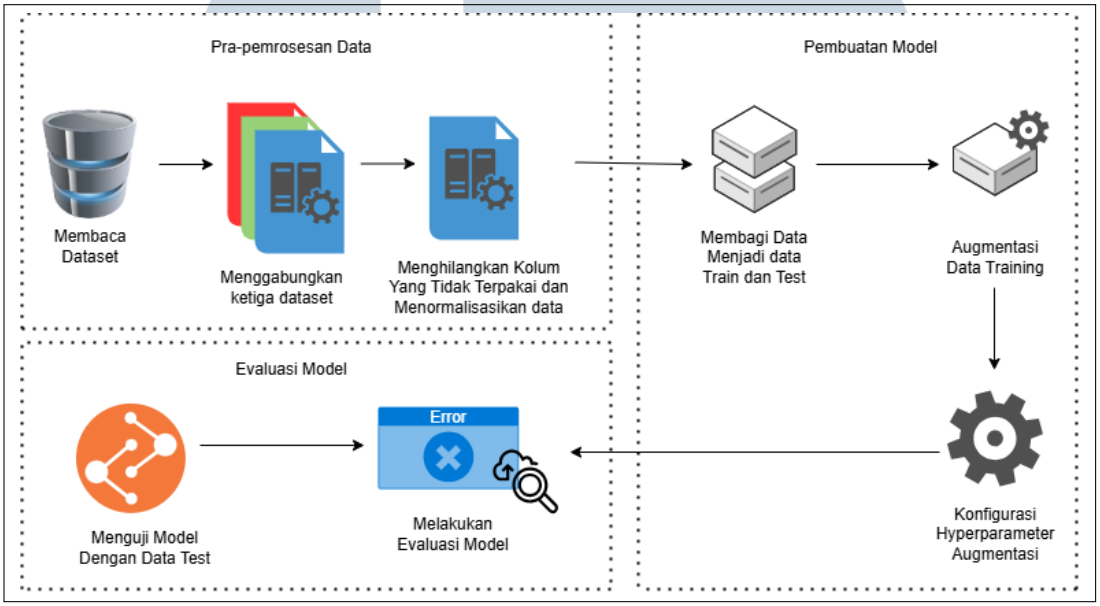
Penelitian ini menggunakan pendekatan eksperimental untuk menguji efektivitas alat augmentasi audiomentations pada sistem deteksi suara untuk bahasa Jawa dan Sunda. Eksperimen dirancang untuk mengukur peningkatan kinerja model ASR setelah penerapan teknik augmentasi audio tersebut, meliputi analisis komparatif sebelum dan setelah penggunaan teknik augmentasi untuk menilai perbaikan yang spesifik dalam aspek-aspek seperti akurasi, kecepatan respons, dan kemampuan mengenali dialek atau aksen yang beragam. Pada Gambar 3.3 dapat dilihat struktur kerja pembuatan sistem pengenalan ucapan.



Gambar 3.3. Pembuatan sistem pengenalan ucapan

Gambaran mengenai aplikasi model pengenalan ucapan seperti yang digambarkan pada Gambar 3.4 dimulai dari pra-pemrosesan data. Pada tahap pra-pemrosesan data akan dibaca dan persiapkan untuk masuk ke tahap pembuatan model. Tahap ini lebih terfokus pada penggabungan dan pembersihan data bahasa Jawa dan Sunda. Lalu pada tahap pembuatan model, data yang telah kita gabungkan

akan kita pisah kembali menjadi dua yaitu data *train* dan data *test*. Setelah pemisahan ini, data *train* akan diaugmentasi agar dapat beradaptasi lebih optimal pada kondisi audio apapun. Menggunakan data audio yang telah diaugmentasi, data akan digunakan untuk melakukan proses *training* pada model wav2vec2. Dengan hasil akhir yang akan diuji menggunakan data *test* yang telah disiapkan dan dievaluasi kinerjanya berdasarkan WER yang didapatkan.



Gambar 3.4. Aplikasi model pengenalan ucapan

3.3.1 Pra-pemrosesan

Pra-pemrosesan atau kerap dipanggil *preprocessing* merupakan kumpulan teknik yang diterapkan pada data sebelum data tersebut digunakan dalam proses pembelajaran mesin. Tujuan utama dari proses ini adalah untuk mempersiapkan data agar lebih mudah dan efektif dalam penggunaan model atau algoritma analisis. Dalam proses *preprocessing*, secara umumnya terdapat tahap-tahap yang dilakukan sebagai berikut:

1. Pembersihan Data (*Data Cleaning*) : Proses ini merupakan proses identifikasi dan perbaikan data. Seperti mengatasi nilai yang hilang, mengoreksi input yang tidak benar, dan juga mengidentifikasi dan menghapus data yang duplikat.

2. Transformasi Data (*Data Transformation*) : Proses ini mencakup pengubahan data ke dalam format yang lebih sesuai untuk analisis. Tahap ini mencakup

normalisasi audio, pemotongan kebisingan (*noise reduction*), pengaturan amplitudo, ekstraksi fitur. Selanjutnya, tokenisasi dan normalisasi data pada teks, hal ini mencakup pengubahan data dan dihapusnya simbol-simbol seperti tanda seru (!), koma (,), titik (.), dan sebagainya seperti yang dapat dilihat pada Tabel 4.1.

Tabel 3.1. Perbandingan data sebelum dan sesudah normalisasi

No.	Data Sebelum	Data Sesudah
1	Saya mendengarkan cerita membosankan dari tema	saya mendengarkan cerita membosankan dari tema
2	halo dunia!	halo dunia
3	Sudah makan? sudah sholat...?	sudah makan sudah sholat
4	mau pergi kemana hari ini?	mau pergi kemana hari ini
5	udah keluar hasil testnya?	udah keluar hasil testnya

3. Pengurangan Data (*Data Reduction*) : Proses ini mencakup pada pengurangan dimensi, agregasi data (menggabungkan data), dan sampling data.

4. Pemisahan Data : Proses ini mencakup pada pemisahan data menjadi set pelatihan (*training*), validasi (*validation*), dan pengujian (*testing*).

3.3.2 Augmentasi Suara

Menggunakan teknik augmentasi dapat meningkatkan ketahanan dan efektivitas model pengenalan suara Anda. Beberapa teknik augmentasi yang digunakan meliputi:

1. Penambahan Kebisingan Gaussian
2. Perpanjangan Waktu (*Time Stretch*)
3. Perubahan Pitch (*Pitch Shift*)
4. Pergeseran (*Shift*)

Variabel independen yang akan dimanipulasi dalam penelitian ini adalah perubahan *pitch* (*Pitch Shift*) dan pergeseran (*Shift)*). Lalu, penambahan kebisingan gaussian dan perpanjangan waktu akan menjadi variabel dependen yang tidak akan dimanipulasi pada saat skenario augmentasi diterapkan.

3.3.3 Pengaturan Model XLSR Wav2Vec2

Pada umumnya, Model XLSR Wav2Vec2 memiliki banyak sekali varian seperti contohnya XLSR-53 yang memuat 53 bahasa, dan XLSR-96 yang memuat 96 bahasa. Pemilihan model ini dapat bergantung dari kestabilan model dimana XLSR-53 sudah cukup teruji dalam aplikasi dan telah dipublikasikan lebih lama dibandingkan XLSR-96.

3.3.4 Pelatihan Model

Komponen pertama dari Wav2Vec2 terdiri dari rangkaian lapisan CNN yang digunakan untuk mengekstraksi fitur yang bermakna secara akustik-namun independen secara kontekstual-dari sinyal ucapan mentah. Bagian dari model ini sudah cukup terlatih selama pra-pelatihan dan tidak perlu lagi di lakukan proses *fine-tuning* [55]. Namun, hal ini tidak menutup kemungkinan untuk proses *fine-tuning* setelah *training*.

3.3.5 Pengujian dan Evaluasi

Pada evaluasi sistem ASR terdapat dua metrik evaluasi yang sering digunakan yaitu *Character Error Rate* (CER) dan *Word Error Rate* (WER). Namun karena bahasa Jawa dan Sunda tidak terikat dengan karakter-karakter khusus seperti implementasi nada pada bahasa Cina. Maka, model ini hanya akan diuji dengan menggunakan metrik evaluasi WER. Metrik yang menghitung berapa banyak kesalahan kata pada suatu kalimat.

3.3.6 Dokumentasi

Sistem akan didokumentasikan melalui kode program yang dibuat serta foto-foto hasil deteksi yang dilakukan. Dokumentasi ini akan mencakup proses pengembangan sistem serta hasil-hasil yang diperoleh dari penelitian ini.