BAB I PENDAHULUAN

1.1 Latar Belakang

Automatic speech recognition (ASR) atau jika diterjemahkan dalam bahasa indonesia adalah pengenalan ucapan otomatis yang merupakan salah satu cabang dari *deep learning* yang telah diadopsi oleh teknologi ini (Deng et al., 2013). *Automatic speech recognition* (ASR) adalah proses penggunaan algoritma dalam mesin komputasi untuk memodifikasi, menganalisis, dan mengenali pola tertentu dalam sinyal audio (Maruf et al., 2020). *Automatic Speech Recognition* (ASR) adalah teknologi yang memungkinkan perangkat untuk mengenali dan mentranskripsikan ucapan manusia menjadi teks. *Automatic speech recognition* (ASR) berhubungan dengan pemrosesan sinyal digital yang berkorelasi dengan pengenalan orang berdasarkan suara atau ucapannya ("Automatic Speaker Recognition Using MFCC and Artificial Neural Network," 2019). Salah satu tahapan penting dalam membangun sistem ASR yang akurat adalah proses ekstraksi fitur, yaitu metode untuk mengambil informasi penting dari sinyal audio. Pada penelitian (Kurzekar et al., 2014) membahas teknik ekstraksi fitur yang paling umum digunakan, seperti *Linear Predictive Coding* (LPC), *Mel-Frequency Cepstral Coefficients* (MFCC), *zero-crossing with peak amplitudo* (ZCPA), *Discrete Wavelet Transforms* (DTW), dan *relative spectral processor* (RASTA).

Dalam penelitian ini, model ASR dibangun menggunakan *Long Short-Term Memory* (LSTM) sebagai algoritma untuk menangani data sekuensial seperti sinyal suara. LSTM, sebagai salah satu jenis *Recurrent Neural Network* (RNN), efektif dalam memproses sinyal ucapan dengan durasi panjang karena mampu mempertahankan informasi penting dari urutan waktu yang panjang dan mengatasi masalah *vanishing gradient*. Selain itu, penelitian ini akan membandingkan satu non ekstraksi fitur dan tiga teknik ekstraksi fitur yang dipakai dalam membangun model ASR yaitu non-fitur ekstraksi, MFCC, RASTA-PLP, dan kombinasi keduanya (hybrid).

Mel-Frequency Cepstral Coefficients (MFCC) dikenal efektif dalam menangkap karakteristik frekuensi suara dan umum digunakan dalam bidang pengenalan suara karena MFCC memiliki akurasi pengenalan yang tinggi dan diskriminasi yang baik serta korelasi koefisien yang rendah (Anusuya & Katti, 2011) sehingga MFCC sangat baik dalam mengidentifikasi frekuensi suara. Namun, memiliki beberapa kelemahan dengan kelemahan atau kerugian utama dari MFCC adalah ketahanan yang buruk terhadap sinyal derau, karena sinyal derau mengubah semua MFCC meskipun setidaknya ada satu pita frekuensi yang menyimpang. Sedangkan, RASTA-PLP (*Relative Spectra–Perceptual Linear Prediction*) mengombinasikan teknik RASTA dengan metode PLP untuk meningkatkan ketahanan atau kekokohan pada fitur PLP, dikenal efektif dalam mengurangi pengaruh noise dan variasi lingkungan yang sering muncul saat rekaman audio dilakukan di tempat yang tidak ideal sehingga RASTA-PLP lebih unggul dalam mengurangi pengaruh noise.

Pada beberapa penelitian sebelumnya, penelitian (Labied & Belangour, 2021) telah membandingkan RASTA-PLP dengan teknik ekstraksi fitur LPC dan MFCC. Hasil yang diperoleh menunjukkan bahwa RASTA-PLP lebih baik dari MFCC dan LPC pada sinyal audio bising dengan akurasi 73%, sedangkan MFCC memiliki akurasi 60%, dan LPC memiliki akurasi 53%. Pada penelitian lainnya (Tazi & El Makhfi, 2017), telah membandingkan beberapa fitur ekstraksi yang telah dilakukan pada penelitian tersebut, dengan fitur ekstraksi yang dibandingkan adalah MFCC, PLP, RASTA-PLP, dan Hybrid (MFCC & RASTA-PLP). Berdasarkan hasil penelitian tersebut, setiap metode ekstraksi fitur menunjukkan akurasi yang berbeda pada berbagai level RSB (dB) dan hasil yang didapatkan menunjukan Hybrid sedikit lebih unggul dari beberapa fitur ekstraksi lainnya, dengan metode Hybrid mencapai akurasi rata-rata tertinggi sebesar 60,28%, diikuti oleh sebesar MFCC 56,09%, RASTA-PLP sebesar 55,73%, dan PLP sebesar 55,03%.

Dalam penelitian ini, pendekatan hybrid diterapkan pada ekstraksi fitur dengan menggabungkan keunggulan dari MFCC dan RASTA-PLP. MFCC tidak terlalu kompleks, akurasi pengenalan yang tinggi (Anusuya & Katti, 2011) dan

sangat baik dalam mengidentifikasi frekuensi suara, sedangkan RASTA-PLP efektif dalam mengurangi kebisingan dan membantu dalam mengurangi distorsi temporal yang disebabkan oleh lingkungan rekaman (Labied & Belangour, 2021). Dengan demikian, kombinasi fitur hybrid (MFCC & RASTA-PLP) diharapkan dapat meningkatkan performa ASR dalam kondisi lingkungan yang penuh noise, meskipun mungkin nantinya terdapat peningkatan kompleksitas pada proses ekstraksi fitur. Untuk itu, penelitian ini bertujuan untuk menguji apakah pendekatan hybrid mampu memberikan hasil yang signifikan dalam konteks bahasa Indonesia tanpa menambah beban komputasi dan kompleksitas secara berlebihan.

Pada penelitian ini, teknik *global dynamic pruning* berbasis low magnitude digunakan sebagai salah satu metode untuk mengoptimalkan kinerja model dengan mempertahankan akurasi yang tinggi namun mengurangi kompleksitas komputasinya. Jika performa yang dihasilkan memang signifikan namun disertai dengan kompleksitas yang tinggi, teknik *global dynamic pruning* berbasis low magnitude ini akan membantu menurunkan kompleksitas tanpa mengorbankan hasil performa pada model. *Dynamic pruning* adalah salah satu metode kompresi model yang efektif untuk mengurangi biaya komputasi jaringan (Shao et al., 2024). Dengan fokus penelitian ini adalah untuk melakukan perbandingan antara penggunaan non-fitur ekstraksi, MFCC, RASTA-PLP, dan hybrid dengan tetap melihat kompleksitas pada penggunaan fitur ektraksi Hybrid pada model ASR, untuk dapat menentukan metode terbaik saat memproses variasi akustik bahasa Indonesia secara optimal. Pendekatan ini akan memungkinkan penggunaan teknik fitur ektraksi Hybrid yang efisien dan tepat untuk konteks bahasa Indonesia, sehingga meningkatkan keandalan dan efisiensi model ASR yang dikembangkan.

1.2 Identifikasi Masalah

Berdasarkan latar belakang tersebut, isu-isu berikut ini yang dapat ditulis pada identifikasi masalah pada penelitian ini, berikut adalah identifikasi masalahnya:

1. Performa Beragam Metode Ekstraksi Fitur: Pemilihan fitur ekstraksi suara merupakan komponen krusial yang dapat memengaruhi performa pengenalan suara. Ada berbagai metode ekstraksi fitur yang dapat digunakan, diantaranya adalah MFCC dan RASTA-PLP, yang merupakan teknik populer. Setiap teknik memiliki karakteristik yang berbeda, sehingga performa yang dihasilkan akan berbeda pada ASR bahasa indonesia berbasis LSTM.
2. Kurangnya Penelitian tentang Kombinasi Ekstraksi Fitur: Selain menggunakan metode ekstraksi tunggal seperti MFCC atau RASTA-PLP, terdapat potensi untuk menggabungkan kedua metode (*hybrid*) guna meningkatkan performa sistem ASR. Namun, masih kurangnya penelitian yang secara spesifik menganalisis dampak kombinasi metode ini dalam sistem pengenalan suara berbasis LSTM.
3. Kompleksitas model ASR berbasis LSTM: Penerapan teknik hybrid ekstraksi fitur pada sistem ASR berbasis LSTM dapat meningkatkan kebutuhan komputasi, waktu pelatihan yang lama, serta memori yang tinggi. oleh karena itu, diperlukan teknik seperti dynamic pruning berbasis magnitude untuk mengurangi jumlah parameter tanpa mengurangi performa pada model ASR.

1.3 Tujuan Penelitian

Berikut ini adalah tujuan dari penelitian model ASR yang berbasiskan *Long Short-Term Memory* (LSTM):

1. Mengevaluasi kompleksitas model ASR yang menggunakan fitur hybrid, terutama dalam hal kebutuhan komputasi dan waktu proses, untuk memastikan bahwa pendekatan hybrid dapat memberikan peningkatan performa yang bagus tanpa meningkatkan kompleksitas secara berlebihan.

2. Menerapkan model LSTM untuk mengenali urutan sinyal suara dengan menggunakan teknik ekstraksi fitur hybrid dan membandingkannya dengan metode ekstraksi non-fitur, MFCC, serta RASTA-PLP secara individual, untuk mengevaluasi apakah pendekatan hybrid dapat meningkatkan akurasi dan ketahanan sistem ASR.
3. Melakukan perbandingan performa model ASR berbasis LSTM untuk bahasa Indonesia dengan teknik ekstraksi fitur non-fitur, MFCC, RASTA-PLP, dan hybrid (kombinasi MFCC & RASTA-PLP) pada data memiliki noise dan tidak memiliki noise, guna mengidentifikasi metode yang paling efektif dalam menghasilkan representasi suara yang mendukung akurasi pengenalan ucapan pada dua kondisi yang berbeda.

1.4 Manfaat Penelitian

Penelitian ini membahas dua jenis manfaat yaitu manfaat teoritis dan manfaat praktis. Manfaat dari penelitian ini antara lain:

1) Manfaat Teoritis

   Penelitian ini memiliki manfaat teoritis yang diharapkan dapat memberikan kontribusi dalam pengembangan pengetahuan tentang metode ekstraksi fitur MFCC, RASTA-PLP dan hybrid dalam sistem pengenalan suara berbasis LSTM, khususnya dalam konteks Bahasa Indonesia.

2) Manfaat Praktis

   Penelitian ini memiliki manfaat praktis dalam pengembangan sistem ASR untuk bahasa Indonesia. Hasil dari penelitian ini dapat dijadikan sebagai acuan untuk perbandingan, yang akan membantu dalam memilih metode ekstraksi fitur yang paling tepat, baik untuk keperluan industri maupun penelitian lebih lanjut.

1.5 Batasan Masalah

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan, berikut batasan masalahnya:

1) Model yang dibangun pada penelitian pengenalan ucapan berbahasa Indonesia ini hanya menggunakan LSTM.
2) Penelitian ini hanya membandingkan satu teknik non fitur ekstraksi dan tiga teknik ekstraksi fitur, yaitu MFCC, RASTA-PLP, dan kombinasi keduanya (Hybrid).
3) Hasil koefisien yang dihasilkan oleh fitur ekstraksi tunggal hanya berjumlah 13 koefisien sehingga hasil koefisien yang didapat dari fitur ekstraksi Hybrid berjumlah sekitar 26 koefisien.
4) Persentase pruning yang dilakukan pada penelitian ini hanya sebesar 30% yang dilakukan pada model ASR untuk pruning rasionya.
5) Penelitian ini membatasi jumlah epoch dalam proses training model, dengan hanya melatih pada rentang 100 *epoch* karena keterbatasan sumber daya sehingga fokus utama hanya untuk membandingkan relatif performa antar metode.