BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pengenalan ucapan otomatis, atau yang lebih dikenal sebagai automatic speech recognition (ASR), memiliki berbagai kegunaan yang berdampak besar [1]. Aplikasi paling populer di antaranya adalah fungsi sebagai penerjemah suara langsung dan penyedia subtitle untuk video [2]. Selain itu, riset-riset terkait terhadap pengenalan ucapan otomatis telah membuka peluang-peluang baru dalam melestarikan bahasa yang hampir punah melalui pengenalan ucapan otomatis [3]. Namun, penerapan sistem ASR pada bahasa-bahasa yang dikategorikan sebagai bahasa berdaya rendah (*low-resource language*) masih menghadapi berbagai tantangan [2, 4].

Sebagai contoh, Indonesia merupakan negara yang kaya akan keragaman linguistik dengan lebih dari 700 bahasa. Tetapi sayangnya, 400 dari bahasanya terancam punah dan sementara itu 12 di antaranya telah dikategorikan punah [5, 6]. Hal ini bukanlah masalah yang kecil, sebab bahasa terpopuler di Indonesia yaitu bahasa Jawa dan Sunda [7] juga masih dikategorikan sebagai bahasa berdaya rendah [8–11].

Dalam upaya mengatasi tantangan bahasa berdaya rendah dalam pengembangan sistem ASR, kekurangan data terlatih atau kebutuhan akan volume data riset yang berkualitas merupakan penghambat utama [2, 4, 12]. Untungnya, dengan semakin canggihnya teknologi saat ini, penelitian terbaru menunjukkan bahwa masalah ini dapat diatasi melalui penerapan teknik augmentasi data [2, 4, 13–15]. Metode ini melibatkan penambahan variasi pada *dataset* yang ada atau penciptaan data sintetis untuk meningkatkan kinerja model ASR dalam berbagai kondisi [8, 16]. Pendekatan augmentasi data tidak hanya memperkaya *dataset* pelatihan tetapi juga memperkuat kemampuan adaptasi model terhadap variasi linguistik dan akustik, membuka jalan bagi peningkatan aksesibilitas dan efektivitas teknologi ASR di kalangan komunitas bahasa yang lebih luas [17–20]. Teknik ini menjadi komponen penting dalam pengembangan sistem pengenalan suara yang efisien dan adaptif [4, 21, 22].

Secara umum, pendekatan teknik augmentasi pada data audio sangat bervariatif. Hal ini dapat melibatkan sejumlah metode seperti penambahan

Gaussian noise untuk meningkatkan ketahanan model terhadap gangguan latar belakang [23], modifikasi kecepatan audio melalui time stretch untuk membantu sistem mengenali ucapan dengan variasi tempo yang beragam [24], *pitch shift* untuk mengubah *pitch* audio tanpa mengubah durasi [25], dan shift untuk menggeser sinyal audio, memperkenalkan variasi temporal pada data [24].

Terkait dengan riset augmentasi pada bahasa berdaya rendah, sebuah riset menunjukkan bahwa penggunaan strategi augmentasi dan ensembling dapat meningkatkan skor BLEU sebesar 40% [2]. Lalu, pada riset lain dengan tolak ukur persentase kalimat yang salah atau kerap disebut sebagai *word error rate* (WER), terjadi peningkatan dari nilai 30,1%-53,3% menjadi 6,3%-13,9% pada bahasa seperti Gronings, West-Frisian, Besemah, dan Nasal [4]. Selain itu, penambahan distorsi pada data text-to-speech juga berhasil menurunkan WER dari 31,48% menjadi 25,13% [15]. Implementasinya juga sudah dipermudahkan dengan adanya Inovasi-inovasi yang telah terbukti dapat memberikan peningkatan kinerja model seperti Audiomentations, SpecAugment, TorchAudio, dan MixSpeech [15, 26, 27].

Selain riset yang membahas tentang augmentasi pada bahasa berdaya rendah, juga terdapat riset yang membahas tentang implementasinya ASR pada bahasa sumber daya rendah seperti bahasa Indonesia, Jawa, dan Sunda. Pada studi ini, pengembangannya menggunakan dua model yang berbeda yaitu model XLSR-53 dan XLS-R 300m. Pada akhir pengembangan model, ketika kedua model dibandingkan ditemukan bahwa model XLS-R 300m memberikan performa yang lebih baik dibandingkan model XLSR-53 dimana XLS-R 300m mendapatkan WER sebesar 5.43 dan XLSR-53 mendapatkan WER sebesar 5.77. Performa ini dicapai melalui penggabungan tujuh *dataset* yang berbeda-beda, di antaranya adalah TITML-IDN, Magic Data (Indonesian Scripted Speech Corpus—Daily Use Sentence), Common Voice, OpenSLR—Large Javanese & Sundanese ASR training data set (SLR35 & SLR36), serta OpenSLR—High-quality TTS data for Javanese & Sundanese (SLR41 & SLR44). Perlu diketahui bahwa terdapat dua faktor yang mempengaruhi penelitian ini. Faktor pertama merupakan data XLS-R 300m yang lebih kompleks dan membutuhkan *dataset* yang lebih besar dibandingkan XLSR-53, dan kedua model juga menggunakan teknik lanjutan seperti KenLM 5-gram untuk optimasi model [8].

Pada riset yang menggunakan bahasa yang berbeda dan menggunakan *dataset* yang lebih sedikit. Pada tiga model yang berbeda yaitu Kaldi, DeepSpeech, dan Wav2Vec2, ditemukan bahwa performa WER yang paling baik didapatkan pada model Kaldi dengan angka 17,9% yang dilanjuti oleh wav2vec2 dengan

WER sebesar 22,9% dan posisi terakhir diambil oleh model DeepSpeech dengan WER sebesar 41,1%. Setelah dilakukan pembelajaran mesin, riset ini juga melakukan pendekatan *real-life* dimana model diuji dengan ucapan secara langsung. Menariknya, riset ini berkata bahwa model dengan WER terbaik pada pendekatan *real-life* bukanlah model dengan WER terbaik pada pembelajaran melainkan model kedua yaitu Wav2Vec2. Hal ini bisa dikarenakan *character error rate* (CER) model Wav2Vec2 yang jauh lebih baik dibandingkan Kaldi dimana Wav2Vec2 memiliki CER sebesar 6,1% dan Kaldi sebesar 7,5%. Tetapi perlu diingat, bahwa berbeda dengan bahasa Indonesia, bahasa Maori banyak memperhatikan nada pada suatu karakter sama halnya seperti bahasa mandarin.

Terakhir, pada riset yang serupa dengan riset pengenalan ucapan berbahasa Indonesia, Jawa, dan Sunda [8], sebuah sistem dilakukan pada data *dataset* yang berbeda tetapi dengan model yang sama yaitu wav2vec2. Dengan menggunakan model wav2vec2.0-pt(pre-trained) dan wav2vec2.0-ft(fine-tuned) didapatkan hasil kinerja pada beberapa kondisi yaitu *Single-task training*, *Monolingual Multi-task Training*, dan *Multilingual Multi-task Training* pada *dataset* NusaASR. Dapat dikonklusikan pada tiga kondisi datas bahwa model wav2vec2.0-ft bekerja lebih baik dibandingkan wav2vec2.0-pt pada *multilingual multi-task training* dengan rata-rata WER sebesar 20% [28].

Berdasarkan latar belakang dan beberapa riset banding, dilakukan eksplorasi lanjutan terhadap kinerja hasil augmentasi audio pada model wav2vec2.0 yang menujukkan performanya yang lebih stabil dan optimal dibandingkan model lainnya seperti Kaldi dan DeepSpeech.

1.2 Rumusan Masalah

1. Bagaimana cara mengimplementasikan augmentasi audio untuk pengenalan ucapan pada model Wav2Vec2 dalam bahasa Jawa dan Sunda.
2. Apa dampak penggunaan augmentasi audio pada bahasa Jawa dan Sunda.

1.3 Batasan Permasalahan

1. Penelitian ini menggunakan Audiomentations untuk implementasi augmentasi audio.
2. Model yang digunakan pada penelitian ini adalah wav2vec2-large-xlsr-53.

1.4 Tujuan Penelitian

1. Mengimplementasikan augmentasi audio untuk pengenalan ucapan pada model Wav2Vec2 dalam bahasa Jawa dan Sunda.
2. Menganalisa dampak penggunaan augmentasi audio pada bahasa Jawa dan Sunda.

1.5 Manfaat Penelitian

1. Dengan menerapkan strategi augmentasi data, penelitian ini berkontribusi pada peningkatan efektivitas sistem ASR pada bahasa yang hampir punah.
2. Penelitian ini menunjukkan bagaimana teknologi ASR dapat dimanfaatkan untuk pelestarian bahasanya dan penerapannya untuk mendokumentasikan sebuah bahasa yang hampir punah.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini terdiri dari enam bagian utama, yaitu latar belakang masalah yang memberikan konteks penelitian, rumusan masalah yang membahas pertanyaan penelitian, batasan permasalahan yang mendefinisikan ruang lingkup penelitian, tujuan penelitian yang menggambarkan hasil yang diharapkan, manfaat penelitian yang menjelaskan dampak dan kontribusi dari penelitian ini, serta sistematika penulisan yang memberikan gambaran struktur keseluruhan tesis.
- Bab 2 LANDASAN TEORI
  Bab ini menyajikan penjelasan mendalam tentang teori-teori, konsep-konsep dasar, dan arsitektur algoritma yang digunakan dalam penelitian ini. Bagian ini bertujuan untuk memberikan landasan teoretis yang kuat dan menjelaskan bagaimana teori-teori tersebut mendukung proses perancangan sistem serta implementasi algoritma yang akan diterapkan.
- Bab 3 METODOLOGI PENELITIAN
  Bab ini merinci tahapan-tahapan yang akan dilakukan selama penelitian,

termasuk alur kerja yang sistematis serta rancangan sistem yang lengkap dengan implementasi algoritma. Bagian ini juga disertai dengan gambar, diagram, dan tabel yang memvisualisasikan proses dan desain penelitian secara komprehensif.

- Bab 4 HASIL DAN DISKUSI
  Bab ini berfokus pada penjelasan tentang sistem yang digunakan untuk menjalankan penelitian, hasil implementasi algoritma, serta evaluasi akurasi yang dihasilkan. Tampilan dari sistem yang telah dibuat juga akan disajikan untuk memberikan gambaran konkret mengenai hasil penelitian ini.
- Bab 5 KESIMPULAN DAN SARAN
  Bab ini berisi ringkasan dari temuan-temuan penelitian yang telah dilakukan serta memberikan saran untuk penelitian lanjutan. Bagian ini bertujuan untuk menutup penelitian dengan menyajikan kesimpulan yang jelas dan rekomendasi yang informatif untuk pengembangan penelitian di masa mendatang.