BAB I

PENDAHULUAN

1.1 Latar Belakang

Emosi merupakan bagian penting dalam interaksi manusia yang turut berperan dalam pengembangan sistem interaksi manusia dan komputer. Salah satu implementasinya adalah *Speech Emotion Recognition* (SER), yaitu teknologi yang memungkinkan sistem mengenali kondisi emosional pengguna melalui sinyal ujaran. Teknologi SER telah diterapkan pada berbagai bidang seperti layanan pelanggan, pendidikan, kesehatan, dan sistem tanggap darurat (Waleed Alsabhan, 2023). Meskipun perkembangan SER telah cukup pesat, penerapannya pada bahasa Indonesia masih menghadapi berbagai tantangan. Salah satu penyebabnya adalah karakteristik prosodi, intonasi, dan pola pengucapan bahasa Indonesia yang berbeda dengan dataset internasional seperti RAVDESS dan EMO-DB (Alluhaidan dkk., 2023; Sailor, 2025). Penelitian oleh Bustamin dkk. (2024) mengembangkan dataset *IndoWaveSentiment* yang mencakup emosi netral, senang, terkejut, jijik, dan kecewa sebagai salah satu upaya pengembangan sumber dataset lokal berbahasa Indonesia.

Salah satu penelitian SER berbahasa Indonesia dilakukan oleh Septiyanto dkk. (2025) menggunakan fitur *Mel-Frequency Cepstral Coefficients* (MFCC) dan arsitektur *Bidirectional Long Short-Term Memory* (BiLSTM) pada dataset *IndoWaveSentiment*. Penelitian tersebut menunjukkan bahwa BiLSTM mampu mempelajari dependensi temporal dua arah pada sinyal ujaran dengan baik sehingga menghasilkan performa klasifikasi emosi yang cukup menjanjikan. Penggunaan fitur MFCC secara tunggal masih memiliki keterbatasan dalam

merepresentasikan informasi emosional secara komprehensif. MFCC umumnya hanya merepresentasikan karakteristik spektral dasar dan kurang mampu menangkap dinamika harmonik, selubung energi, serta karakteristik perseptual pendengaran yang berpengaruh terhadap emosi pada sinyal suara (E & Kulkarni, 2025).

Beberapa penelitian mulai mengeksplorasi pendekatan *early fusion* dengan menggabungkan beberapa fitur akustik agar representasi emosi yang dihasilkan menjadi lebih kaya. Penelitian oleh Gumelar dkk. (2020) menunjukkan bahwa kombinasi fitur *Hilbert Spectrum* dan *Cochleagram* mampu meningkatkan representasi spektral emosi pada sinyal suara melalui pendekatan *deep learning* berbasis CNN. Penelitian oleh Peng dkk. (2023) menunjukkan bahwa fitur berbasis *Cochleagram* mampu merepresentasikan mekanisme pendengaran manusia secara lebih biologis sehingga efektif dalam meningkatkan performa pengenalan emosi. Kombinasi fitur MFCC, *Hilbert Multispectrum*, dan *Cochleagram* berpotensi menghasilkan representasi emosi yang lebih komprehensif. Keunggulan tersebut diperoleh karena ketiga fitur tersebut mampu menangkap informasi spektral, harmonik, dan karakteristik pendengaran secara saling melengkapi.

Arsitektur model juga mempengaruhi performa SER selain pemilihan fitur. Meskipun BiLSTM efektif dalam mempelajari dependensi temporal, arsitektur tersebut memiliki keterbatasan dalam mengekstraksi pola lokal dan karakteristik spasial pada representasi spektral audio. *Convolutional Neural Network* (CNN) diketahui mampu mengekstraksi pola lokal pada sinyal audio secara efektif.

Beberapa penelitian mulai mengembangkan arsitektur *hybrid* CNN-BiLSTM untuk menggabungkan kemampuan CNN dalam ekstraksi fitur lokal dan kemampuan BiLSTM dalam mempelajari dependensi temporal jangka panjang (Hussein dkk., 2025; X. Lu, 2022; Tiwari, 2025). Penelitian terbaru menunjukkan bahwa penambahan mekanisme *attention* pada arsitektur CNN-BiLSTM mampu meningkatkan fokus model terhadap bagian sinyal yang mengandung informasi emosional paling relevan sehingga performa klasifikasi emosi menjadi lebih baik (Zennou, 2026). Mekanisme *attention* juga dinilai efektif pada kondisi dataset terbatas karena membantu model memfokuskan proses pembelajaran pada informasi emosional yang dominan.

Pemilihan fitur, arsitektur model, serta metode evaluasi juga memegang peranan penting dalam menilai kemampuan generalisasi sistem *Speech Emotion Recognition*. Banyak penelitian masih melakukan evaluasi menggunakan pembagian data secara acak sehingga data dari pembicara yang sama berpotensi muncul pada data pelatihan dan pengujian (Ibrahim dkk., 2021). Kondisi tersebut dapat menghasilkan estimasi performa yang terlalu optimistis karena model tidak hanya mempelajari karakteristik emosi, tetapi juga karakteristik individu pembicara (Gjoreski dkk., 2014). Penelitian ini menggunakan metode *Leave-One-Speaker-Out* (LOSO) *Cross Validation*. Metode *Leave-One-Speaker-Out* (LOSO) *Cross Validation* mengevaluasi model menggunakan data pembicara yang tidak pernah digunakan selama proses pelatihan. Performa model yang diperoleh lebih merepresentasikan kondisi penerapan di dunia nyata. (Ibrahim dkk., 2021).

Kajian penelitian terdahulu menunjukkan masih terdapat beberapa celah

penelitian pada pengembangan *Speech Emotion Recognition* berbahasa Indonesia, khususnya pada dataset *IndoWaveSentiment*. Sebagian besar penelitian sebelumnya masih menggunakan fitur tunggal seperti MFCC dan belum mengeksplorasi penggabungan beberapa representasi akustik yang saling melengkapi. Penerapan arsitektur *hybrid* CNN-BiLSTM yang dipadukan dengan mekanisme *Temporal Attention* pada dataset berbahasa Indonesia masih terbatas. Aspek evaluasi pada sebagian besar penelitian masih menggunakan pembagian data secara acak sehingga kemampuan generalisasi model terhadap pembicara baru belum dievaluasi secara optimal. Penelitian ini mengusulkan integrasi *early fusion* fitur MFCC, *Hilbert Multispectrum, dan Cochleagram* menggunakan arsitektur CNN-BiLSTM berbasis *Temporal Attention* yang dievaluasi melalui metode *Leave-One-Speaker-Out* (LOSO) *Cross Validation*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, diperoleh rumusan masalah sebagai berikut:

1. Bagaimana meningkatkan representasi fitur akustik agar mampu menangkap karakteristik emosi secara lebih komprehensif pada *Speech Emotion Recognition* berbahasa Indonesia?
2. Bagaimana meningkatkan kemampuan generalisasi model *Speech Emotion Recognition* terhadap pembicara yang belum pernah digunakan selama proses pelatihan?

1.3 Tujuan Penelitian

Penelitian ini dilakukan menggunakan arsitektur CNN-BiLSTM sebagai

pendekatan dalam sistem pengenalan emosi suara berbahasa Indonesia. Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Mengevaluasi efektivitas penerapan early fusion fitur akustik MFCC, *Hilbert Multispectrum*, dan *Cochleagram* pada arsitektur CNN-BiLSTM dengan *Temporal Attention* dalam meningkatkan representasi fitur dan performa *Speech Emotion Recognition* berbahasa Indonesia.
2. Mengevaluasi kemampuan generalisasi model menggunakan *Leave-One-Speaker-Out Cross Validation* pada skenario *speaker-independent.*

1.4 Manfaat Penelitian

1.4.1 Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan manfaat secara teoritis sebagai berikut:

1. Memberikan kontribusi dalam pengembangan kajian *Speech Emotion Recognition* (SER), khususnya terkait integrasi fitur akustik MFCC, *Hilbert Multispectrum,* dan *Cochleagram* untuk merepresentasikan emosi pada sinyal ujaran berbahasa Indonesia.
2. Memberikan kontribusi terhadap pengembangan metode *deep learning* untuk *Speech Emotion Recognition* melalui penerapan arsitektur *hybrid* CNN-BiLSTM berbasis *Temporal Attention* dalam proses klasifikasi emosi ujaran.
3. Menjadi referensi ilmiah bagi penelitian selanjutnya yang berfokus pada pengembangan sistem *Speech Emotion Recognition* berbasis dataset lokal, khususnya pada kondisi *low-resource* dataset dengan pendekatan fusi fitur akustik dan pemodelan temporal-spasial.

1.4.2 Manfaat Praktis

Penelitian ini diharapkan dapat memberikan manfaat secara praktis sebagai berikut:

1. Memberikan pendekatan alternatif dalam pengembangan sistem *Speech Emotion Recognitio*n (SER) berbahasa Indonesia melalui integrasi fitur MFCC, *Hilbert Multispectrum*, dan *Cochleagram* menggunakan arsitektur CNN-BiLSTM berbasis *Temporal Attention.*
2. Menjadi referensi awal untuk pengembangan dan penelitian sistem *Speech Emotion Recognition* pada kondisi dataset lokal yang terbatas.
3. Menjadi dasar pengembangan penelitian lanjutan terkait penerapan *feature fusion* dan *attention-based deep learning* pada pengenalan emosi ujaran berbahasa Indonesia.

1.5 Batasan Penelitian

Penelitian ini dibatasi pada beberapa aspek berikut:

1. Penelitian menggunakan dataset *IndoWaveSentiment* yang terdiri dari lima kategori emosi, yaitu netral, senang, terkejut, jijik, dan kecewa, tanpa mempertimbangkan tingkat intensitas emosi sebagai kelas terpisah.
2. Fitur akustik yang digunakan dibatasi pada MFCC, *Hilbert Multispectrum*, dan *Cochleagram* yang dikombinasikan menggunakan pendekatan *early fusion* melalui proses konkatenasi fitur.
3. Arsitektur model yang digunakan dibatasi pada *hybrid* CNN-BiLSTM berbasis *Temporal Attention* tanpa melakukan perbandingan dengan arsitektur lain seperti GRU, *Transformer*, maupun model *attention* berbasis *self-*

attention.

4. Penelitian menggunakan teknik augmentasi audio berupa penambahan *noise, time stretching, pitch shifting, time shifting,* dan *speed perturbation*.
5. Evaluasi model dilakukan menggunakan metode *Leave-One-Speaker-Out* (LOSO) *Cross Validation* untuk mengukur kemampuan generalisasi model terhadap pembicara yang belum pernah dilihat selama proses pelatihan.
6. Evaluasi performa model difokuskan pada metrik *accuracy*, *precision, recall, F1-score, confusion matrix*, serta analisis performa berdasarkan emosi dan pembicara.
7. Penelitian ini hanya berfokus pada analisis emosi berdasarkan karakteristik akustik sinyal suara dan tidak membahas aspek linguistik, semantik, maupun konteks kalimat dalam ujaran.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran mengenai alur pembahasan dalam laporan penelitian. Adapun sistematika penulisan skripsi ini adalah sebagai berikut.

Bab I Pendahuluan berisi latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, serta sistematika penulisan.

Bab II Landasan Teori berisi kajian teori yang mendukung penelitian, meliputi *Speech Emotion Recognition,* pengolahan sinyal audio digital, augmentasi audio, fitur akustik, *early fusion*, *Convolutional Neural Network* (CNN), *Bidirectional Long Short-Term Memory* (BiLSTM), *Temporal Attention,*

serta penelitian terdahulu yang relevan.

Bab III Metodologi Penelitian menjelaskan tahapan penelitian yang dilakukan, meliputi pengumpulan dataset, pra-pemrosesan data audio, augmentasi data, ekstraksi dan penggabungan fitur akustik, perancangan model CNN-BiLSTM berbasis *Temporal Attention,* serta metode evaluasi menggunakan *Leave-One-Speaker-Out* (LOSO) *Cross Validation.*

Bab IV Hasil dan Pembahasan berisi hasil eksperimen yang diperoleh dari penerapan model yang diusulkan serta analisis pengaruh penggabungan fitur akustik terhadap performa model dalam mengenali emosi ujaran berbahasa Indonesia.

Bab V Penutup berisi kesimpulan dari hasil penelitian yang telah dilakukan serta saran untuk pengembangan penelitian selanjutnya.